NCEA ONLINE RESEARCH: MORE THAN ONE DIGITAL NCEA EXTERNAL ASSESSMENT OPPORTUNITY PER ANNUM

Final report

Don Klinger, Suzanne Trask,
Bronwen Cowie

August 2020

Wilf Malcolm Institute of Educational Research
Te Pūtahi Rangahau Mātauranga o Wilf Malcolm
THE UNIVERSITY OF WAIKATO

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Acknowledgements

Our thanks go to members of the Expert Review Panel who provided feedback on this report:

Professor Christine Harrison, King's College London

Professor Dawn Penny, Edith Cowan University, Perth

Associate Professor Nigel Calder, The University of Waikato

Associate Professor Maurice Cheng, The University of Waikato

Dr Frances Edwards, University of Waikato

Dr Phillipa Hunter, The University of Waikato

Dr Elizabeth Reinsfeld, The University of Waikato

TABLE OF CONTENTS

LIST OF TABLES

Executive Summary

As part of its mandate outlined in the Our Future State portfolio *Quality for the Future World – Kia noho takatū ki tō āmua ao*, the New Zealand Qualifications Authority (NZQA) is actively exploring an online digital assessment programme to inform future structures, tools and practices of summative digital assessment in the context of the National Certificate for Educational Assessment (NCEA). Such innovations within NCEA have the potential to provide when ready assessments that provide more timely and accurate feedback to students, enhance the monitoring of students' learning, and provide support for subsequent learning and teaching. Globally, there has been a shift towards both digital and "on demand/when ready" assessment. Assessment that provides formative feedback that is sufficiently frequent and/or has stakes attached has long been positively associated with increased learning and achievement. Current digital technologies can further enhance the positive benefits of such assessments. Surprisingly, there has been little systematic review of practices in place, the challenges and potentials of such assessment practices, or the opportunities for innovative models to enhance the value of such assessments beyond the summative measurement of student achievement.

The purpose of this review was to examine the current body of research literature and summarise relevant digital and "when ready" models of large-scale summative assessment currently in practice. The combined findings provide a foundation for the implementation of a large-scale assessment programme that can increase personalisation of learning (individual learning progression and responsiveness to test readiness), enhance school and sector flexibility in curriculum decision-making and scheduling of testing, and use digital/digitised test administration and management to effectively develop, implement and monitor testing processes and test takers' achievement, across administrations. Such assessments have the potential to offer learners greater ownership of their learning progress.

1. How does the timing and type of summative assessment/assessment impact sustained, deep learning?
2. What is the impact on reliability/decision consistency and the validity of interpretations when assessments are administered at multiple time points?
3. What are the practical, structural and measurement challenges and opportunities of summative, when ready assessment processes?

Testing when ready and multiple testing opportunities have implications for curriculum, the assessment programme design, and the use of formative assessment to determine test readiness.

The determination of readiness is a key pillar for design and implementation. What is readiness and who determines readiness? A second pillar is the format and structure of the assessment programme. Currently, the majority of digitised assessments use multiple-choice and open response items, however, as evidenced in this report, testing programmes are being expanded to include digital portfolios or performance assessments to effectively measure student performance in relation to learning standards. These assessment programmes require an administrative and school infrastructure to support the flexibility required to cope with on-demand or when ready assessment. Digital systems provide a mechanism for such implementation but must work across platforms. In the absence of dedicated assessment centres, schools must be able to provide suitable seating/space for greater or fewer numbers of students accessing online assessment at any given time, and provide staffing to monitor and moderate assessments at multiple time points.

When ready testing requires policies and procedures that assure security, reliability, fairness, and accuracy of assessments, and strengthening any one of these may not necessarily impact positively on the others. As a result, decisions in relation to these aspects will likely result in compromises. As an example, digital assessments can improve construct validity and deliver pedagogically rich assessment environments that assess higher order thinking and information processing; however, these methods can complicate the judgement process, increasing the potential for error or bias and reducing reliability.

Attention to the technical aspects of when ready digital assessments are paramount and will shape the policies and practices for the assessment programme.

- The accuracy of on-demand testing is dependent on the provision of safeguards such as test security and proctoring, and both in-person and remote models of live or delayed (video) proctoring have been investigated. eExams are no more or less secure than paper-based exams but the level of security depends on the type of exam and the protections that are in place.
- When ready testing can use different systems for marking; computer-marked or person-marked/moderated, or both. Automated scoring is commonly used for closed or multiple-choice or short open response questions but there are systems for computer-marking of more complex formats.
- Accessibility and usability are important in the design specification of digital assessments. Tools and mechanisms must be in place that accommodate flexible assessment formats and pathways and offer possibilities for learners to exercise agency. Nevertheless, flexible and/or individualised programmes of learning and assessment do not automatically ensure equity.
- More flexible or individualised programmes of learning and assessment depend on equitable and appropriate resourcing to ensure that all learners including low performing or minority groups have equal access and opportunities to achieve and demonstrate their learning. Attention needs to be given to context, concepts, and linguistic demands.

1. Introduction

This research review was commissioned by the New Zealand Qualifications Authority (NZQA) in January 2020. The NZQA is working towards long-term goals for schools and learners as outlined in the Our Future State portfolio *Qualify for the Future World – Kia noho takatū ki tō āmua ao* (NZQA, 2018). Part of this work is the development of the NCEA Online digital assessment programme (NZQA, n.d.-a).

The purpose of this review is to investigate areas of potential interest and prioritise ideas for future development and innovation trials under the NCEA Online programme. The research team was tasked with conducting a general review and synthesis of key findings from research studies and grey literature to inform the prospect of providing multiple digital external assessment opportunities per annum to students in high-stakes summative assessment for issuing qualifications (credentialing).

It is intended that findings of the review will be used to inform future structures, tools and practices of summative digital assessment in the context of the National Certificate for Educational Assessment (NCEA) in Aotearoa New Zealand.

The first part of this report outlines findings from the literature review that address three global questions:

- How does the timing and type of summative assessment/assessment as a core component in the measurement and awarding of qualifications in senior secondary school, impact on sustained, deep learning?
- What is the impact on reliability and decision consistency and the validity of the interpretations made from the results of summative assessments when administered at multiple time points, including the impacts teachers' and students' pursuit of learning?
- What are the practical, structural and measurement challenges and opportunities associated with implementing a summative assessment when ready process?

The second part of this report consists of a discussion paper which analyses potential benefits and challenges for the New Zealand context of offering some externally assessed assessments (examinations or other forms) at mid-year rather than at year end, focussing on the digital mode.

2. Framing/scope and methodology of review

Over the past thirty years New Zealand assessment policy has foregrounded the formative role of assessment - assessment for learning (Absolum et al., 2009; Bell & Cowie, 2001; Crooks, 1988, 1993, 2011; Department of Education, 1994; Hipkins, 2005; Ministry of Education, 1993, 2007, 2011). Assessment has been positioned as integral to effective teaching and learning, including student self-regulation, with parents able to make a contribution to their children's learning through their involvement in assessment (Ministry of Education, 2011). From the mid-1970s, New Zealand has used a high trust model with partial or total internal assessment a feature of all senior secondary qualifications, including past University Entrance and University Bursary credentialing processes. A combination of internal and external standards have been part of the NCEA since its introduction in the 2000s. As an overview, we explain the current structure and assessment philosophies of NCEA (Section 2.1) along with the relevant aspects of the 2019 NCEA Review and Change Package (Section 2.2). We consider the guiding principles from the NZQA Digital Assessment Vision (Section 2.3). Along with the key criteria for summative assessment from the literature (Section 2.4), we consider these together with future directions for NCEA and digital assessment to propose criteria for NCEA summative assessment. It is these criteria that frame the thinking and analyses in this report.

2.1 NCEA current structure

Achievement objectives in the National Curriculum (consisting of the English medium New Zealand Curriculum (NZC) and the Māori medium counterpart Te Marautanga o Aotearoa) provide the basis for assessment in New Zealand schools. Senior secondary learning (Years 11 to 13, Curriculum Levels 6 to 8) is assessed at levels one to three of the National Qualifications Framework (NQF) by the standards-based National Certificate of Educational Achievement (NCEA) which is administered by the New Zealand Qualifications Authority (NZQA) (Darr, 2019). Students typically complete Level 1 in Year 11, Level 2 in Year 12 and Level 3 in Year 13, although it is also the norm for students to complete credits across different levels within one year of study (Hipkins et al., 2016). The NCEA summarises and reports on students' achievements for students themselves, their whānau, the school, future learning providers, employers, and the Ministry of Education (Education Review Office, 2007; Ministry of Education, 2007a; Mutch, 2012). The NCEA has a widely accepted status as a high quality, rigorously monitored qualification system with high validity and reliability of results (Hipkins et al., 2016; OECD, 2011; Wylie & Bonne, 2016).

Students gain credits towards the NCEA by completing separate but discipline-related achievement standards. These are usually grouped together into courses such as English, science, or history. Achievement standards are typically 'worth' three or four credits. A course usually comprises 18 to 24 credits. A minimum of 80 credits is required for award of the qualification at each level. Achievement standards specify a range of internal or external assessment tasks, each consisting of detailed achievement criteria which are assessed at Excellence, Merit, Achieved, or Not Achieved. High performing students can gain course and/or certificate endorsement. For example, course endorsement at Merit level is awarded if at least 14 Merit credits are achieved. A certificate endorsement is awarded if students gain at least 50 credits at the level of endorsement across all courses (NZQA, n.d.-d; Hipkins et al., 2016).

Internal achievement standards offer teachers the freedom to teach and assess using learning contexts and assessment tasks, formats and timing that best suit their learners (Hipkins et al., 2016). Guidelines for internal assessment state that "students should not be assessed for a standard until the teacher is confident that achievement of the standard is within their reach, or until the final deadline for assessment, if there is one" (NZQA, n.d.). Thus, student learning and preparation of evidence for summative internal assessment typically involves cycles of formative assessment and feedback as part of level-appropriate guidance or supervision from teachers. However, there is evidence that under some conditions, such as teachers and students under pressure to collect credits or improve results, (Hipkins, 2015; Wylie & Bonne, 2016), the validity of some types of internal assessment is undermined by practices aimed at helping students 'get through' (East, 2014; Gillon & Stotter, 2012; Hume & Coll, 2009; Moeed, 2010; Thorpe, 2012).

Evidence of achievement in internal standards is evaluated and moderated by teachers themselves. Random samples of student work are selected for external moderation. External moderation of internal assessment tasks serves the dual purpose of ensuring the assessment task is suitable and reflects the requirements of the standard, and moderating teachers' judgements of the evidence presented (Cowie & Penney, 2015; NZQA, n.d.-b; n.d.-c). Schools can decide to offer one further assessment opportunity for any standard at their discretion for students who receive Not Achieved to move to Achieved, or from Achieved to Merit or Excellence after further learning has taken place. Immediately following an assessment event, one resubmission is permitted for students to identify minor issues or errors where these prevent the awarding of a higher grade. No teacher guidance in the form of specific feedback or teaching is permitted (NZQA, n.d.).

External standards are usually assessed in the form of written, externally set and marked examinations with up to three achievement standards per course being assessed. For courses with a practical component such as art, the assessment may be via portfolio submission (NZQA, n.d.-b). There is one

opportunity per school calendar year with external examinations held over a 3-week period beginning in early November. Marked examination booklets are returned to students and they may apply for a reconsideration of their results (NZQA, n.d.-e). Student learning and preparation for external examinations characteristically takes the form of formative assessments such as recall quizzes and practice essays or long-answer exam-type questions with feedback as elements of good teaching.

The NZC states that assessment must be "suited to purpose" and "chosen to suit the nature of the learning being assessed" (Ministry of Education, 2007a, p. 40). The Ministry of Education assessment position is that progress does not look the same for all learners and the system must adjust to meet learners' diverse needs, rather than the other way around (Ministry of Education, 2010a). Achievement standards are thus designed to credential meaningful learning in a wide range of knowledge, skills, and competencies which connect to and extend beyond school into the workplace and tertiary institutions (NZQA, n.d.-a; Wylie & Bonne, 2016). Therefore, a strength of the modular, credit based NCEA system is the ability to meet the needs of diverse learners with a flexible qualification. A weakness is that the focus on assessment for credits has resulted in practices that allow NCEA assessment, rather than curriculum objectives, to drive teaching and learning in senior secondary years (Absolum et al., 2009; Cowie et al., 2011; Cowie & Penney, 2015; East, 2014; Gillon & Stotter, 2012; Hipkins, 2015; Johnston et al., 2017; Thorpe, 2012). Other issues such as students gaining far in excess of credits required conflict with NZC advice against over-assessment and contribute to teacher workload issues (Ministry of Education, 2007a).

2.2 NCEA review and change principles

A wide-ranging review of the NCEA was undertaken in 2019 as part of a wider Kōrero Matauranga or Education Conversation about the future of education in New Zealand. The purpose of the review was to find out how the NCEA can be strengthened to "meet the needs of 21st century learners". A Ministerial Advisory Group with input from an NCEA Review Reference Group developed a Discussion Document that proposed six Big Opportunities as a starting point for public engagement. Students, parents, teachers, members of all education sectors and the wider public were engaged in the process (Darr, 2019; Kōrero Mātauranga, 2019).

The outcome of the review and engagement process was the announcement of the NCEA Change Package, a set of seven principles to guide changes to strengthen NCEA assessment (New Zealand Government, 2019). The changes are intended to contribute to delivering the Government's vision for an education system that equips learners with the knowledge, skills and capabilities needed to be successful in future education, employment and life in a global economy (New Zealand Government, n.d.) and a system that learns.

The seven changes are:

1. Make NCEA more accessible (end NCEA fees and ensure achievement standards are accessible for all, including students with disabilities or special learning needs).
2. Mana ōrite mō te mātauranga Māori (ensure there is equal status for mātauranga Māori in NCEA and greater opportunities for students to follow mātauranga Māori pathways).
3. Strengthen literacy and numeracy requirements (co-requisite requirement of 20 credit literacy and numeracy package which may be assessed whenever students are ready and as early as Year 7).
4. Fewer, larger achievement standards (standards will be rebuilt so that each covers a broader range of knowledge, with 50:50 split between internally and externally assessed credits).
5. Simplify NCEA's structure (remove 'carry-over' credits and make each level a 60-credit qualification. Maximum numbers of credits able to be entered set at 120 for Levels 1 and 2 and 100 for Level 3. Only allow resubmissions from a Not Achieved to an Achieved grade).

6. Show clearer pathways to further education and employment (create graduate profiles for each level of NCEA and develop a vocational entrance award).
7. Keep NCEA Level 1 as an optional level (Level 1 is the highest-level qualification for 10% of students. Keeping Level 1 as optional gives teachers the opportunity to innovate approaches to a broad education at Level 1) (New Zealand Government, 2019, pp. 6–16).

A Review of Achievement Standards (RAS) is currently underway. The RAS will redesign and develop achievement standards for each subject to align with the principles in the NCEA Change Package. Drafting, trialling and revision for Level 1 subject matrices, achievement standards, teaching and learning guides and assessment resources is in progress in 2020, with all new standards at Levels 1–3 fully implemented by 2025.

2.3 Digital assessment: NCEA online

The NZQA Future State portfolio sets out the innovations and outcomes needed to keep pace with changing modes of learning and changing workforce skills and knowledge bases in technology-rich environments. NZQA is moving NCEA external assessment, marking and the moderation of internal assessment as well as student Records of Achievement on the National Qualifications Framework (NQF) to digital and online formats (NZQA, 2018). The proposition is that paper-based, end-of-year external examinations do not take advantage of 21st century teaching and learning approaches and constrain students to be ready to demonstrate their learning at a specific point in time, irrespective of where they are on their learning journey.

Since 2014, NZQA has worked with schools to co-design and trial external digital assessments. In 2020, digital external NCEA examinations will be offered in 21 subjects. The following six principles guide the NZQA *Digital Assessment Vision for NCEA Online:*

1. Assessment integrity: students can authentically and securely show evidence of learning (knowledge, skills and abilities)—psychometric analysis of results. External standards assess a wide range of capabilities, on demand where possible.
2. Te Ao Māori: ensure equitable experience and outcomes for Māori learners—prioritise a move towards flexible personalised pathways—ability to use Reo and Mātauranga Māori in assessment, co-design/redesign assessments with Māori teachers and students.
3. Accessibility and usability: manage security and track learner progress e.g., digital dashboards—user experience—no one disadvantaged e.g., assistive technologies.
4. Adaptability: the assessment experience is adaptable enough for students to personalise their experience—timing of assessments is driven by student readiness—remotely supervised and 'open book' assessments.
5. Digital first: time and geographical constraints cease to be relevant, more continuous access to examiners and markers, human and computer marking.
6. Data as an asset—learning analytics inform assessment development and inform teaching and learning.

(NZQA, n.d.)

2.4 Criteria for summative assessment from the measurement literature

The design of educational assessments needs to be informed by the curriculum (including consideration of the features of the subject/disciplinary domain), the construct to be assessed, and theories of learning and pedagogy. This increases the validity of the interpretations made from an assessment, given that validity is a function not of the test itself but of the procedures and interpretations involved (Black & Wiliam, 2018; Messick, 1989; Perrenoud, 2006). Remembering that assessment is a process of collecting and using evidence to make different kinds of inferences and decisions about students' learning, "one validates, not a test, but an interpretation of data arising from a specified procedure" (Cronbach, 1971, p. 447, as cited in Black & Wiliam, 2018). The Standards for Educational and Psychological Testing (American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.), 2014) and The Classroom Assessment Standards; Klinger et al., 2015) provide a valuable set of guidelines and principles to ensure sound and defensible assessment practices, regardless of the format. With summative assessments related to online testing and testing when ready, some of these key criteria include:

- a clear description of the purposes of the assessment, along with a rationale for the format to be used and its fit to the intended purposes;
- an explicit description of the framework (norm-referenced or criterion-referenced) in which student achievement will be judged;
- sufficiently high levels of reliability (norm-referenced) or decision consistency (criterion-referenced) relative to the stakes of the assessment for students and/or others;
- a well-articulated assessment blueprint in which all forms of the assessment follow;
- multiple test form equivalency in terms of content, difficulty and scores (e.g., equating);
- procedures to ensure fair, accessible, and equitable administration;
- an administration procedure that is stable across multiple platforms and relevant contexts;
- evidence that the assessment and its components are not subject to construct irrelevant variance (e.g., Haladyna & Downing, 2004);
- current and ongoing evidence of the validity (accuracy) of the interpretations made regarding student achievement on the assessment;
- current and ongoing evidence that the resulting decisions, actions and consequences of the assessments are appropriate; and
- procedures to ensure and monitor test and administration security across multiple sites and times.

2.5 Principles for summative NCEA assessment

Taking into account criteria for summative assessment from the measurement literature, the current design of NCEA, the NCEA Change Package and current RAS, key principles for summative NCEA assessment are distilled below (e.g., Absolum et al., 2009; Hipkins & Cameron, 2018; New Zealand Ministry of Education, 2011; Norcini et al., 2011; van der Vleuten, 1996). These, along with the recommendations in the *NZ Digital Assessment Vision*, inform the discussion in this review.

- Assessment as part of pedagogy, contributing to deep, sustained learning.

- Students at the centre of teaching and learning, and learning and assessment as inextricably linked—students are active participants in the assessment of their learning.
- Assessment to be undertaken *when the student considers they are ready* to provide a true representation of their knowledge, skills and competencies to be demonstrated (validity).
- Achievement standards and assessment tasks should be of comparable demand as the NCEA makes no distinction between sources of contributing credits in award of the qualification.
- Coherence:
  - timing and type of assessment aligns with the purpose of the learning;
  - alignment of curriculum objectives, pedagogy and assessment task (NCEA assessments measure what a student knows or can do against the registered criteria of an achievement standard and as such, the achievement standards lie between curriculum objectives and assessment task); and
  - alignment with/leads to next steps for learners including tertiary or vocational training or workplace.
- Equity and inclusion—flexible assessments with opportunities to acknowledge/credential a wide variety of learning outcomes via multiple formats/modes of representation, including opportunities for students to follow mātauranga Māori pathways.
- Equitable pathways to improve the equity of outcomes for Māori and Pasifika students through richer data about student experience and different ways of capturing student responses.
- Fewer assessments so less stressful for students and teachers.
- Reliable/consistent assessment:
  - evidence collected leads to replicable/reproducible outcomes.
- Evidence collected leads to valid/fair inferences and decisions.
- All stakeholders have confidence in NCEA as a robust and flexible qualification.
- Security and integrity of all aspects of the assessment process is preserved.
- Manageable assessment—administration—workable—cost-effective outcomes easily reported and understood by stakeholders.

2.6 Terms and definitions

The following terms are used in this review:

Digital vs. digitised assessment: It is important to distinguish between different ways of conceptualising technology-enabled assessment. Digitisation is when existing procedures or examination formats are substituted for online or digital versions. Using the SAMR (substitute, augment, modify, redefine) framework, digitised assessment could be characterised as substitution and augmentation where there is a direct substitution with no or very little functional change (Puentedura, 2013). These are computerised tests that mirror paper-based tests and allow for similar response options.

Digital assessment: also, e-assessment, e-exams, electronic assessment, online assessment, computer-based assessment—under the SAMR 'modify and refine' framework involves significant innovation, allowing for the creation of previously unattainable formats or test items. Examples are on-demand,

anytime anywhere assessment, use of software to expand the scope of what can be assessed, or stealth assessment (Fluck, 2019; Masters, 2013).

High-stakes assessment: assessments that have substantial consequences for stakeholders, for example, used for the purposes of individual qualifications or selection, institute evaluation or reporting for accountability (Oppose et al., 2020).

On-demand assessment or assessment when ready: the availability of an assessment at the time (and preferably the place) when learning is complete or at students' or teachers' discretion (anytime/anywhere assessment).

Standardised assessment: assessment tools where procedures and scoring enable student results to be compared against results of a sample group/norm.

Summative assessment: an end-point event (internal or external) to demonstrate overall outcomes of learning in relation to curriculum progress and achievement criteria. Can be one-off, point-in-time or summarised over a period of time with several smaller events (van Groen & Eggen, 2020).

2.7 Structure of review

This review is set out in two parts as follows:

Sections 1 and **2** have presented the introduction and a summary of criteria for quality NCEA assessment which provide a framework for analysis/discussion. **Section 3** explains review methods. Overall findings of the research review are presented in **Section 4**. The focus is primarily on how the **timing** (Section 4.1) and **type** (Section 4.2) of summative assessment not only impacts teachers' and students' pursuit of learning but also impacts the reliability and decision consistency of such assessments and the validity of the interpretations made from the results. Section 4.3 is a **systems review of current practices** in other educational jurisdictions that can be considered comparable in terms of educational philosophy, broad educational structures and outcomes. Examples include Australia/South Australia, Canada and Finland.

Section 5 synthesises findings and sets out some implications. It incorporates insights from the review and feedback from NZQA and the Expert Review Panel. The discussion of the pros and cons of different models of externally assessed assessments takes into account the findings of this review, cross-cutting themes such as equity, workability, security/reliability, and collectively, the current design of NCEA, principles of the NCEA Change Package and current RAS, and key principles for summative NCEA assessment, with each aspect informing the other.

3. Review focus and methods

In this section we detail the processes used to search, select, and synthesise papers and reports. Stakeholder consultation took the form of initial meetings to understand the focus for the review and refine research scope and questions. Next, a comprehensive literature search was conducted, drawing on systematic techniques outlined by Hagen-Zanker and Mallett (2013) and Kable et al. (2012). Processes of mapping, appraising and synthesising research evidence from peer-reviewed academic journals and grey literature are explained below. An initial scoping review based on the research questions was used to refine search terms and compile appropriate search strings. The search was broad, encapsulating the main ideas specified in the research questions, and a large number of results from a number of relevant fields were identified. The research team then met to agree on inclusions and exclusions. The final search was restricted to papers published post-2010 to reflect the speed of change in technological innovation and digital learning/assessment applications in schools in recent years (Johnson et al., 2017).

A search of the following academic education databases was conducted:

- A+ Education (Informit),
- Education Database (Proquest),
- EBSCO—Academic Search Complete and Education Research Complete,
- ProQuest: Education Database and ERIC,
- Scopus,
- Taylor & Francis Education,
- NZCER database,
- ACEReSearch,
- ARC, and
- Web of Science.

Secondary sources such as literature reviews and meta-analyses were accessed where recent and combed for suggestions of relevant primary sources (n = 8).

Note: The key focus for this report was research studies conducted in New Zealand and internationally that address the nature and impact of assessment timing, type and testing when ready in the secondary school sector. However, research or evaluations of high-stakes summative digital assessment in secondary schools is an emerging, disparate field with a small number of countries and jurisdictions involved, and so searching uncovered a relatively small number of relevant results (n = 21). Hence the search was widened to incorporate studies from the wider body of literature in higher education where these studies could potentially inform the research.

Search strings were constructed using appropriate Boolean logic operators from combinations of four main categories as specified by the research questions:

1. Digital assessment ("digital assessment" OR eassessment OR e-assessment OR electr* assessment OR "computer-assisted testing" OR "computer-based testing").
2. High-stakes summative assessment (summative OR "high stakes" OR "summative evaluation" OR standardised OR practices).
3. Timing and type or format of testing (test item OR instrument* OR type OR format OR design OR tool OR "reporting tool") and (timing OR time delay OR "timing of assessment") and (frequency OR "assessment frequency" OR "frequency of assessment") and (test delay OR spacing gap OR spacing effect OR lag effect OR testing effects) and ("on demand" OR "when ready").
4. Secondary school ("senior school" OR "secondary school" OR "high school" OR "vocational school").

Searching continued until saturation was reached.

For the systems review, a search of the grey literature was undertaken. This included a scan of government testing agencies, technical reports, ministry of education websites and government-funded education research for Australia/South Australia, Canada (British Columbia, Alberta and Ontario) and Finland. In addition, a search for dissertation and thesis (masters and doctoral) research available on university websites was conducted to access recent studies that may not be in the public domain (for example: nzresearch.org, Trove, These Canada, and Theseus).

The full search resulted in 449 items based on screening of titles and abstracts. Closer screening of full text resulted in information or findings from 135 items being included in the final report. Items were organised and tagged in a citation management database (Zotero).

A wide range of study types were considered with the proviso being that findings informed the research questions rather than being restricted to certain methodologies or sample size, although some studies were excluded on the basis of quality and high rated journals were prioritised. Study types included qualitative case studies (e.g., see Broadbent et al., 2020; Buchan & Swann, 2007; Chian et al., 2019), mixed methods quasi-experimental studies (e.g., Parker et al., 2012; Pretorius et al., 2017; Walker & Handley, 2016) as well as quantitative and quasi- experimental work (e.g., Chang et al., 2012; Ling, 2016; Rezaei, 2015; Sulan et al., 2018; Weiner & Hurtz, 2017). Grey literature included reports (e.g., Boyle, 2010; Hillier, 2014) and government web pages (e.g., Government of Alberta, 2019; Ministry of Education and Culture, Finland, n.d.; SACE Board South Australia, n.d.).

The literature review and jurisdiction case studies have been used to create a synthesis of findings and implications for practice section. Our developmental process for this has aimed to ensure our work meets the needs of the client. Following this protocol, the initial scoping review has been shared simultaneously with the Expert Review Panel and members of NZQA who requested the work to discuss the contents of the review and the potential implications for development and implementation of the NCEA Online digital assessment programme. The Expert Review Panel was tasked with reviewing the paper to ensure: (a) the relevant literature has been cited, (b) the paper provides a sufficient review of the issues, and (c) the primary issues have been included in the review. Simultaneously the NZQA team commentary on key issues from their perspective. Section 5 takes account of this feedback.

4. Findings

The findings are presented in three sections. Section 4.1 explores themes from the review and analysis of literature focussed on the timing of summative assessment. Key questions pertaining to the timing of summative assessment were:

- If a working assumption is that summative assessment should contribute to the process of sustained deep learning, how does the timing (and type) of assessment impact on this?
- If a working assumption is that summative assessment is a core component in the measurement and awarding of qualifications, how does the timing (and type) of assessment impact on this?
- What does the research say about assessment when ready?
- What does the research say about assessment immediately after the learning has taken place vs after a break and a revision period?

Section 4.2 explores themes focussed on types of digital assessment, considering both tools and instrumentation. Lastly, section 4.3 examines what is occurring in other jurisdictions with regard to summative assessment for credentialing purposes.

4.1 Timing of summative assessment

Three themes were distinguished from an analysis of papers that included a focus on the timing of summative assessments, or *when* an assessment as a one-off event (internal or external) takes place in relation to the learning opportunity, with this, as might be expected, overlapping with the type of assessment. These were:

1. Flexible timing of assessment events:
   - on-demand and/or when ready, immediately after learning.

2. Fixed timing of assessment events:
   - fixed time once per school year,
   - multiple fixed times per school year e.g. mid and end year,
   - reassessment or resubmission opportunities.
3. The interaction between formative/summative intent:
   - a sequence of overlapping/integrated formative and summative tasks that on their own or as part of a suite that includes a concluding summative task contribute to a student's summative grade.

4.1.1 Flexible timing of the summative assessment event

This section reports on findings in literature associated with timing, or when summative events take place: on-demand or when ready, and following a break. Opportunities and challenges for the different types of assessment that can be used for this are discussed (studies on type of digital assessment are discussed in detail in 4.2). Impacts on student learning and individual responsiveness, school sector flexibility and assessment management are considered.

On-demand or when ready assessment

On-demand or when ready assessment is defined as the availability of assessment with a high degree of flexibility. In its most flexible form, on-demand assessment takes place at the time (and preferably the place) of students' or teachers' choosing (anytime/anywhere assessment). For example, teachers could opt to schedule a summative event at the conclusion of a module or students on individualised programmes could enter for an event once they feel prepared to do so (Bargh, 2011; Boyle, 2010, Wheadon et al., 2009). On-demand assessment is used widely in a formative sense and can be integrated with digital learning (see for example, Buchan & Swann, 2007; Deed et al., 2019; Friyatmi et al., 2010; Yamaguchi & Hall, 2017). Few empirical studies that demonstrated the implementation or impact of high-stakes, summative on-demand or when ready assessment in the secondary school sector were located in the wide-ranging review conducted for this report. Weiner and Hurtz (2017) have similarly noted that there is limited empirical or evaluative research in this space, while Tarricone and Newhouse (2016) claim that educational technologies, which might support this approach, tend to be "seriously underused" in high-stakes assessment (p. 7). This said, some review reports which consider the implications of technological advancements for digital assessment have made reference to possibilities and the inherent advantages of on-demand testing (e.g., Hipkins & Cameron, 2018; Wheadon, 2009). There was no evidence that other jurisdictions surveyed for this report offer on-demand, high-stakes summative assessments in the high school/secondary education sector. It was usual for large provinces/territories/states to offer multiple set windows for external examinations (see Section 4.3). These were more associated with alignment to school year semesters or holiday breaks than enabling on-demand events to be held immediately following learning as a means/approach that might advantage students through the timely demonstration of their learning.

The potential to deliver on-demand or when ready assessment events is perceived to be a necessary support for:

1. The increasing personalisation of learning, including individual learning progression pathways and responsiveness to test readiness.
2. Enhanced school and sector flexibility in curriculum decision-making and scheduling of testing.
3. Enhanced digital/digitised test administration and management, for example, ease of opportunity for test retakes.

(For example, see Boyle, 2010; Csapó & Molnár, 2019; Masters, 2013; Ripley, 2007; Timmis et al., 2016; van Lent, 2009; Wheadon et al., 2009).

Newhouse (2016) distinguishes between i) computer-based response exams where students respond to questions; ii) production exams, where students produce an artifact under timed and invigilated conditions; and iii) performance assessments where students perform or demonstrate a skill set. One form of an on-demand summative assessment event is a timed, supervised (remote or on-site) examination (Fluck, 2019). Implementing these typically requires the capacity to generate multiple unique assessments because the reuse of an assessment form can create security challenges. A common approach is to reuse items from a central test bank, with procedures to monitor the item for over-exposure (e.g., changing test statistics). Examinations of mathematics or the sciences require large item banks capable of producing randomised question sets of a comparable difficult with manageable levels of item re-use (Friyatmi et al., 2020; He, 2012; Wheadon et al., 2009). Integrating on-demand features with less flexible scheduled examinations would still require a large number of items or assessments.

Some studies have shown that computer adaptive testing (CAT) can be used to achieve on-demand assessments that respond to a candidate's ability level (e.g., Thompson, 2017). Computer-adaptive testing commonly employs item response theory (IRT) to effectively use an item bank to generate tailored assessments (e.g., Gierl et al., 2013; Zhang et al., 2019). IRT (modern test score theory) provides psychometric information about the items (e.g., difficulty, discrimination) which then uses the candidates' responses to provide ability estimates with a defined level of precision (van Groen & Eggen, 2020; Weiss, 2011). As a simple example of CAT, testing begins with a question of medium difficulty or a candidate-selected question level, with subsequent questions of greater or lesser difficulty selected from the item bank depending upon each preceding answer (Istiyono et al., 2020; Öz & Özturan, 2018). The test typically continues until a predefined level of precision is reached (Zhang et al., 2019). Information generated from CAT is often used for formative purposes, for example, the Australian NAPLAN (National Assessment Program—Literacy and Numeracy) and New Zealand e-assTTle (Assessment Tools for Teaching and Learning) (Brown, 2019; Thompson, 2017).

One disadvantage of CAT is that students are not able to return to questions to correct them (Kimura, 2017). Additionally, examination security and integrity are key concerns (Fluck, 2019). It is possible to argue that e-Exams including CAT can improve academic integrity, as each examination paper consists of a unique set of questions and individual tests can be of different lengths (Adegbija et al., 2012). The issues of item leakage also exist in CAT environments and these can impact the adaptive nature of the assessment. Further, high-quality discriminating items tend to be overused, resulting in a gradual decline in the quality of the usable items in the bank (Liu et al., 2019).

Test banks are expensive to develop and maintain (Zhang et al., 2019). Balancing test accuracy, efficiency and length with the difficulty threshold in CAT can be challenging, as at 50% probability of answering correctly (which maximises test information and minimises items to be covered) (Kimura, 2017, p. 2), candidates can lose motivation and their perceived lack of achievement can affect their self-efficacy. Computer adaptive multistage testing where candidates receive test sections or sets of items can be a useful strategy for achieving optimum test length and difficulty threshold while allowing opportunity to review answers before moving on (Becker & Bergstrom, 2013; Kimura, 2017; Masters, 2017). This approach therefore has the potential to optimise motivation and demonstration of learning. The use of testlets has also become a more common approach in a variety of online testing formats. The item formats within CAT must be amenable to computer-based administration and immediate scoring, raising questions as to the scope and nature of what can be assessed in this way and how this aligns with what the curriculum values. As a result, CAT, while a continued area of research interest, has not been widely adopted for use in secondary education high-stakes testing.

CAT has had more uptake in relation to formative assessment. Online tutoring software augmented by on-demand computer-adaptive or game-based formative assessment tools enable learners to seek and use feedback within self-assessment of learning (Buchan & Swann, 2007; Yamaguchi & Hall, 2017).

If combined with summative, on-demand assessment when ready, this approach presents the potential for continuous and integrated approaches to the formative and summative purposes of assessment in support of learning (Hipkins & Cameron, 2018; Timms, 2017) (and see Section 4.1.3).

Returning to the Newhouse (2016) category of performance assessments, the ability to enter for summative performance-type assessments 'when ready' is common in professional accreditation or certification (Weiner & Hurtz, 2017). Examples are readiness for licence to practice within industry sectors and driver licence exams (Murgatroyd, 2018a; Houston & Thompson, 2017; Price et al., 2018). A form of performance assessment in the school sector is the oral defence of a project or paper, the results of which Guha et al. (2018) propose can be used to demonstrate development of the skills and competencies needed for tertiary education or entry into the workforce. Guha et al. (2008) suggest that digital portfolios can be used to capture students' performance more effectively than standardised tests for a range of knowledge and skills and can better reflect the achievements of priority or underserved learners. Digital or ePortfolios can be used in assessment of project-based or problem-based learning where evidence is captured cumulatively and assessed against standards when students and/or teachers deem it complete/ready (Boyle, 2010). Digital or web-based self-assessment tools and formative feedback from teachers can assist students in the compilation of such supporting evidence (Binkley et al., 2012) (and see Section 4.2.4).

There are a number of implications for on-demand and when ready assessment arising from this analysis.

- Depending on the curriculum, and an assessment programme design that includes a number of formative or practice events, on-demand assessment could increase or decrease the overall time spent on assessment (for example, five summative events of between 30–60 minutes across a year in which students complete a number of practice tests for each, versus one examination of 180 minutes that comprehensively tests the taught curriculum).
- Assessment systems and schools need an infrastructure to support the flexibility required to cope with on-demand or when ready assessment. For example, an assessment system must work on a variety of digital platforms, both in terms of hardware and software. Schools must be able to provide suitable seating/space for greater or fewer numbers of students accessing online assessment at any given time, and provide staffing to monitor and moderate assessments at multiple time points.
- Key related questions for implementation of when ready assessment are: What is "readiness", Who makes the decision that the student is "ready", and What are possible unintended consequences for students, teachers and schools of being deemed "ready" or not? Is a student ready when they demonstrate a specified level of "competence" or meet a performance standard, or is readiness achieved when competence or achievement exceeds a performance standard? In all cases, an informed answer is likely to rely on multiple milestones or formative events (Binkley et al., 2012; Boyle, 2010, Guha et al., 2018; Houston & Thompson, 2017). Wesolowski (2014) proposes that decisions be tied to the use of structured rubrics that track consolidation of skills or capabilities and thus progress towards the goal of readiness for summative assessment. How schools and teachers might manage teaching for a mix of students and student groupings in terms of progress towards readiness, being ready and beyond is another key question for implementation.
- On-demand and when ready assessments have the potential to offer learners greater ownership of their progress in learning (Irwin & Hepplestone, 2012; Pretorius et al., 2017; Rideout, 2018). However, as noted by Wheadon et al. (2009), on-demand systems where students progress at varying rates when gaining high-stakes credentials can aggravate unhelpful competitive aspects such as parental pressure or peer-to-peer comparison. This may feed negatively into wider community discourse linked to school ratings and comparisons. Further, there is potential for

these approaches to perpetuate, or even exacerbate, already entrenched inequalities in progress and achievement, with some student groups more or less likely to be able to access resources or out of school support which might help them progress more quickly (Leadbeater, 2006) (and see Section 4.2.2).

- The facility for on-demand assessment immediately following a learning module could reinforce issues with 'washback' in which the curriculum is essentially narrowed to focus only on what is/will likely be assessed (e.g., Absolum et al., 2009; Fensham & Rennie, 2013; Gillon & Stotter, 2012; Johnston et al., 2017; von Heyking, 2019; Wallace & Priestley, 2017). In this case, the focus might turn to improving test performance rather than the curriculum and learning. There is some evidence of this in relation to NCEA internal assessment as the formative tasks essentially mirror the summative task, undermining both purposes (East, 2014; Hipkins, 2015; Hume & Coll, 2010; Moeed, 2010).

- eExams can preserve academic integrity as digital answer papers are able to be more easily checked using anti-plagiarism software (Fluck, 2019). Yet there are several security risks with online or digital exams. For example, Dawson (2016) lists five possible methods of cheating and mitigation strategies in online exams with bring-your-own-device (BYOD) administration (p. 596). However, based on findings of a recent international review of eExam technologies, Fluck (2019) and Sindre and Vegendla (2015) argue that overall, eExams are no more or less secure than paper-based exams but that the level of security depends on the type of exam and the protections that are in place.

- On-demand testing requires policies and procedures that assure all stakeholders of security, reliability and fairness of assessments; however, it needs to be noted that strengthening any one of these aspects does not necessarily impact positively on the others. As one example, higher stakes examinations typically require higher levels of security, standardisation, and reliability; yet these can add to increased anxiety and pressure for some students, reducing the accuracy of the results. For high-stakes, examination-based assessment, on-demand testing is dependent on provision of requirements for safeguards such as test security and proctoring (supervision, invigilation) to maintain integrity in online and on-demand assessment environments. Some experimental research has investigated concerns that unsupervised remote or online testing may result in higher incidences of cheating (Ladyshewsky, 2015; Michael & Williams, 2013). Weiner and Hurtz (2019) describe a quasi-experimental study in which online professional licensing exams were conducted either onsite or by remote. Their findings suggested that remote online examinations can mimic a traditional testing environment, depending upon setup and controls. Test content can be delivered to candidates using their own computers and a video recording can capture their engagement with the test via the computer web camera. The test can then be completed without live supervision. Another option is to have an online remote supervisor who interacts with the candidate in real time. Yet another option is live remote proctoring at designated sites which candidates attend (Weiner & Hurtz, 2017). Irrespective of the option chosen for remotely invigilated online exams to be successful, real time technical support must be available and candidates ideally would have the opportunity to become familiar with the exam format/instrumentation prior to working on any tasks (Cramp & Medlin, 2017). Importantly, remotely invigilated online exams demand careful, systematic design, preparation and piloting for successful implementation (Cramp & Medlin, 2019).

- Online on-demand testing can require different systems for marking; computer-marked or person-marked/moderated, or both. Automated scoring is commonly used for closed or multiple-choice questions/quantitative answers and short simple answers (Murchan & Oldham, 2017; Reinersten, 2018). Systems have been developed for computer-marking of more complex writing tasks, but the costs prohibit such systems in many jurisdictions. As an alternative, assessment tasks that require rich responses such as open-ended answers or extended writing

can use distributed marking models (Stödberg, 2012; Vista et al., 2015). For example, identification and highlighting of structural criteria such as sentence length and paragraph complexity, readability, use of critical keywords, are automated but the final judgement as to whether these indicators point to particular skill levels is still done by humans (Vista et al., 2015).

- Baird et al. (2017) evaluated 'rater accuracy' using data from high-stakes tests examinations in England. Their data was drawn from the training and monitoring of 576 raters in 110 teams, across 22 examinations. The authors found variations in rater accuracy (consistency) and propose that while face-to-face training and supervisor-based monitoring of raters is the norm, online training and expert-based monitoring has the potential to produce more accurate depictions of rater judgement effects. The authors also point out that two-thirds of marking for public examinations in England is now on-screen and suggest that the advent of marking technology holds the promise of more accurate and efficient rating by markers.

4.1.2 Fixed timing of the summative assessment event

As noted in Section 4.3, it is common in larger education jurisdictions such as those in Canada and Australia to offer fixed assessment timing but multiple set windows for external examinations per school year. Also as noted, although multiple fixed windows in effect offer examinees some flexibility in when they can complete an assessment, this scheduling system is more to do with practicalities of aligning with different school semester cycles and managing student numbers in online environments than optimising students' demonstration of learning. The spacing between examination windows and the frequency or number of events offered, including reassessment opportunities, are two aspects to be considered in connection to fixed timing.

Assessment following a break or revision period

The timing of summative assessment according to fixed yearly or multiple set cycles often results in a break between an initial learning period and summative testing. This is not necessarily detrimental to students' learning or the demonstration of their learning. The findings of some studies highlight the opposite effect, with wider spacing between testing offering benefits for student achievement and long-term learning, especially when used in conjunction with formative events and revision in the lead-up to the examination. This "spacing effect" has been found to increase long term retention of information.

Vlach and Sandhofer (2012) investigating spaced lessons in primary science and Chen et al. (2018) investigating spaced primary mathematics lessons and Bird (2010) working with English-learning adults found the spacing of lessons led to better results. Other studies in a range of disciplines have shown that test performance is improved if learning and revision sessions are spaced, but that optimal spacing varies according to the type of learning. The spacing effect has a positive impact on achievement over a range of curriculum objectives, including for learning that requires memorisation of declarative knowledge, is conceptually difficult, demands problem-solving skills, procedural know-how or the development of motor skills (Carpenter et al., 2012; Moulton et al., 2006; Rohrer & Taylor, 2007; Sobel et al., 2011).

Student preferences for timing of assessments in a first-year undergraduate business programme was the focus of an investigation by McManus (2016). Students (n = 263) were asked to indicate preferences for when they took a first assessment worth 20% of their final grade: shortly after completion of the learning module, after a gap of eight weeks or no preference. Each examination was unique but of equal complexity and students were assured there would be no knowledge or teaching advantage for those who chose the later assessment. Fifty two percent of students opted for the early testing option while 42% opted for a testing delay. The remaining students indicated that they had no preference (2%) or did not make a decision (5%). A significant number of students highlighted conflicts such as the desire to

get the exam out of the way or having extra time to review content and master conceptual knowledge. Students who took the later option performed more poorly than those in the earlier group, however, there was no statistically significant relationship between timing of the assessment and achievement if influences attributed to poorer attendance by the later testing group were controlled for. Students who opted for the earlier test went on to perform better in subsequent assessments. It was hypothesised that this group was "more comfortable and engaged" in their learning (McManus, 2016, p. 214).

The impact of assessment and assessment frequency on learning and achievement

The positive impacts of testing and increased testing frequency on student learning and achievement are well established in literature (Figueroa-Cañas & Sancho-Vinuesa, 2018; McDaniel et al., 2011). However, there are other factors to consider, such as the effects for teachers and students of added stress and workload, and reduced time for teaching and learning. Variables such as subject area and stakes (high or low stakes, formative or summative role) likewise need to be taken into account when investigating the effectiveness of testing frequency and assessment type.

The impact of testing on improved knowledge recall can be partly explained by the 'generation effect' as described by Bertsch et al., (2007). Bertsch et al. conducted a large meta-analysis of 445 effect sizes from 86 studies to conclude that students who actively engage with or interact with learning materials to generate new information (for example, produce synonyms for words) or practice mathematics problems have better recall than if they simply read a textbook or listened to a lecture. It is thought that the associated effort or urgency associated with testing leads to improved or enduring retention of information (Beagley & Capaldi, 2016; Phelps, 2012). For example, Angus and Watson (2009), reporting on the introduction of three-weekly online quizzes (worth 2% each of the final course mark) in a first-year university applied mathematics course claimed that more frequent exposure to online testing led to increased student achievement. This result was achieved after controlling for other factors such as gender, effort and previously assessed aptitude.

A meta-analysis on the effect of testing on student achievement conducted by Phelps (2012) included 177 quantitative studies with a variety of study designs, conducted between 1910 and 2010. Results were that testing with feedback produced the strongest positive effect on achievement, however, adding stakes or higher testing frequency was also found to strongly and positively impact achievement. Studies in which the treatment group were tested more frequently than the control group produced a mean $d \approx 0.85$. Studies in which the treatment group was tested with higher stakes than the control group produced a mean $d \approx 0.87$. Studies in which the treatment group was made aware of their test performance or course progress (and the control group was not) produced a mean $d \approx 0.98$. Studies in which the treatment group received some other type of remediation or feedback based on their test performance produced a mean $d \approx 0.96$ (p. 34). Phelps explained that an optimal intervention would employ a number of these strategies at the same time.

Başol and Johanson (2009) conducted a comprehensive review and meta-analysis of 78 studies to calculate an overall effect size for frequency of testing over achievement in United States secondary schools and universities. Studies with similar testing frequencies were classified as low (every other week or less), medium (weekly) or high (more than once weekly). An aggregated mean effect size was calculated. Overall, the findings indicated that frequent testing increases academic achievement with a cumulative mean effect size g=0.41, i.e., assuming normal distribution, a student receiving frequent testing achieved at a higher level than 66% of students who were not frequently tested. Interestingly, overall effect sizes did not vary significantly between lower, medium and higher frequencies, which raises the question of determining optimal testing frequency. When analysing effect sizes for separate independent variables, there was no significant difference between the studies using frequent graded tests and frequent un-graded tests, contradicting the hypothesis that frequent summative testing would have a greater positive impact on achievement than frequent formative testing. The studies were also

found to differ in their effect sizes according to subject matter, with mathematics having the largest mean effect size.

Bergmann (2014) focussed on two questions: why some nations require students to take large-scale standardised tests more frequently than others, and whether there is a correlation between the frequency of large-scale standardised testing and academic achievement. Standardised testing was defined as tests that were designed and scored externally, uniformly administered and scored, typically given to large groups at once, and where results are used for a variety of reasons (p. 17). Data was collected from responses of 40 sample nations to OECD school questionnaires about testing frequency for the 2003 and 2009 administrations of the Program for International Student Assessment (PISA). Results from this study indicated that the frequency of large-scale standardised tests is most likely to be associated with testing consequence or stake and that the frequency of large-scale standardised tests is not significantly related to academic achievement. Arguing that no current theories exist for why some jurisdictions test more frequently than others, Bergmann identified four key variables from a correlation analysis of what might be associated with a high frequency of testing: (a) data is used to compare the school to the national performance, (b) data is used to compare the school to other schools, (c) achievement data is posted publicly (e.g. in the media), and (d) achievement data is used in the evaluation of the principal (p. 45). Bergmann suggested that in the States these findings demonstrate the impact of policy requirements such as No Child Left Behind. It is interesting that none of these variables relate to enhanced student learning or to increased validity of the assessment scores.

Sulan et al., (2018) investigated the psychological and achievement effects of testing frequency in a range of university music performance and theory courses in Cyprus. The test group was 59 students in their first to third year of study. A previous pattern of one mid-term exam and one final exam within a semester (1–1) was changed to include two mid-term exams per semester (2–1). It was assumed that a positive relationship exists between testing frequency and academic achievement, especially for students who are exam and achievement oriented. While a number of variables such as self-confidence, tendencies to perfectionism, prior training and level of accomplishment influenced performance results, overall average grades for both performance and theoretical exams in the 2–1 structure were higher than for the 1–1 format. However, the extra assessment increased performance anxiety to the extent that the initiative was abandoned after one semester. This was especially significant for female students. Likewise, Vaessen et al. (2017) found that frequent assessments increased stress for students enrolled in a compulsory first-year statistics course (n = 219) but students found the assessments to be motivating for study and revision purposes.

A quasi-experimental study by Rezaei (2015) also found that student achievement improves with more frequent testing. A total of 288 California State University students taking quantitative research methods courses were followed over a period of five years. The goal was to determine if there were improvements in students' conceptual learning using a weekly quiz which was graded for summative purposes when compared to using mid-term and final examinations only. Frequent testing was associated with increased motivation for learning and better understanding of factual knowledge as demonstrated by progress in quizzes. Higher order critical thinking and problem-solving skills also showed improvement. These results were reflected both in the final examination and in a final project which involved developing a quantitative research proposal. An interesting finding was that frequent testing reduced individual differences between initial and final test scores. Rezaei thus suggested that frequent testing might act as an equaliser in overall student achievement. Another goal was to compare students' performance in collaborative group quizzes with individual quiz scores. Open-book collaborative weekly quizzes resulted in better test scores, less test anxiety, and greater general student satisfaction. The results also showed that when quizzes are open-book and students have a chance to collaborate (discuss how to answer the quiz questions), they perform significantly better in their final examinations and final projects. Rezaei argued that this finding is an indicator of enhanced conceptual learning.

The frequency of high-stakes assessment events can also be considered in terms of impacts on the time available for teaching and wider student learning, and on teacher workload (Polesel et al., 2014). In the context of English A levels, the introduction of resits in 2000 was flagged by Scott (2012) as a reason for a significant improvement in results between 2001 and 2009. Scott interviewed managers, university admissions tutors, teachers and students, and administered a student questionnaire (n = 267) to explore implications for student learning of resit policies in relation to high-stakes exams. At the time of the study, A level examinations used a modular format where students were permitted to re-sit units within a course. Resit rates were high with 74% of Year 13 students resitting at least one unit in one or two subjects. While acknowledging that resit opportunities avoided the issue of students being disadvantaged in a one-off exam situation, Scott argued that allowing unlimited resits produces undesirable consequences for learning. These included students being overly relaxed about exam preparation and the time dedicated to teaching for resits disrupting student learning going forward and adding to teacher workload. Overall, undesirable practices included a focus on surface learning for extrinsic rewards rather than on deep learning. Scott concluded that "due to the high stakes nature of A levels the resit system has given rise to some questionable practices by colleges and students as well as undesirable effects on teaching and learning, rendering it unfit for the purposes of the exam" (p. 447). NB: from 2017, A Levels assessment changed with students now permitted only one resit in the year following their last scheduled examination sitting.

4.1.3 Multiple or spaced summative assessments with formative feedback

Many recent studies were located that suggest or advocate for assessment models that employ a sequence of overlapping/integrated formative and summative tasks with dual goals of strengthening student learning and contributing to an overall summative grade (Barana et al., 2019; Day et al., 2018; Maclean & McKeown, 2013). A number of these studies used the affordances of digital tools to capture, curate and evaluate data on student learning.

In the context of higher education, Broadbent et al. (2018) report on case study research which demonstrates how formative elements can be integrated with summative assessment in a large first-year psychology course at an Australian university. Students studied fully online (n = approx. 300) as well as using blended learning models (n = approx. 1200) across different campuses. Over an 11-week semester, assessment consisted of ten online quizzes of ten multi-choice questions each (10% of final summative grade), a 100-question multi-choice examination (45%) and three journal tasks (45%). Broadbent et al. argue that for summative assessment to benefit learners it should be integrated with formative elements such as personalised feedback, however, they acknowledge that multiple assessment tasks can mean issues for teacher workload. Formative design features intended to build student capacity to make judgements about their own learning and improve future performance included online availability of exemplars and rubrics and time-efficient personalised audio rather than written feedback. Further to this, in a large course with many tutors and markers it was vital that shared assessment and marking procedures supported high inter-rater reliability to ensure fairness and avoid student complaints related to perceptions of inequitable workloads or grading practices. To ensure consistency in feedback messages and grading, tutors and markers participated in rigorous training in the provision of feedback, grading and moderation. The authors argue that the case study demonstrates a "proof-of-concept" design for a large multi-mode, multi campus course that features high-quality formative feedback and "defensible summative grading" (p. 320).

The processes involved in the design of an integrated formative summative assessment system are also described in an illustrative case study by Chian et al. (2019). The achievement of constructive alignment of assessment tasks with year-level and program-level learning outcomes was seen as an important indicator of valid and reliable assessment. Assessment tasks were aligned with learning outcomes in a problem-based learning (PBL) cycle in a first-year undergraduate dental programme. The assessment was an adaptation of a Triple Jump Assessment (TJA), an approach used in PBL which evaluates both

knowledge and process over three distinct stages or jumps (McTiernan et al., 2007). The TJA concept designed by Chian et al. first involved assessment of independently written individual responses to an unfamiliar problem scenario under formal, timed examination conditions and individual contributions to an examiner-facilitated group discussion. This first stage was intended to assess students' ability to explore and discuss problems and generate ideas. The second stage involved individual and collaborative research and was intended to assess students' ability to research an issue and discuss current knowledge and learning resources. The final 'jump' was intended to assess students' problem-solving abilities in applying new learning and sharing new knowledge. This was assessed in the form of an individual written examination and an oral 'viva' in front of two examiners. The insights gained from the project facilitated the identification of four core principles and design elements for valid and reliable TJA: (a) viewing the assessment design process as a collaborative and collective faculty endeavour, (b) recognising the assessment design process as dependent on shared understandings of learning, (c) highlighting the centrality of ongoing review, and (d) prioritising student learning in the development of the TJA as an assessment system.

Houston and Thompson (2017) argue that the context of paramedic education provides a clear example of the challenges of balancing assessment purposes. This is because graduating paramedics need to be certified as competent and "road-ready" as well as able to function as critically reflective practitioners who are focussed on learning and performance improvement. They describe the recalibration of assessment in a capstone (final-year, final-semester) project, with dual goals of integrating formative and summative purposes and repositioning assessment as a communication process about learning. The previous summative-only system was seen as inhibiting teaching and learning in that it promoted "grade-seeking" behaviours at the expense of continuous, and personalised, learning and self-assessment. Neither did it meet the needs of a student population that was diverse in terms of understanding and mastery of curriculum, with differing levels of maturity and life experience (p. 4). Assessed learning activities focussed on knowledge and application and on practical skills. Two diagnostic and formative multi-choice examinations were held at the beginning and mid-point of the semester. These were intended to highlight initial gaps in knowledge and understanding. In a subsequent formative/summative cycle, students engaged in problem-based learning and were required to research knowledge gaps and contribute findings to a wiki site, where class members could collaboratively share and edit information. Students were summatively assessed on their contributions to the problem scenario and wiki. Learning materials that students contributed to the wiki were featured in the mid-semester multi-choice examination, meaning that students effectively contributed to the design of this assessment. The development of a student-tutor consensus marking approach to a final assessment was designed to "encourage students to apply a paramedic lens to critique their own work (p. 7). Half of the final grade was derived from a performance assessment where a tutor made a judgement against set criteria. This was withheld from the student while they critiqued and graded their own performance. Any disagreement was resolved by a clarifying conversation. A final oral assessment was linked to individual learning needs indicated by the formative multi-choice exams, meaning that students were able to pay attention to specific areas in their preparation. The formative focus continued even with this assessment as students participated in a post-oral exit interview where they were given advice for their ongoing development. Assessment in this integrated formative/summative design was seen as a two-way stream of communication and on-going dialogue (p. 11) and was used on an individualised basis for benchmarking as well as signalling needs going forward.

New Zealand tertiary teachers who collaborated to review and develop their online assessment practices identified that effective online assessment offered a range of benefits to both students and teachers including: more interactive assessment and opportunities for formative feedback, increased efficiency and reduced workload, the ability to meet the needs of increasingly diverse learners, and the opportunity to use new technical and pedagogical skills (Terrell, 2016, p. 4). The team developed guidelines and an online assessment tool selector (OATS) for use in planning and selecting appropriate assessments that are aligned to assessment purpose and programme of learning. Twelve case studies of tertiary teachers'

use of online assessment tools detail the use of a range of online tools including ePortfolios in Google, Moodle and WordPress, blogs, forums, gaming, quizzes, videos and wikis.

It is of note that the previous examples are all from tertiary settings and while some involved comparatively large numbers of students, the contexts and student numbers did not approach those that would be involved in a nationally organised sequence of formative and summative assessments in secondary schools. It is possible that the affordances of digital technologies might allow for using some of the strategies outlined. Shute and Rahimi (2017), based on their review of research on the effective use of computer-based assessment for learning for primary and secondary education (Kindergarten–grade 12), claim that advancements in computer-based assessment could eventually blur the boundaries between teaching, learning and assessment to the point where assessment is continuous and formative, obviating the need for high-stakes assessment. They argue that ongoing, data-driven computer-based assessment "holds great promise for creating high quality and personalised learning experiences" (p. 15). Stealth assessment, for example, can be used to analyse the data generated by students' interactions with technology in game-based learning contexts. Shute and Rahimi state that assessment quality (validity, reliability and efficiency) are not impacted by the integration of formative and summative aims.

How does formative quiz performance link to final summative assessment achievement?

Linking back the advantages of increased testing frequency, regular quizzes were found to contribute positively to students' summative grades, although a positive impact could not be taken for granted. Marden et al. (2013) found that students who performed poorly in quizzes were more likely to fail the examination, suggesting that formative online quizzes may be a useful tool to identify students in need of assistance. In contrast, Gokcora and DePaulo (2018) found formative quizzes improved student performance in summative end-of-year examination-based assessments. Kennedy and Fiesta (2020) explored the impact of different formats of formative quiz performances (in-class versus online; different time limitations and ability to use notes) on student summative performance. Students performed significantly better on the summative examination if they had completed formative quizzes with rigid time restrictions before the summative event. Untimed quizzes, delivered online, led to increased formative performance but decreased performance on summative examinations.

Rezaei (2015) found that when quizzes were open-book and students had a chance to collaborate (discuss in pairs how to answer the quiz questions), they performed significantly better in their final examinations and their projects. Rezaei argues that while the literature indicates many students and teachers believe that frequent testing is stressful for students, their study shows that if the test is open-book and if students are able to collaborate on the quiz, this leads to "positive, meaningful, and sustainable effects" (p. 195).

4.1.4 Section summary

This section provides a short summary of key points from the papers reviewed in this section in terms of timing (flexible on-demand when ready or fixed), frequency and the interaction of formative and summative tasks.

On-demand and/or when ready assessments:

- offer potential for personalised learning pathways and student ownership,
- offer potential for flexibility in school administration and management,
- require structures/processes for determining when a student is ready to be assessed, and
- can increase or decrease the overall time spent on assessment depending on design/format.

For on-demand and/or when ready digital assessment:

- Remote exams are possible, with remote supervision systems available.
- eExams are no more or less secure than paper-based exams but security depends on the type of exam and the protections that are in place.
- Different systems for marking are required depending on assessment type; computer-marked or person-marked/moderated or both.
- No evidence was found of on-demand when ready high-stakes national summative assessment systems currently employed in the high school/secondary education sector.

For computer-adaptive testing (CAT):

- Computer adaptive of game-based and stealth assessment tools present potential for continuous and integrated formative/summative approaches are compatible with on-demand/when ready assessment.
- On-demand CAT examinations require large item banks.
- Multiple exam formats from CAT can be tailored to each individual, to multiple testlets, to automated multiple test form assembly.
- Security in CAT is a prime concern, especially in terms of item exposure.
- CAT can be used to increase test validity, but students are not able to return to questions to correct them - multistage CAT or embedded testlets offers a way around this issue.
- On-demand and CAT can involve high costs in development and maintenance of item banks.

Frequency of assessment:

- Large provinces/territories/states offer multiple set windows for external examinations (see Section 4.3) but these are more associated with alignment to school year semester or holiday breaks.
- Frequent assessments can have a positive impact on student learning, motivation and achievement but this can depend upon the subject area and type and level of assessment.
- For low-scoring students, frequent testing may act as an equaliser in overall student achievement.
- Formative review quizzes with time restrictions were found to have a significant positive effect on summative examination achievement when compared to quizzes with less rigid restrictions.
- Open-book, collaborative quizzes reduced stress and had positive impacts on student achievement in the final examination.
- Frequent assessment can be associated with negative impacts such as increased anxiety and stress for students, increased teacher workload, less time available for teaching and wider student learning, and a focus on teaching to the test.
- There is a learning advantage over a range of curriculum objectives (knowledge recall, procedural and conceptual knowledge, motor skills and problem-solving) associated with cumulative and repeated exposure to learning materials leading to a summative event, for example, after spaced learning or a testing delay.

Overlapping formative/summative goals:

- For summative assessment to benefit learners it should be integrated with formative elements such as personalised feedback, with design features intended to build student ownership and autonomy in seeking to improve achievement, however, this can create issues for teacher workload.
- Advancements in computer-based assessment could eventually blur the boundaries between teaching, learning and assessment to the point where assessment is continuous, formative, and data-driven, with potential to create high quality and personalised learning experiences.

4.2 Type of assessment

The selection of an appropriate assessment tool is a complex process. As noted in Section 2.4, to make valid inferences about student learning, assessment must align with curriculum objectives and consider the nature and pedagogy of different disciplines and learning areas. The purposes of an assessment and use/s to which the reported results are put (for example, grading and programme completion, admissions and selection to further tertiary study, employment or certification) are also critical to validity. In considering factors related to equity and inclusion, assessment pathways need to be appropriate for all learners. Ideally, assessment practices are student-centred and position students as active participants in their learning (see Section 2.5).

Decisions about overall assessment design and instrumentation necessarily encompass the following aspects:

- mode of capture and representation of evidence, for example, digital or paper-based assessment, oral or written, examination, performance, portfolio;
- conditions of assessment including timing and supervision, external or internal, open or closed book, collaborative or individual;
- assessment administration and security including supervision/invigilation, distribution and submission processes; and
- marking processes including achievement criteria, and marking rubric or schedule, and moderation.

This section explores themes from the review and analysis of literature focussed on types of digital assessment, considering both tools and instrumentation. Firstly, we outline issues associated with concepts of reliability and validity related to assessment type. We consider matters of equity and inclusion in digital assessment for diverse student populations. Next, modes of evidence capture are discussed in relation to computer-based vs. paper-based examinations, computer adaptive testing, and educational data mining including issues connected to test administration and security (see also Sections 4.1.1 and 4.4). Lastly, types of assessment that involve multiple forms of representation such as ePortfolios and performance assessment, along with implications for marking and moderation, are discussed (see also Section 4.2.4).

There is a large body of literature that discusses advancements and innovations in e-assessment in higher education and within large testing companies (e.g., ACT, ETS, Pearson). We have drawn on findings of these studies to inform this discussion of digital assessment as fewer studies were located that focussed on the secondary school sector.

4.2.1 Reliability and validity in summative assessment

Reliability is a measure of replicability, consistency and fairness in assessment. That is, the extent to which an assessment is free from bias and consistently measures learning (Chian et al., 2019; Harlen,

2005; Ministry of Education, n.d.) such that "if the assessment were to be repeated, the second result would agree with the first" (Harlen, 2000, p. 111, as cited in Harlen, 2005). The validity of an assessment tool is the extent to which the evidence produced supports the making of valid or accurate inferences. There are many forms of validity including consequential validity (the consequences for learners and teachers) and criterion validity (the criteria for judging the performance of a learner) (Harlen, 2005; Newhouse, 2016). Messick (1989) and James (1998, as cited in Harlen, 2005), however, argue for construct validity as an overarching or unifying concept. According to Messick, this unified view of construct validity "comprehends both the scientific and ethical underpinnings of test interpretation and use" by integrating "considerations of content, criteria, and consequences" (1989, p. 5).

In sum, a valid assessment measures what it was designed to measure (Dembitzer et al., 2017; Harlen, 2005; Ministry of Education, n.d.) and results in defensible and accurate interpretations for the intended purposes. When collecting evidence to be used in making summative judgements, decisions about the type of assessment need to take into account the tensions between goals of achieving optimum reliability and preserving validity (Harlen, 2005).

Advances in digital technologies present opportunities for more automation and scope for conducting flexible, personalised "anywhere, anytime" summative-type events (Masters, 2013, p. 27). Likewise, digital contexts improve the ability to deliver "pedagogically rich assessment environments" which assess higher order thinking and information processing skills (Fluck, 2019, p. 5), potentially increasing construct validity. Similarly, capturing multiple forms of digital evidence from a performance or ePortfolio arguably allows learners to better show what they know and can do. However, this increased scope can complicate assessment, with more factors to be separately assessed or requiring overall or holistic judgements (Newhouse, 2016). There can be more potential for errors or bias, resulting in reduced reliability (Harlen, 2005). On the other hand, increasing reliability by tightening the range or methods of evidence collection (for example, from assessment of project-based learning to assessment by examination) can decrease validity. A further point to consider is the mode of marking or scoring. There are different implications for dependability/consistency in marking and moderation for national examinations and computer marking than standards-based or criterion-referenced internal/teacher summative judgements (Harlen, 2005; Smaill, 2013).

Thus it is necessary to pay attention to the conditions that differentially impact manageability, reliability and validity when considering assessment type. A high-stakes assessment system designed to offer flexibility and negotiate these tensions typically consists of a combination of internally and externally assessed tasks (see Sections 2.1 and 4.3).

4.2.2 Equity and inclusion in digital assessment

Important factors when considering equity and inclusion in digital assessment are test accessibility and usability (Fluck, 2019; Scalise et al., 2018). Begnum and Foss-Pedersen (2018) conducted an exploratory case study of universal learning design quality of two major digital assessment solutions in the Norwegian higher education system. Adherence to principles of universal design means that digital assessment systems do not rely on "one-size-fits-all solution(s) but rather flexible approaches that can be customised and adjusted for individual needs" (p. 793). The goal of universal design is to ensure that courses and assessments are inclusive and accessible to all students; that is, that opportunities exist for participation and contribution regardless of (dis)abilities. Begnum and Foss-Pedersen concluded that current universal design requirements for digital assessment are lacking in specificity and do not challenge the providers of digital assessment solutions to strive for universal usability. Features and criteria for usability include that information is presented in a way that is simple and understandable, accessible (for example, easily visible), the user is physically able to interact with and navigate the system, and that the system is compatible with external assistive technologies. Based on their findings, Begnum & Foss-Pedersen claim that fundamental to strengthening universal design quality is a focus

on accessibility and usability aspects in the design specification stages. They propose a set of revised mandatory and desired specifications and outline a systematic feature analysis approach for evaluating the universal design quality of digital assessment solutions. Based on his experiences with introducing the online standardised multi-choice testing system e-asTTle, Brown (2019) also asserts the need to ensure equity and quality of experience for all test-takers. By this, Brown means that there is equity of opportunity for all students to learn how to use and interact with the testing tools and software, and that construct-irrelevant factors such poor quality devices or lack of motor skills when operating a pointer or mouse do not impact scores.

Also important in equitable assessment are tools and mechanisms that accommodate flexible learning and assessment pathways and offer possibilities for learners to exercise agency (Begnum & Foss-Pedersen, 2018; Dembitzer et al., 2017; Ministry of Education, 2007; 2010; New Zealand Government, n.d.; Timmis et al., 2016). An underlying assumption is that multiple modes of evidence capture and representation results in more valid assessment and promotes equity and fairness because this allows diverse learners to best show what they know and can do. It needs to be acknowledged, however, that more flexible or individualised programmes of learning and assessment do not always offer an equitable frame. Some forms of project-based assessment or assessment of learning in extended inquiries may disadvantage students who have yet to develop the self-efficacy or level of skills and competencies required to follow their learning through to completion and submission of a final assessment task (Campbell et al., 2007; Leadbeater, 2006; Trask, 2019). Furthermore, the level of achievement attained often depends on the availability of external (teacher or other) resourcing. For example, Wilson and McNaughton's (2014) analysis of NCEA assessment literacy demands illustrates the complexity of generic and disciplinary literacy skills required across a range of subject areas. In an equitable system, ensuring that all learners have equal opportunities to achieve in NCEA necessitates high-level teacher knowledge of individual student literacy needs and the ability to provide appropriate and targeted support.

May (n.d.) and Hipkins and Cameron (2018) point out that assessment systems and structures can construct inequalities in achievement and contribute to underlying inequities in education overall. May (n.d.) draws on the concept of cultural equivalence (or fairness) to argue that context, concepts, and linguistic aspects of an assessment may all be biased towards certain dominant cultural norms, thus disadvantaging students from minority groups. Extending this to the context of performance assessments, Hipkins (2018) suggests that while in theory performances might allow students to express their learning in culturally relevant ways, because teachers are immersed in their own social and cultural worlds they may be less familiar with other insider knowledges and understandings needed to provide students with appropriate support and feedback.

A case study conducted by Thorpe (2012) highlights another, different equity issue related to assessment systems and structures in NCEA music. Thorpe reported on an achievement standard where students were required to be individually assessed on a collaboratively composed group performance. It was found that students who were better able to analyse and articulate their individual contributions to the composing process were more likely to achieve highly than students who were perhaps more musically talented but less able to document their contribution. The implication here for valid and equitable assessment practice is that on the one hand students were valued as creative contributors in the learning process, yet on the other, summative NCEA assessment rewarded those with well-developed literacy skills over the creative, contributory process.

Threats to assessment validity and reliability associated with marking and moderation processes by nature carry implications for equity and fairness. For example, compared to out-of-context external judgements where a student is not known to an assessor, there are many sources of possible bias that arise in internal teacher judgements. These include the influences of teacher knowledge of students' socioeconomic circumstances, expectations based on prior achievement, behaviour or motivation, as

well as teachers' cognisance of pedagogical approaches employed and student interactions during the learning process (Harlen, 2005; McMahon & Jones, 2014; Wyatt-Smith & Castleton, 2005).

Tensions between different policies that at once broaden and constrain opportunities for flexibility and inclusion in high-stakes assessment have been previously noted internationally and nationally by many (e.g., Bolstad & Gilbert, 2012; Griffin & Care, 2015; Griffin et al., 2012). A situation where a focus on comparability, consistency and integrity in assessment is at cross-purposes with policies that aim to be inclusive is described by Baak et al. (2020). This critical policy analysis and case study research focussed on ways in which two South Australian schools supported students from refugee backgrounds to complete their senior secondary education and SACE (South Australian Certificate of Education) qualification. They report that school staff and students were constantly navigating the competing demands of high-stakes assessment where achievement is key, and policies directed towards flexibility and inclusion of diverse learners. Flexibilities within assessment tasks were mostly associated with opportunities to present responses using alternative formats, however, for this to happen in practice, additional support was required from teachers, which increased workload. Additionally, teachers were concerned with performance outcomes and wanted to "get their students through" (p. 12), but felt constrained by uncertainty around moderation processes and lack of diversity in assessment exemplars. In curriculum areas that included external examinations, teachers were focussed on preparing students for these and were less likely to use diverse assessment approaches. Baak et al. (2020) surmised that refugee students' diverse, multilingual competencies and skills sets were "made invisible" by structures that work to narrow the scope for flexibility in assessment tasks (p. 14). They recommend that access to a range of exemplars and increased training of staff involved in moderation could improve understanding of task responses that represent culturally diverse work.

In sum, accessibility and usability are important in the design specification of digital assessments. Also important in equitable assessment are tools and mechanisms that accommodate flexible assessment formats and pathways and offer possibilities for learners to exercise agency. This said, flexible or individualised programmes of learning and assessment do not automatically ensure equity (Harris & Dargusch, 2020). Some forms of project-based or performance assessment may disadvantage some students who have yet to develop the skills and competencies or self-efficacy needed to complete the assessment task. Teachers and the assessment task may not allow a student to express their learning in culturally relevant ways. The assessment literacy demands inherent in the different disciplinary and subject areas can be a barrier, as can the literacy skills required to read and represent ideas. In practice, flexibilities within assessment tasks can be productive but students may require additional support with implications for teacher workload and system provision.

4.2.3 Modes of evidence capture

Computer-based vs. paper-based assessment

The mode effect refers to the impact on achievement of administering assessment via a computer-based or paper-based medium. Assessment mode can also refer to the type of digital device used, for example, desktop computer, laptop or tablet. Advantages of digital or computer-based examinations compared to paper-based include administrative convenience, greater efficiency and lower costs (Fluck, 2019; van Groen & Eggen, 2020; Masters, 2013; 2017). Computer-based examinations conceivably allow for a greater range of response options, for example, multiple choice and long answer, drag and drop activities, highlighting, drawing, talking, selecting hotspots or response to digital or visual media (Newhouse, 2016; O'Leary et al., 2018). Fluck (2019) cites examples where digital exams might offer more and different opportunities than paper-based exams. These include multimedia applications such as virtual microscopes and enhanced ability to assess higher order thinking skills as well as information processing and writing skills. One example examination required access to a combination of software application, external policy documents and video media. Students were provided with fish identification software, an International Law of the Sea reference document and video media showing fish caught

using latitude/longitude to pinpoint specific locations. They were asked to determine whether the catch was legal and to provide reasons. Fluck argues that affordances such as these increase contextual authenticity of examination items (see also Section 4.3).

Computer-based and paper-based examinations differently enable fundamental tasks of reading and response. Thus, each mode provides a distinct experience, and there are implications for equity (OECD, 2015). For example, navigation when reading on a screen requires scrolling rather than turning pages in a booklet, which changes students' abilities to read or scan backwards or forwards. In a question-response computer-based format, a student might use split screens or change between windows, depending on the screen size available and font-size required to see comfortably. A paper-based examination on the other hand may consist of separate question and answer papers, giving students the ability to browse papers and set them out on a desk (Brown, 2019; Carpenter & Alloway, 2018; Jeong, 2012; OECD, 2015; O'Leary, 2018).

Prisacari and Danielson (2017) found that students in an undergraduate chemistry class made more use of working paper in paper-based rather than computer-based assessments when completing conceptual or algorithmic questions. They surmised that in a computer-based exam, students needed to either copy questions from screen to working paper or continue to refer to both, thus requiring more time and effort. Jackel (2014) found that students had more difficulty with tasks that relied on diagrams or tables in a computer-based tertiary entrance test. Students reported that this was due to lack of ability to freely annotate or draw on screen.

The representation of young people as "digital natives" (Facer, 2011, p. 18) and the prevalence of technology-rich teaching and learning in schools are both key drivers in a migration towards digital assessment (Hillier, 2019). However, the assumption that all, or even most, young people are enthusiastic, confident and competent users of technology is contested, and this needs to be kept in mind when pursuing digital assessment as a strategy that enhances equity (Thomas, 2011). As noted above, the assumption that all students have reliable and easy access to technology is also contestable. At the same time, some argue that familiarity with digital technologies and achievement in computer-based assessment environments are not necessarily related (Jeong, 2012). Backes and Cowan (2019) examined the rollout of online testing for the Partnership for Assessment of Readiness for College and Careers assessment (PARCC) in the United States. They found strong evidence that students scored lower on computer-based tests and that this represented true test mode effects, with online achievement lower by around 0.10 standard deviations in mathematics and 0.25 standard deviations in English language arts. These results were most pronounced for students at the bottom of score distributions. On the other hand, some studies have found that students write significantly more in computer-based assessments than in hand-written paper-based modes (Lovett et al., 2010; Ministry of Education and Culture, Finland, n.d.). Other advantages include the suggestion by Tian et al. (2019) that online testing and marking systems mean teachers have more time to focus more on formative aspects of teaching and learning, and that the learning gains from this are significant when compared to paper-based delivery and marking. However, other studies have found no significant difference in mode effect. In other words, little difference in test scores was found between participants who took computer-based and paper-based assessment (Chua & Don, 2013; Herrmann-Abell et al., 2018; Hewson & Charlton, 2019; Öz & Özturan, 2018; Tian et al., 2019). This is consistent with Hillier et al. (2019) who argue that young people are accustomed to computer-based learning and easily adapt to computer-based assessment.

Hillier and colleagues (2019) report on the results of a multi-university Australia-wide partnership project: *Transforming Exams across Australia*. The four-year project sought to identify the necessary underlying conditions and contextual factors for successful face-to-face e-assessments in higher education and deliver an eExam technical infrastructure. An eExam technology platform was developed consisting of a USB booting system that supported large-scale, robust, authentic assessments. This was tested in 35 research trials across the ten partner institutions involving participation by students, lecturers and administrators and measuring factors such as implementation, use, and user experience.

The eExam platform was found to be secure and reliable, resistant to network outages and able to function offline, thus reducing disruptions. On the whole, students reported that they preferred typing to writing and that the eExams aligned with their usual ways of working. However, some students faced challenges with writing in the online environment and some needed support with accessing appropriate technology. Careful organisational planning and logistical support was key to success of the examinations. Development in terms of digital literacies and alignment with digital pedagogies was also seen to be important in scaling e-assessment, as was the development of policies that support digital learning and appropriate technological requirements such as BYOD strategies. Hillier and colleagues note that the e-Exam platform is well placed for use in school assessments and high-stakes examinations (p. vi). Citing the Finland national matriculation digital examination system, a national approach was recommended for Australia as further work continues to refine policies, processes and infrastructure and continue professional learning and development work.

Writing from England, and based on results from their nationally representative sample of students in England (n = 1339) across several verbal and visuospatial tests, Carpenter and Alloway (2018) provide a number of hypotheses to explain the differences in working memory performance (the ability to process and recall information) between computer-based and paper-based testing. These include that computer-based testing increases cognitive load, the possibility of technological mode effects due to socioeconomic status (SES) and differential access to quality devices, and developmental and biological gender differences. With respect to cognitive load, Carpenter and Alloway propose that lower levels of automaticity and lower working memory capacity decreases efficiency of memory search and retrieval (p. 11), which possibly in conjunction with the computer-based environment, negatively impacts performance, especially for students with lower automaticity. In contrast to these findings, Prisacari and Danielson (2017) found no significant differences in cognitive load between computer-based and paper-based assessment when analysing question type or overall test level.

Beyond equity of access, computer fatigue from technological mode effects such as screen resolution can influence performance and create higher cognitive load, for instance, if students require more time to read and digest material from a low-resolution screen. Technological mode effects also include browsing and tactile influences, such as the ability to locate and point to objects or count them using the hand, or easily remember object locations. These tasks are different for paper-based formats as opposed to viewing them on a computer screen. Carpenter and Alloway also suggest that males may play more computerized video games than females, providing them with an advantage on visuospatial tasks. Importantly, Carpenter and Alloway point out that reviewing and changing answers is not permitted in either test mode, and that this may impact achievement especially for underperforming students as they are unable to review and improve their answers.

While Carpenter and Alloway (2018) proposed that mode effect might be related to the type and quality of device used, this did not seem to be an issue in studies of device mode effect by Ling (2016) and Hamhuis et al., (2020). Ling (2016) observed no noticeable disadvantage for US eight-grade students (n = 403) between reading, mathematics and writing testing on an iPad and a PC (personal desktop computer) for students who were experienced in the use of these devices. Hamhuis et al. (2020) used item response theory to explore mode effects in the Equivalence Study of TIMSS 2019. They found that girls slightly outperformed boys when using a tablet device compared to paper in the 2017 mathematics and science TIMMS assessment in Dutch primary school students (n = 532). There was no significant difference for the paper-based test. Hamhuis et al. observed that the extent and direction of the gender gap in mathematics and science achievement has traditionally fluctuated over the years, and that as findings of their study suggest the gender gap in computer skills is also changing, further investigation is needed. However, in contradiction to the Carpenter and Alloway study, Hamhuis found no significant mode effects for items with high reading demand and no overall significant score effects between paper and tablet for mathematics and science.

Bearing in mind the differences in findings reported above, the consensus by many seems to be that with careful assessment design and student preparation including prior practice and familiarity with the computer-based mode, achievement is enhanced (Dosch, 2012; Nardi & Ranieri, 2019; Oduntan et al., 2015; Walker & Handley, 2016; Wallace & Clariana, 2005).

Computer adaptive testing

Computer adaptive testing (CAT) has gained traction as a consequence of the affordances of digitised tasks and computer-based assessment. As discussed in Section 4.1.1, computer adaptive testing has the potential to achieve on-demand and personalised testing, although there are questions about the extent to which the breadth of the curriculum can be assessed through this approach. This section reviews a small number of additional studies which address this matter.

Adaptive/interactive e-tasks that lend themselves to automated scoring include activities such as selected response or drag and drop ordering, and simulations and virtual laboratories, for example, in integrated STEM tasks or when graphing mathematical functions (Scalise et al., 2018; Timms, 2017; Yerushalmy et al., 2017). Simulations can involve complex tasks where product and process data can be collected from each task or situation to assess a range of capabilities (van Groen & Eggen, 2020). Similar to stealth assessment, interactive assessments can be used to capture, in real time, student actions as evidence of their decision-making and problem-solving skills (Shute & Rahimi, 2017; van Groen & Eggen, 2020). However, van Groen and Eggen (2020) propose that adaptive learning environments, simulations and games may not deliver the level of precision required for individual high-stakes summative assessments and further that as summative assessment is an end-point assessment, these approaches are more compatible with formative approaches.

One form of learning analytics is embedded or continuous stealth assessment, where information is extracted from the processes that students follow. For example, when working with a spreadsheet, transaction-level data such as mouse activity, number and type of key-stroke, sequences followed, options explored, corrections made, and time spent on activities can be used to assess student skill and gaps in understanding (Masters, 2017a; Nyland et al., 2017; Tate & Warschauer, 2019). Timmis et al. (2016) suggests that this approach can reduce test anxiety, is less disruptive to the flow of learning and can reduce overall assessment.

In terms of mastery and progression, computer-adaptive tutorial systems can restrict access to subsequent levels until a student has mastered the previous level, with this level of restriction leading to a significant improvement in achievement and engagement (Paiva et al., 2017).

Out-of-level assessment (for example, used with high or low scoring students) in computer adaptive testing (CAT) is useful for measurement accuracy and for accuracy of feedback for formative assessment purposes. In high-stakes testing, out-of-level testing is useful for increasing test reliability for students who live with disabilities or those performing at high levels. However, tensions that arise such as the labelling of students and implications for exam preparation and instructional practice must be considered (Wei & Lin, 2015). Wei and Lin administered out-of-level items to grade four "main-stream" mathematics students in CAT (p. 52). Aims were to explore the extent to which an expanded item pool increased measurement accuracy and efficiency for high- and low-performing students and to explore the impact of out-of-level testing on other behaviours. The item pool used came from a commercial formative assessment program (Grades 3–8) and included multi-choice and open-ended items. Findings were that allowing out-of-level items for administration in CAT led to significant improvement in measurement accuracy and test efficiency for advanced and struggling students at a given grade. The out-of-level testing provided more specific information about students' abilities and progress. However, out-of-level testing did not seem to make a difference to the ability estimates for middle-ability students (p. 64).

Systems level data management and analysis

Due to the nature of online assessment, which includes administration of multiple assessments (CAT, embedded testlets or fixed forms), management of item banks, and the psychometric techniques to ensure consistency of scores and validity of inferences, organisations administering these assessments must develop systems to manage and quickly analyse large amounts of data, even at the item level. The addition of testing when ready and multiple testing opportunities also requires a level of student record data management to ensure a student is completing the assessment when appropriate and the student receives an alternate form in any subsequent repeat. As a result, significant resources are required to manage online assessment programmes and analyse these data in a timely manner, with larger teams required if short timelines are in place for score reporting. For high-stakes assessments, extra resources and procedures are required to ensure sufficient reliability and accuracy for the intended purposes. Along with the creation to produce and publish a sufficiently defined set of test specifications, online testing programmes should be supported by a Technical Report that addresses the procedures used in the development, administration, analysis, test-security, score reporting, data management and monitoring with respect to the purposes of the assessment.

An advantage of this level of data storage and analysis is the potential for more in-depth analyses of student-level data. Aldowah, Al-Samarraie, and Fauzy (2019) note that educational data mining (EDM) and learning analytics (LA) can "establish an ecosystem that can consecutively collect, process, report and work on digital data continuously in order to improve the educational process (p. 13). Shute and Rahimi (2016) describe the potential to monitor students as they interact with digital tasks, and this can then be used to assess understanding and capabilities. This can increase the reliability and efficiency of data collection for skills and capabilities that were previously difficult to assess or went unheeded (see also Newhouse, 2016; Redecker & Johannessen, 2013; Shute & Rahimi, 2017). One form of learning analytics is embedded or continuous stealth assessment, where information is extracted from the processes that students follow.

A developing field of learning analytics is that of cognitive diagnostic assessment (CDA). With sufficiently large datasets, models and procedures have been created that use student response patterns to identify and report on individual students' specific strengths and weaknesses relative to the outcomes being assessed. As a result, the reporting of summative assessment results can be accompanied by detailed individual student formative feedback. Although the current use of CDA is largely restricted to large-scale testing companies using primarily multiple-choice assessments, research on expanding the value and item formats supporting CDA continues.

4.2.4 Modes of evidence representation

ePortfolios

A portfolio is the purposeful collection of student work over a period of time to document, collect and reflect on learning outcomes over a period of time. Portfolios can be compiled by teacher or student and are appropriate as a tool for evidence collection for most curriculum areas. Digital portfolios or ePortfolios enable the collection of diverse and multi-dimensional evidence that can be used for formative and summative purposes, and this can improve assessment validity (Chang et al., 2012; Hung, 2012; Newhouse, 2016; O'Sullivan et al., 2012). Artifacts might include student work samples, teacher assessments, student self-assessments and reflections, or recorded performances (Newhouse & Tarricone, 2014).

Wesolowski (2014) names four purposes of portfolio assessment: (a) to showcase the student's work, (b) to showcase student growth, (c) to provide evidence of student self-assessment, and (d) provide documentation of a student's collected work (p. 83). Sufficient structure and scaffolding is necessary in portfolio assessment to ensure key learning outcomes are met and to avoid evidence gaps, however, too much scaffolding can reduce the validity of the assessment if it is not open ended enough to allow a

variety of methods for individuals to best demonstrate their learning and understanding (Newhouse & Tarricone, 2014; O'Sullivan et al., 2012).

Conley (2014) reports on the 40-school New York Performance Standards Consortium which has gained permission for students to complete a graduation portfolio in lieu of some requirements for the New York Regents state-wide standardised high school examination. Portfolios are evaluated against clear standards and scored against common rubrics. Tasks include a scientific investigation, a mathematical model, a literary analysis, and a history/social science research paper. Portfolios can also include an arts demonstration or analysis of a community service or internship experience (p. 16).

A study which explored the consistency of assessment decision-making in web-based portfolios found that teacher and student self- assessments demonstrated high consistency with the final course examination and thus had adequate validity (Chang et al., 2013). Portfolio work from 72 students in a Taiwanese senior high school computer application course was assessed by three teachers who were familiar with the teaching and learning programme, scoring rubric and criteria, and the portfolio software. Students also were familiarised with the assessment rubrics and were given opportunity to offer suggestions and modifications. Portfolio elements included learning goals, written reflections, and uploaded artifacts. The coefficients of Pearson's correlation revealed that the overall teacher portfolio scores and end-of-course scores were highly significant ($r=0.84$, $p<0.001$). Additionally, the student self-assessment scores were highly consistent and correlated with the exam scores ($r=0.71$, $p<0.01$).

Performance assessment

Performance-based learning and assessments can provide a bridge between the high school environment and tertiary study by strengthening student learning and promoting the development of higher-order thinking skills (Guha et al., 2018). As noted in Section 4.1.1, there are many types of performance assessment which can be suitable in situations where learning is not easily assessed by examination. This includes outcomes of project-based or problem-based learning where knowledge development as well as research, collaborative and problem-solving skills are important. In other words, performance-based learning requires performance-based assessment. Performance-based assessment is compatible with portfolio-based assessment and it is possible to integrate formative learning and summative assessment purposes (Chian et al., 2019; Williams & Penney, 2011) (and see Section 4.1.3).

The studies reviewed in the paragraphs below incorporate findings from a three-year ARC (Australian Research Council) project conducted by the Centre for Schooling and Learning Technologies (CSaLT) at Edith Cowan University, in collaboration with the Curriculum Council of Western Australia. The wider project investigated the feasibility of using digital representations of work for authentic and reliable performance assessment in senior secondary courses including foreign languages, engineering, information technology, arts and physical education (Tarricone & Newhouse, 2016).

Williams and Penney (2011) report on research investigating the use of various forms of summative external digital assessment in two courses (Year 11 Engineering Studies and Year 11 Physical Education Studies) in terms of manageability, cost, validity and reliability. While both of these courses have significant performance components, teachers have tended to focus mainly on the theory component which is assessed by external examination. Williams and Penney investigated the use of ePortfolios as a means of collating authentic representations of learning for assessment. In the Engineering Studies course, a computer-managed Extended Production Examination was designed where students ($n = 94$) were given a design brief and completed timed activities including written tasks, annotated sketches and video-recorded reflections (p. 33). These were curated in portfolio form and marked externally by two assessors using a standards-referenced rubric, while at the same time, the teacher marked students' work using their own assessment criteria (p. 34). Moderately significant correlations between both external assessors and between the assessors' and teacher's scores suggested that student competency was consistently recognised across the different marking systems. Digital portfolios were similarly used to curate student responses in the Physical Education Studies assessment

(n = 148). An explicit intent of this study design was to prompt integration of theoretical and practical dimensions of physical education (see Penney, Jones, Newhouse, & Campbell, 2012). The assessment in physical education also allowed students the capacity to draw or annotate diagrams to demonstrate their knowledge and understanding of strategies in games as an alternative to written responses (see Penney, Newhouse, Jones, and Campbell, 2012). In this case, evidence was collected from a four-part integrated digital assessment which included text and graphic responses and field-based video-recorded practical performances. Performances were again assessed by two external assessors and by the class teacher, this time with assessor-assessor and assessor-teacher scores significantly but only weakly correlated. Conclusions were that the affordances of digital technologies and formats can be used to enhance authenticity of assessment in courses that have a significant performance dimension by capturing and storing a variety of different evidence types and better enabling integration of theoretical and practical elements. However, it was concluded that further work was needed to address assessment reliability.

As part of the same CSaLT project, Newhouse (2013) investigated the feasibility of summative digital performance assessment in a senior secondary Applied Information Technology course. Aims were to support authentic yet manageable forms of external assessment. A computer-based production examination was conducted with 16 teachers and 17 final-year classes. Data were collected from a total of 163 students. The 2-hour examination was divided into five sub-tasks that began with a planning phase to develop ideas for a prototype product. Subsequent steps involved creating a video or animation and images to be used in the prototype, further prototype development so that it complied with specified requirements and a reflective evaluation. Students produced a variety of final presentation formats including web sites, movies or Flash animations, slideshows and static documents (p. 270). A marking rubric set out criteria for assessment from which assessors made a holistic judgement. Correlation analysis showed moderately high inter-rater reliability between the two assessors. Again, there was a moderately significant correlation between external assessor scores and teacher marks in spite of different marking criteria. An online tool which generated pairs of portfolios together with assessors scores and comments was then used for comparative pairs marking, which was continued for 11 rounds until an acceptable level of reliability was reached (Cronbach's Alpha = 0.90). Issues to note were that when students failed to follow instructions, for example, if they did not provide required information about their designs, it was difficult to make consistent judgements. Some assessors perceived that the variation in final presentation format was a limitation. In pairs marking, the number of subtasks meant that there needed to be consensus between assessors as to the relative importance of each task for holistic judgement. Student voice data indicated that most (although not all) preferred the focus on practical work to theory and that they were able to demonstrate their skills in the production exam. Around 50% of students did not, however, feel confident of achieving well. It was thought that this was due to various issues to do with lack of experience with the exam format, lack of time in the examination or lack of technical skills.

A variety of performance tasks were used as part of the CSaLT project to assess oral languages (Italian Studies) with the goal of improving assessment reliability and manageability. These involved audio or audio-visual recordings of oral presentations aimed at simulating authentic conversations. For example, students were engaged in conversation with an assessor or were asked to talk for two minutes about a holiday destination. Another task involved students first listening to a dialogue or interview before responding orally into a microphone to a series of computer-based examination questions. Digital recordings provided a record of performance that enabled judgments to be checked and moderated, whereas prior assessments were conducted face-to-face in real time and proved less reliable. Digital assessment also enabled assessments to be conducted at a student's own school, whereas previously students were required to travel many hundreds of kilometres to a testing centre (Newhouse, 2016).

Submitting digital representations in the form of online portfolios is easier in terms of manageability and cost than submission of bulky portfolio boards; however, representations need to be of a quality

that enables valid and reliable assessment. For example, it is more difficult to capture two- and three-dimensional qualities of works in digitised versions. This was the focus of a first phase of the CSaLT project for summative assessment of digitised works created by students in Visual Arts (n = 75) and Design courses (n = 82) (Newhouse, 2014). Firstly, an aim was to explore techniques and processes appropriate for conversion of visual arts and design portfolios to digital form. Scanners and cameras were used to create digital image files of various types. This proved difficult and time consuming for the Visual Arts course and the quality of some representations in terms of resolution and colour reproduction was inadequate. Secondly, the feasibility of paired comparison scoring was investigated. The digital portfolios were marked using both analytical and paired comparison methods. Consistency of these scores were compared against scores from official external assessment of the original portfolios. Overall, the paired comparisons scoring provided reliable scores for both courses. For the Visual Arts course, paired comparison scoring was judged to be better suited than analytical methods, where the inter-rater reliability coefficients were low (p. 218). The results of scoring for digitised portfolios strongly correlated with official external results in Visual Arts, but not in the Design portfolios. However, the approach lacked support from Visual Arts teachers and students who wanted the original artworks to be assessed. By contrast, the attitudes and perceptions of Design teachers and students were very supportive.

In drawing together overall conclusions from the wider ARC/CSaLT project, Tarricone and Newhouse (2016) argue for the increased use of digital technologies for high-stakes assessment and press the advantages of comparative judgement approaches in performance assessments of creative works as a viable and reliable alternative to analytical marking.

Student choice in assessment

Allowing multiple modes of representation or offering flexibility in the format of evidence provided for assessment can increase student ownership and control of the learning process. Offering students a say in some assessment conditions or adapting conditions to accommodate specific needs can also improve motivation and engagement and have a positive impact on assessment results (Irwin & Hepplestone, 2012; Pretorius et al., 2017; Rideout, 2018; Schofield, 2017).

For instance, students can be offered choices in different assessment conditions such as timing, individual or collaborative tasks and open or closed book (Gharib et al., 2012). von Heyking (2019) lists a range of modifications to assessments that include creating an audio version of the exam, reader-writer provisions, control of ambient noise, frequent breaks with the examination delivered in sections, altered font size, and text-to-speech technology (p. 31). Open-book or resourced examinations have been shown to reduce examination anxiety and can promote deep learning and enjoyment of learning, and there is little overall influence on test scores, especially when examinations require conceptual understanding, critical reasoning or problem-solving and higher order thinking. Negative effects are that students tend to rely too heavily on notes and thus spend less time in learning and preparation for examinations (Block, 2012).

An experimental study by Leithner (2011) found that students' learning styles significantly affected their performance across four different examination formats in a first-year political science/research methods course. All students first completed the Solomon and Felder Learning Style Index questionnaire. A control group (n = 45) was then given no choice of exam format while students in an experimental group (n = 45) were permitted a choice of two out of four exam formats of multi-choice, short answer with an essay, visual chart, and applied case study. Leithner argued that by offering students a choice of examination format, students were more able to show what they know rather than show how well they are able to respond to a certain examination format. Nonetheless, Leithner cautioned that students should not be exempted from the requirement to develop skills and engage in challenging tasks, for example, improving writing in written assessments. A point of interest was that many students were unaware of their learning strengths and styles, and thus less able to benefit from

the opportunity to choose a "testing style" that might benefit them (p. 424). This would suggest the need for teacher learning and student support in this area.

Irwin and Hepplestone (2012) reviewed studies of flexible assessment practices in higher education. They discuss a number of issues and opportunities associated with allowing student choice in assessment format. They note that flexibility or choice in assessment format can increase validity, as "some assessments measure a students' ability to engage with a particular form of assessment and assessment constraints rather than learning in a more general way" (p. 776). For example, having to produce a written essay under timed conditions may disadvantage some student groups, as would a requirement to master blog styles or navigate wikis in order to complete an assessment. Irwin and Hepplestone proffer the suggestion that flexibility might enhance learning because students are free to focus deeper learning if they are not struggling with a format that is difficult or unfamiliar. On the other hand, similar to Leithner (2011), they note that some assessments require students to demonstrate skills that they will need for future life and therefore development in a range of areas is important. Importantly for validity, Irwin and Hepplestone stress that the overall assessment needs to align with learning goals. If students are required to present a structured, critical commentary on an issue, formats that could be appropriate include essay, oral presentation, reflective blog, video, animated slide presentation. Regardless of the mode of representation, they reinforce that clear achievement criteria must demonstrate progress and achievement against learning outcomes.

4.2.5 Section summary

This section provides a short summary of key points from the papers reviewed in this section in terms type and mode of assessment, including considerations of reliability and validity, and equity and inclusion.

Validity and reliability, equity and inclusion:

- When collecting evidence for summative judgements, decisions about assessment type need to take into account the tensions between goals of achieving optimum reliability and preserving validity.
- Digital contexts have potential to improve construct validity with the ability to deliver pedagogically rich assessment environments which assess higher order thinking and information processing and capture multiple forms of evidence.
- Gathering multiple forms of evidence can complicate the assessment judgement process, increasing the potential for error or bias and reducing reliability.
- Test accessibility and usability are important design factors for ensuring equity and inclusion in digital assessment, including equity of opportunity for all students to learn how to use and interact with the testing tools and software.
- Flexible or individualised programmes of learning and assessment depend on equitable and appropriate resourcing and valid, fair outcomes of marking and moderation to ensure that all learners have equal opportunities to achieve.
- Valid and equitable assessment practice relies on ensuring that all aspects of assessment including context, concepts, and linguistic demands do not disadvantage students from low-performing or minority groups.

Modes of evidence capture and marking:

- eExams are consistent with students' usual ways of working and learning.
- In digital assessment environments overall, achievement can be enhanced, with the provisos that attention is paid to careful assessment design, students have the opportunity to familiarise

themselves with the exam environment and devices, there is equitable access to quality devices, and there is sound organisational planning and logistical support.

- Digital assessment enhances possibilities for 'rich', contextually authentic assessment tasks and enhanced ability to assess higher order thinking or problem-solving skills as well as information processing and writing skills by: allowing a greater range of response options; using interactive or adaptive tasks to capture assessment data in real time; and offering the ability to link to a combination of software applications, external documents and media, databases and ebooks within the same assessment.
- Offering flexibility and student control over the format of evidence provided and processes and conditions of assessment can increase student ownership, engagement and motivation as well as achievement, but this does not mean that students should be exempted from the requirement to develop skills and engage in challenging tasks.
- Online testing and marking systems can mean teachers have more time to focus more on formative aspects of teaching and learning.

Modes of evidence representation:

- The affordances of digital technologies can be used to enhance the authenticity of assessment in courses that have a performance dimension by capturing and storing a variety of different evidence types and better enabling integration of theoretical and practical elements.
- Performance-based assessment is compatible with portfolio-based assessment and it is possible to integrate formative learning and summative assessment purposes.
- Digital portfolios enable the collection of diverse and multi-dimensional evidence including the outcomes of project-based or problem-based learning and this can improve assessment validity.
- Sufficient structure and scaffolding is necessary in portfolio assessment to ensure key learning outcomes are attended to and avoid too evidence gaps, however, too much scaffolding can reduce the validity of the assessment, making it not open ended enough to allow individuals to best demonstrate their learning and understanding.
- For oral language, drama or music performances, digitised recordings can provide a permanent record of performance that enabled judgments to be checked and moderated.
- For art and design subjects, digitised representations in the form of online portfolios are more manageable and less costly than submission of bulky portfolios boards, however, the process of digitisation can be difficult and time consuming and is preferably managed within school, by the student. For arts subjects, there can be issues with the quality of some representations in terms of resolution and colour reproduction.
- Further work is needed to address assessment reliability in performance and ePortfolio assessments, however, comparative judgement approaches show potential as a viable alternative to analytical marking for providing consistent and reliable scores.

4.3 Operational case studies

This section examines what is occurring in other jurisdictions with regard to summative assessment for credentialing purposes. Characteristics of senior secondary assessment systems in Finland, South Australia and three Canadian provinces (British Columbia, Ontario, and Alberta) are described.

Key questions for this part of the review were:

- Do other jurisdictions offer more than one opportunity per year?

- If yes, what are the reasons given for offering more than one opportunity?
- If yes, what is participation like across the various sessions offered—are the participant cohorts different?
- If yes, how is reliability/dependability and validity managed?

4.3.1 Finland

There is very little national standardised testing in Finland. The digital matriculation examination is the only national assessment in upper general secondary school (ages 16–19). The move to digital matriculation examinations began in 2011. The first online tests were held in 2016 in geography, philosophy and German, with the final examinations going digital in 2019. Digital assessment was seen to be consistent with students' ubiquitous use of digital technologies outside of school and in line with workforce demands for skills in information and digital technologies. Computer-based assessment also enabled the use of a variety of audio and visual media. An additional advantage was that there are fewer logistics associated with digital modes compared with paper-based examinations. Digital examinations are more secure, can be backed up and are less likely to be lost (Ilomäki & Lakkala, 2018; Savolainen, 2017; von Zansen, 2016).

The matriculation examination assesses students' skills and knowledge against the National Core Curriculum for General Upper Secondary Education and is used as eligibility for entry to higher education (Kupiainen et al., 2016; Ministry of Education and Culture, Finland, n.d.; Savolainen, 2017). The exam is taken by all general track students (the matriculation examination acts as the official certificate of upper secondary studies together with a final report card compiling grades for all courses) (Kupiainen et al., 2016). In Finland, 50% of students opt to go to vocational institutions and 50% to general track secondary school. If the vocational track is chosen, students can still participate in the matriculation examination, however, this can be difficult because the exam is based on the upper general secondary curriculum. The system is designed to be as simple as possible and can be administered in special environments such as prisons or hospitals. It is possible to retake the exam even after students have left school (Vikburg, 2018).

Examinations are offered twice a year in spring and autumn. The examinations are held at the same time in all upper secondary schools. The largest schools have up to 200 students at a sitting, but this depends on the school and examination subject. Over a three-week period, there are nine days of exams. One examination has a length of 6 hours including around four hours of exam time and meal and toilet breaks. Exams are supervised by teachers.

Thirty-six different subjects are offered. Students must take four exams to get matriculation. The only subject that is obligatory is Finnish, Swedish or Sami, depending on the candidate's first language. Students can complete the whole examination over one three-week examination period or do it in parts over up to three consecutive periods (Ministry of Education and Culture, Finland, n.d.; Ylioppilastutkintolautakunta, 2017b). The foreign language test of advanced English is one of the most popular subjects in the matriculation examination with 51% of all candidates registered for this examination in the spring of 2016 (Savolainen, 2017).

The digital environment means it is possible to test using a variety of materials and tools with the test items such as pictures, video and audio. For example, a test item may contain spreadsheet data to be analysed using statistical tools. All students have access to the same large range of software applications including computation, graphics editors, word processing, spreadsheet and structural formulas for chemistry, and there are five different versions of online calculators available. This means that teachers have the freedom to choose the tools that best suit their students. Software is integrated meaning students can copy and paste into the exam from applications. For example, a screenshot from an online calculator can be pasted into an answer field. The examination is an html page split screen format with one side for questions and answers and one side for materials. For example, an Art History examination

may include a video of an expert discussing art, with students able to watch the video as many times as they wish before attempting the question.

Students complete their examinations on laptops. Laptops must have a power supply as batteries do not last. Students must also have headphones. It is a Finnish Matriculation Examination Board (MEB) requirement that students have a laptop and headphones or that schools will support students with loan equipment. The laptop models are not limited but for the stipulation that all devices used have a USB port. USB sticks are used as convenient and cheap external devices to boot student laptops for the examination so that each student has the same operating system and application suite (Fluck & Hillier, 2016). At the start of the examination, students boot their laptops from the USB sticks into a live Linux environment complete with the examination and suite of programmes and applications. The Linux operating system locks students out of their local files and programmes leaving only pre-installed examination applications and materials available. The USBs are delivered to the school prior to the exams. Around 47 000 USB sticks are sent to schools, with a very small fail rate. Schools receive more USBs than needed in case of failure.

The examinations and all information are on the local school server which students are directly connected to via a wired network. The system is set up to work directly from the local network rather than the internet meaning there are no issues with internet speed or reliability. Since the local servers are not connected to the internet, test items are delivered to schools. The school Principal goes online and gets the encrypted examinations from the MEB web server using sealed decrypting codes. There is an unlocking step that supervisors must do before students can log in and begin the exam. Candidates' answers and files are automatically backed up on the local server. After the examination, candidates' answers are exported to a USB memory and later uploaded by the Principal or nominee to the MEB web service where they are marked first by teachers and then by the Board's censors (Ylioppilastutkintolautakunta Studentexamensnamden, n.d.).

Interestingly, as part of ensuring security of the Digabi examination system, the release candidate system was exposed to a 'hackathon' from August 7 to September 1, 2013 where any security inadequacies discovered were rewarded with prizes (Fluck & Hillier, 2016).

Students and teachers can practice for the digital exams using the MEB-developed system called Abitti. Abitti provides a complete examination practice environment, with practice exams as well as help and support in loading the student USB sticks, setting up a server, authoring test items, carrying out the course exam in the local network and assessing the students' answers. Practice exams are well-used and help teachers and students to prepare for the matriculation examination. During development of the digital examination system, Abitti gave teachers and students direct access to the exam environment for trials and feedback. New features and programmes were communicated from social media, with feedback and requests also received this way (Vikburg, 2018).

Student experiences

At first there was strong opposition by students to digital exams. High-stakes exams are a significant source of stress and anxiety for students and technical difficulties during digital examinations can exacerbate this (Pollari, 2017). There were frustrations for students during early digital testing around typing and use of formulas in subjects such as mathematics, however, most issues have since been eliminated. Students have since stated that the digital environment feels natural and aligns with how they usually communicate. Advantages included faster completion and ease of error correction, but some students noted that faster typing can leave some errors unnoticed. Some were bothered by the sound of multiple keyboards with many people typing in a large space. Students mentioned technical difficulties such as accidental removal of the USB, which means re-starting the computer. Students pointed out issues with computer scoring where the system was programmed to accept only one correct answer when there could be multiple answers, for example, the English word 'coat' was accepted but not 'jacket' (Tarvainen, 2014).

4.3.2 South Australia

The South Australian Certificate of Education (SACE) is a two-stage qualification designed to "equip students with the skills, knowledge, and personal capabilities to successfully participate in our fast-paced global society" (SACE Board South Australia, n.d., para. 1).

The SACE Policy Framework guides subject accreditation, learning, and assessment design, and assessment integrity. It aspires to apply "rigorous and consistent standards" while special provisions support inclusion through acknowledgement of and responsiveness to "a diversity of students, in different places of learning, through the personalisation of learning and assessment (Baak et al., 2020; SACE Board of South Australia 2018, p. 1).

Students usually study Stage 1 in Year 11 and Stage 2 in Year 12. Subjects are worth 10 credits (one-semester course) or 20 credits (two-semester or full-year course). Students need 200 credits to complete the SACE, with at least 60 at Stage 2. Compulsory requirements include a total of 50 credits at a C grade or higher for a Personal Learning Plan (10 credits, Stage 1), a numeracy requirement (10 credits at Stage 1 or 2), literacy requirement (20 credits at Stage 1 or 32), and a Research Project (10 credits, Stage 2) (SACE Board South Australia, n.d., paras. 3–5).

Students complete a range of assessments according to performance standards. These contribute to subject grades (A–E) and overall results. All Stage 1 assessments are marked by teachers and moderated externally. There are three types of Stage 2 external assessment: investigations, performances and examinations. Examinations are held once a year in early November. In Stage 2, 70% of assessments are marked against performance standards by teachers with 30% marked externally (grades range from A+ to E-). Stage 2 external assessment materials for investigations (any components other than performances and examinations) are submitted for online marking after being marked by the teacher. School assessment materials are submitted for online moderation.

Electronic exams for Stage 2 have been progressively introduced between 2016 and 2020 in a number of stage 2 subjects. The move to electronic exams was in order to provide better, more authentic and relevant examinations which recognise that students are accustomed to using computers in their everyday lives (SACE Board South Australia, n.d.).

In 2020, electronic exams are available for:

- Biology,
- English Literary studies,
- Geography,
- Indonesian (continuers),
- Legal Studies,
- Modern History,
- Nutrition,
- Psychology,
- Tourism.

Students use a laptop or desktop computer with minimum specifications (for example, 13-inch screen, able to connect to the internet via wired or wireless connection, minimum of three hours connected power or battery life). A locked exam browser is installed on devices used for the examinations which prevents access to the internet or other resources. Some schools have BYOD policies that allow students to use their own devices when minimum setup and technical requirements can be met. Exams are supervised on the day by invigilators who monitor progress and device connectivity and manage situations such as device failure as they arise during the exam (SACE Board South Australia, n.d.).

Practice exams are available so that students and teachers can familiarise themselves with the format and key features (for example, spellcheck, countdown timer, single-panel or multi-panel screens,

dictionary, scribble paper, zoom or hide screen, navigation, highlighting). Subject-specific assessment tips and advice for students are provided on the SACE website. The SACE Board has developed training programmes for school staff as they plan to manage technical and administrative aspects of the electronic exams.

There are special provisions to support students who are living with disabilities or managing illness or injury (SACE Board South Australia, n.d.).

4.3.3 Canada

Education in Canada is under the control of each provincial and territorial government. Each Ministry or Department of Education develops curriculum, implements educational policy, allocates funds to schools and school boards, and ensures educational outcomes and goals are met. Each province and territory contains regional school boards that vary in size based on location and current provincial policies. Throughout Canada there is also a small set of private schools that may or may not receive partial public funding. In most provinces, school boards are tasked with policy implementation, teacher and administrator hiring, and student education, wellbeing and records. Children aged four or five years can enrol in Kindergarten, and elementary schooling begins in Grade 1. Education is commonly broken into Elementary or Secondary programs, although the change varies across provinces/territories, generally between Grades 6 to 9. Secondary schooling extends to Grades 11 (in Quebec) or 12. While systemic differences exist across the country, there tends to be shared structures and goals reflecting societal needs (McEwen, 1995; Volante & Ben Jaafar, 2008).

Of relevance for the current report, every province has some form of large-scale testing. Klinger, DeLuca, and Miller, (2008) summarised the provincial examination programmes across the country, highlighting the ongoing changes in assessment programmes across provincial and territorial jurisdictions. The review also classified the various assessment programmes based on their primary purposes: accountability, gatekeeping, instructional diagnosis, and system monitoring. High-stakes examinations (gatekeeping) were only found in the upper secondary grades. These testing/examination programmes varied in stakes, from those having little impact on students' final grades to graduation examinations that contributed up to 50% of a student's grade for a specific subject. Some provinces (e.g., Ontario, New Brunswick) also administered (and continue to administer) a literacy examination that students must successfully complete prior to graduation. Since that review in 2008, provinces have continued to modify their assessment programmes, and have also started to explore the use of online testing and increased multiple administrations. This report provides a review of ongoing work in British Columbia, Ontario and Alberta, all of whom are actively exploring online testing for summative, high-stakes purposes and expanded administration dates.

British Columbia

Education in British Columbia includes Kindergarten to Grade 12. Secondary education begins in Grade 8. The vast majority of schools are within the public-school system with schools contained inside 60 regional school districts. Private schools (primarily religious schools) receive some provincial funding, depending on the amount of provincial curriculum taught. Of interest, school structure is quite varied in British Columbia. While secondary school begins in Grade 8, many districts will have middle schools (Grades 6 to 8 or Grades 7 to 9). Secondary schools also use different timetable models, most commonly full year or semester, with a smaller number using a quarter system. These different models can be found in the same school district, highlighting a high level of community autonomy.

British Columbia has a long history of large-scale provincial examinations, from scholarship examinations to Grade 12 graduation examinations in academic subjects that contribute 40% to students' final grades in those courses. The General Accounting Office (GAO) in the USA considered the British Columbia Provincial Examination programme as one of two exemplary programs to emulate, the other being Alberta (Cheliminsky & York, 1994). Of interest, the provincial examination programme has

undergone radical changes since the GAO report. Of relevance for this report, the Grade 12 examinations have been discontinued. These were 2 hour written examinations that included multiple choice and open-ended responses and contributed 40% to students' grades in academic subjects. Examinations were centrally marked by practising teachers during marking sessions. Beginning in the early 2000's these examination programmes began to change. Currently, examinations are offered in Grades 10 in numeracy and literacy and language arts in Grade 12 (English, English First Peoples and Français langue). The current Grade 12 language arts examinations are also being replaced by Literacy assessments in 2020/21.

The examinations are administered at multiple time points during the academic year over a one-week period. The time points coincide with timetables and schools that use a quarter, semester, or full-year timetables. There are three administrations of the Grade 10 assessments and four of the Grade 12 assessments. There was also an August supplementary, but this is being phased out.

Examinations are electronic but the numeracy assessment does have a 2-page paper response component. Scoring is completed using a central marking process, but it has the capacity to have marking done online as this has been done with other examinations. As discovered in the times of the provincial examinations, the central marking process has long been considered to be a very effective form of professional development.

The literacy and numeracy assessments are graduation requirements and students are scored on a 4-level proficiency scale. "The assessment instruments are not defined as formative or summative in nature; rather, information from the new graduation assessments can be used both summatively and formatively" (British Columbia Ministry of Education, 2019a, p. 1). Students may choose to rewrite the assessments to increase their reported proficiency level.

Central to the new test are the focus on tasks or a critical thinking scenario. As an example, students taking the Grade 10 numeracy assessment complete a set of common items, a student-choice section and a self-reflection section, all based on four tasks. These assessments represent a new generation in testing for the province and it has garnered the attention of other jurisdictions.

The COVID-19 situation has resulted in an examination of using secure browsers for students to access the examination. The province is also exploring an administration system that would allow students to complete the assessments when they feel ready (Grade 10, 11, or 12). A concern expressed by the Ministry of Education is that students would delay writing the assessments until as late as possible. There is also an acknowledgement that a fully online, electronic assessment program would result in compromises with respect to the test format.

Ontario

Education in Ontario encompasses Junior Kindergarten (2 years of Kindergarten) to Grade 12. Ontario was the last province to remove Grade 13 (1988) and its replacement the Ontario Academic Credit (2003). Secondary education begins in Grade 9. The vast majority of students are enrolled in fully funded public or separate schools (primarily catholic) contained within 76 school boards and 7 school authorities. There is also a small private education sector. The vast majority of secondary schools use a semestered timetable.

Ontario is distinct in that its provincial examination programme is overseen by the Educational Quality and Accountability Office (EQAO), an independent agency of the provincial government. The EQAO annually administers three province-wide examination programs, the Grade 3 and Grade 6 Literacy and Numeracy Assessment, the Grade 9 Numeracy Assessment, and the Ontario Secondary School Literacy Test (EQAO, 2011). The Grades 3 and 6 examinations are administered over a 2-week period late in the school year, and the Grade 9 Numeracy Assessment is administered over a 2-week period at the end of each semester. The OSSLT is administered to Grade 10 students on a single date in spring and is a graduation requirement, although it is largely a minimum competency examination with the peak of its

information (and measurement accuracy) at a low level on the IRT theta scale. While students do receive a score on the test, there is an acknowledgement that the measurement error is very high for scores in the upper 1/3.

Currently, the tests are completed in a paper and pencil format. Online test formats have been explored, especially for the OSSLT. Given it is a Graduation requirement, the EQAO has explored methods that would allow students to write the OSSLT at multiple time points during the year. Currently, students may choose to defer writing the OSSLT if they believe they are not ready or, if unsuccessful, rewrite the test during a subsequent administration. The EQAO has explored the use of an online format for the OSSLT along with the possibility of multiple administrations per year. Given the number of years the OSSLT has been in place and the extensive procedures used for item development, test construction and equating, there is a large number of items available to create multiple test forms that have similar test characteristics. Statistical projections have suggested that the viability of such a testing programme would create compromises, specifically, the need to use a classical test score model and the inability to equate tests outside of a 3-5-year window.

Recently, the government of Ontario added a requirement for teacher candidates to complete the mandatory Ontario Mathematics Proficiency Test (MPT) prior to qualification. Teacher candidates will be able to complete the test at any time during and perhaps even before their teacher education programme. The EQAO was tasked with the development and implementation of the MPT in the winter of 2020.

The MPT will be a computer-based test administered at official, proctored test sites over a one-month period 4 times per year. Tests will be proctored. The MPT consists of 75 items. Four of the items are field test items. The other 71 items represent two components, 50 items on math content (70%) and 21 items on pedagogy (30%). Each applicant's test is unique and drawn from a test bank of 400 items (will expand over time). Items are administered as 5-item testlets. Items are selected based on content and item statistics (Classical test score theory) in order to provide test forms that are equal in terms of difficulty and fit to the text blueprint.

The MPT uses multiple choice items that are automatically machine-scored. Results are released about 10 days later along with some diagnostic feedback on the sub-categories. The delay in the release of the results enables EQAO to check and confirm the results.

Teacher candidates must successfully complete each of the two components with a 70% score on each, although they do not need to meet the 70% score on each of the sub-categories. Unsuccessful candidates can retake the test as many times as they wish during one of subsequent sessions. Procedures are in place to ensure that candidates do not write the same test in any subsequent attempt.

As part of its ongoing work to expand its online testing presence, the EQAO continues to explore the online testing environment. Explorations include:

- The desire to expand the item formats from the current reliance on multiple choice and traditional written response,
- online proctoring,
- multiple administration dates for the OSSLT (testing when ready),
- equating designs for online test administrations, and
- increased and more specific student feedback.

Alberta

Alberta is the fourth-largest province in Canada with a population of 4.3 million in 2018 (Government of Alberta, 2020; Population statistics, n.d.). Education in Alberta encompasses Kindergarten to Grade

12 with elementary education being from Kindergarten to Grade 6, Junior high/Middle school being from Grades 7 to 9, and senior secondary being from Grades 10 to 12. The vast majority of students are enrolled in fully funded public or Catholic schools, contained with close to just under 400 School Authorities (School Boards). Alberta also has a relatively small number of Charter schools. The vast majority of secondary schools use a semestered timetable.

Alberta has a long history of provincial testing with high-stakes diploma examinations occurring in Grade 12 academic subjects. As noted previously, the General Accounting Office (GAO) in the USA considered the Alberta Examination programme as exemplary and one to emulate (Cheliminsky & York, 1994). The Alberta Grade 12 Diploma Examinations Program has three main purposes: (a) to certify the level of individual student achievement in selected Grade 12 courses, (b) to ensure that province-wide standards of achievement are maintained, and (c) to report individual and group results.

The Diploma Examinations system is digitised in that students are able to complete written components using word-processing computer technologies.

Diploma exams are offered in selected Grade 12 courses: Biology 30, Chemistry 30, English Language Arts 30–1, English Language Arts 30–2, Français 30–1, French Language Arts 30–1, Mathematics 30–1, Mathematics 30–2, Physics 30, Science 30, Social Studies 30–1, and Social Studies 30–2. Exams are available in English and French (Government of Alberta, 2020).

From September 2015, the weighting on diploma exams changed from 50% to 30% of a student's final mark, with 70% of the final High School Diploma mark derived from school-based course work. School-based courses assess a broad range of knowledge and skills and the weighting change was to ensure that students' final grades better reflect their performance on the full range of learning outcomes in Alberta's Program of Studies. Students may retake the exam to improve their diploma examination course mark component (Government of Alberta, 2020; von Heyking, 2019). Diploma examination results are widely reported and are an important component of school and school board accountability frameworks (von Heyking, 2019).

There are set times and days for subjects and exams must be administered according to these. Because high schools in Alberta follow a range of schedules (full year, semestered, quarter system), there are two main administration periods in January and June, with further administration periods in November, April and August. Examination schedules are released well in advance; the current scheduling information covers from November 2019 to August 2022 with schedules for 2021-22 considered draft until confirmed in November 2020 (Alberta Education, 2019; Government of Alberta, 2020).

Language arts (English or French) and social studies exams have a written component (Part A), and a multiple-choice component (Part B). Students are permitted to complete the written component of the exams using digital technologies. Students must seek and gain permission to use technology from the school principal under the following conditions:

- students usually produce written work on a device;
- students are proficient at using the devices and word processing applications;
- students understand and can abide by specific rules and procedures for preparing and submitting examination answers using technology;
- technical expertise is available for the duration of the examination period; and
- security, validity and confidentiality of student work and diploma exam materials are in no way compromised.

The principal must ensure all devices and printers used for the diploma exams are appropriately configured to safeguard exam security, validity and reliability and to minimise distraction to students. Quest A+ enables secure online exam administration on both student and school-owned PC and Apple devices. Once a secure exam has begun, the locked browser disallows screen capturing and blocks access to the internet and to local networks. It does not permit students to exit the test environment.

Students can access word processing tools but functions such as autocorrect, and predictive word search are disabled. Practice tests can be found on Quest A+ and are available on demand (Quest A+, n.d.).

The exams are supervised or administered by school staff but not by those who teach the subject being examined. Student and supervisor secure access codes are tied to date, time, and school code (Alberta Education, n.d.; Quest A+, n.d.). Students are required to print and staple their examination responses to their diploma exam booklets and complete verification steps to ensure that what is attached is accurate and complete (Government of Alberta, 2020).

Marking of written exam components takes place in set time periods by nominated teachers. For example, marking sessions for January 2020 in English Language Arts begins January 16th with main marking activity occurring January 23rd - 29th. A small group of teachers who were involved with standard setting grade a representative sample of student papers. This is followed by group leader training. Small groups of markers work in marking sessions under a group leader. Every student paper is marked twice with any discrepancies going to a third marking. Several times during a marking session, there is a standard check where everyone marks the same paper to check for consistency. Machine-scorable answer sheets are marked by scanning machines. For example, the mathematics examination consists of machine-scored, multiple-choice and numerical response questions and a few written response questions. Science examinations consist of numerical response and written response questions that are machine scored (Government of Alberta, 2020).

4.3.4 ACT

ACT (Formally American College Testing) has a long history of high-stakes testing, especially as an alternative to the SAT for college admissions. The test itself consists of four compulsory multiple-choice domains: English, Mathematics, Reading, and Science. The ACT also has an optional writing component. ACT administer its test across international jurisdictions and at multiple time points. Recently, ACT has started to administer testing programmes for certification purposes.

ACT uses proprietary software that allows for automated test assembly of parallel linear forms (fixed forms) meeting the content and "difficulty" blueprint. Thirty operational forms are produced. Given the high stakes of the College admissions test, test security is paramount. As a result, ACT uses a specified distribution and administration protocol of the forms to control and reduce test exposure domestically and internationally.

While it is acknowledged that single date and time administration is the best option, the nature of the College test requires much more flexibility. Tests are administered online at multiple time points and multiple locations. Students complete the examination at specified testing centres. In smaller communities, local businesses serve as examination sites. ACT is exploring the potential of remote proctoring to further enhance its testing capabilities across multiple sites, providing greater flexibility for administration.

4.3.5 Section summary

This section provides a short summary of key points from a review of characteristics of selected senior secondary assessment systems in Finland, South Australia and Canada in terms of number of assessment opportunities and considerations for managing digital assessment.

Number of assessment opportunities:

- Students have increased flexibility in choosing when to complete an assessment either within the same year or across years if multiple testing windows are offered, for example, Finland (two examination windows), British Columbia and Alberta (3-4 examination windows).
- Students are the primary decision makers as to when an examination will be completed except for those assessments linked to final course examinations.

Digital assessment:

- Digital assessment is being progressively introduced and is seen by all jurisdictions reviewed to be consistent with students' use of technologies in school learning and outside of school as well as in line with workforce demands for skills in information technologies.
- The purposes and the stakes of online assessments vary.
- Test formats for online assessment tend to be restricted to multiple-choice and short-answer open response items, although more complex tasks and items are emerging, with the digital environment enabling testing using a variety of materials and tools.
- Test security is a primary concern, especially for high-stakes assessments.
- It is common for access to software applications within an examination (including computation, graphics editors, online calculators, word processing, spreadsheet and structural formulas) to be restricted and to lock students out of their local files by using USB sticks to boot devices into locked browsers.
- Digital examinations can be securely administered using either students' own devices or school-owned devices.
- Internet speed and reliability is a key issue for online examinations, and this issue can be avoided if examinations and all information is on a reliable local school server by which students can be directly connected to via a wired network.
- Minimum specifications apply for devices in terms of battery life, screen-size, screen resolution, internet connectivity, and battery life.
- The manageability of submission processes for online assessments is easier compared to paper-based assessment.
- It is common for practice exams to be made available for students and teachers.
- There are benefits to teacher professional learning when teachers are included in the marking process, with systems of automated scoring or online marking vs the use of central, face-to-face or distributed marking varying across the jurisdictions reviewed.
- Online assessment programmes are backed by sophisticated technical structures to ensure score equality across multiple forms.

5. Synthesis of findings and implications

The first part of this report synthesised findings from research and practice to address three global questions:

- How does the timing and type of summative assessment (with a focus on digital assessment) as a core component in the measurement and awarding of qualifications in senior secondary school, impact on sustained, deep learning?
- What is the impact on reliability and decision consistency, and the validity of the interpretations made from the results of summative assessments when administered at multiple time points, including the impacts on teachers' and students' pursuits of learning?
- What are the practical, structural and measurement challenges and opportunities associated with implementing a summative assessment when ready process?

Overall, the role of digital tools in high-stakes senior secondary summative assessment and the potential opportunities and challenges presented by on-demand or when ready assessment are under-researched. In the higher education sector, instances of digital innovation in assessment including novel, responsive approaches to design and administration that include a focus on the formative and summative functions of assessment have been noted. This review of national and international literature has shown that in the tertiary environment it is possible to undertake online or digital formative/summative or blended models of learning and assessment which accommodate the needs of moderate-sized student cohorts (i.e., $100 \leq n \leq 1500$). However, this field is very much under development in both practice and research. In the final report to the Australian Government on a large-scale, four-year project: *"Transforming exams across Australia: Processes and platforms for e-exams in high stakes, supervised environments"* Hillier et al. (2019) note that across higher education institutions there is "currently no robust, viable method to do authentic e-assessment that will work to align both in-class progressive assessment for learning and higher stakes summative assessment undertaken in large scale exam halls" (p. v). Nevertheless, tertiary level assessment research offers some ideas for ways forward in the senior secondary sector.

While cross- and within-institute collaborations or partnerships may exist within the tertiary education sector, universities and other tertiary providers generally function autonomously when administering assessments, with qualifications approved in New Zealand by quality assurance bodies, the NZQA and Universities New Zealand (Tertiary Education Commission, n.d.). This means that the number of students taking part in any one tertiary institute assessment event is lower relative to the numbers that might participate in standardised national assessment in New Zealand secondary schools. Scale is an issue for secondary school assessment, where a student cohort numbering around 145 000 participates in NCEA external assessments at over 400 examination centres each year (NZQA, 2019). Also, of note when comparing higher education assessment models with secondary school summative assessment is that tertiary calendars are typically based on two or three semester cycles per year, and examination/assessment windows are aligned to these. In New Zealand, the school calendar is based on a yearly cycle, although the four-term structure offers potential for semesterisation.

This discussion considers what is known and relevant about timing and types of digital assessment in the context of the NCEA. The goal is to inform future innovation focused on enhancing flexibility in the qualification system in a way that delivers advantages for students while avoiding unintended negative consequences and minimising operational or administrative challenges. There are tensions and necessary compromises between opportunities and constraints when managing innovation. In designing successful assessment systems, it is necessary to leverage the expertise of qualifications authorities in areas of logistics, administration and quality assurance in partnership with those who have expertise in IT and technical support (Hillier et al., 2019). It is also important to leverage the expertise of those who have an understanding of curriculum and likely student learning pathways and motivations. This approach acknowledges that curriculum, pedagogy and assessment act together to shape teacher and student experiences of teaching and learning (Bernstein, 1977). In the context of NCEA goals it is also important to consider how to ensure a productive interplay/interaction between formative and summative assessment (section 4.1.3; Gonski et al., 2018; The Gordon Commission on the Future of Assessment in Education, 2013). Within this complexity, there is much debate about which of curriculum, pedagogy and assessment is, and should be the main driver of educational reform efforts and direction. We take this debate into account in detailing the affordances and challenges of different assessment regime timings and formats.

The sections below primarily focus on opportunities for assessment innovation in the NZQA's reform efforts and articulate possible drivers and constraints for issues related to assessment tools, systems integration, data management, and student learning motivation and achievement. These drivers and constraints have implications for operational decisions to best design assessment systems to optimally meet the needs of students and educational programmes going forward.

5.1 Timing of assessment

This section focuses on the potential affordances and challenges of offering more than one summative assessment opportunity per year, for example, a second externally assessed or external digital examination window at midyear.

Table 1: Considerations related to timing of assessment

Learning affordances	There is potential to open up learning pathways and provide more opportunities for student demonstration of learning along with more timely advancement to next learning stages (section 4.1.1). It is possible to avoid the negative and demotivating consequences of poor performance if students are tested when ready rather than to a timetable (Harlen & Deakin Crick, 2003). Extra assessments in the form of resubmission/reassessment opportunities are a contributing factor to higher rates of achievement (see section 4.1.2). An early opportunity to demonstrate competence in assessment may build assessment confidence (Kōrero Mātauranga, 2020). Midyear assessment allows courses to be split into semesters (Kōrero Mātauranga, 2020).
Learning tensions and challenges	More frequent testing has been shown to increase student motivation and achievement and to act as an equaliser in overall achievement, but this benefit depends upon the subject and the testing regime (section 4.1.2). Assessment that is timed for an end of year exam period allows the maximum time to consolidate conceptual learning and to develop and practise the skills and competencies required for effective demonstration of learning (Kōrero Mātauranga, 2020, section 4.1.2). More frequent assessment increases student stress (see section 4.1.2). Resubmission/reassessment opportunities can mean that students do not necessarily take full advantage of first assessment opportunities, knowing there is a backup (section 4.1.2). Low marks in an early summative assessment might influence students' self-efficacy for future learning and perceptions of themselves as learners who are able to achieve (Hopfenbeck, 2015). An increased number of summative examination opportunities can mean increased time and focus on assessment at the expense of teaching and learning (section 4.1.2). In high stakes contexts, frequent testing can have detrimental impacts on student motivation for learning and to student attitudes and approaches necessary for sustained 'lifelong' learning. (Harlen & Deakin Crick, 2003). Frequent summative testing can encourage test-taking behaviours that are detrimental to higher order thinking (Harlen & Deakin Crick, 2003).

	Other impacts of frequent testing include: increased pressure (on teachers and students) to do well or improve achievement by whatever means; increased test anxiety; and teaching targeted more and more towards practices which ensure the best results, such as repeated practice tests, which in turn reinforces the importance of the exams. There may be greater pressure from whānau/caregivers for students to be 'ready' at particular times. If students' progress at varying rates when gaining high-stakes credentials, this can aggravate unhelpful competitive aspects such as parental pressure or peer-to-peer comparison and may feed negatively into wider community discourses linked to student stereotyping and school rankings and comparisons. This could lead to issues such as the performance and desirability of schools being associated with the number of early passes they achieve, and students and schools attempting to 'game' the system to maximise pass rates (Flórez et al., 2018; Ingram,et al., 2018; and see section 4.1.1). Shorter teaching/testing cycles can lead to improvements in achievement results in the short term but may not translate to the development of learning dispositions and capabilities that might serve a learner in the long term. For example, while not systematically researched, examination data and anecdotal evidence in British Columbia in the 1990s regarding the 10 week (quarter), 20-week (semester), and 40-week (full year) school teaching models, indicated that students in the quarter system received higher scores on the provincial examinations but fared more poorly in subsequent university classes. Such findings suggest there are complex interactions between curriculum and programme structure, summative assessment results, learning, and subsequent achievement. External assessments can constrain the taught curriculum and limit instructional autonomy to best meet students' learning needs and progression, with more frequent assessment providing stronger constraints. External achievement standards that are only available for assessment mid-year could, depending on the nature of the Standard, constrain the taught curriculum, which would run counter to the NZC press for local, customised and responsive curricula.
Administrative affordances	The development of a digital platform enables new infrastructure to be built that takes advantage of current and evolving technologies. Administrative workload is distributed more evenly over the course of a year. Digital assessment systems are better able to use psychometric methods and models to monitor assessment quality and ensure assessment consistency across multiple administrations. Markers would develop expertise more quickly. Using agile, distributed digital marking systems (e.g., papers scanned and uploaded, markers meet online), more teachers could have the opportunity to benefit from professional learning as a consequence of being involved in marking and moderation (Frost, 2010; Smaill, 2020).

	Online marking could increase the number of markers available and reduce the need for travel for moderation meetings (Frost, 2010).
Administrative challenges	Systems timeframe for developments and scalability challenges include increased costs due to the need for more resources and larger teams required to manage online assessment programmes in a timely and efficient manner. For high-stakes assessments, extra resourcing is required for development, administration, test-security, data management and marking, score reporting, and monitoring to ensure sufficient reliability and accuracy of psychometrically comparable tasks (Section 4.2.3). For digital exams, technical challenges include examination or assessment security, and reliable, high-speed internet connectivity. If digital assessments are intended to operate on multiple platforms, efforts must be made to ensure similarity of administration (Section 4.3.1). For on-demand examinations at a distance, delivery systems need to include security mechanisms that ensure robust identity verification procedures are in place to allow student access to an assessment and to monitor its completion. It is possible to use biometric markers such as facial, voice, keystroke, however there are associated ethical issues (Fluck, 2018). Multiple tests/test form equivalency: offering additional examinations means more investment in developing item banks, managing security/item leakage and processes of test equating. For example, if there are two external examination windows per year - two separate assessments need to be designed and equated (Sections 4.1.1, 4.2.3 and 4.3.3). School systems and infrastructure would need to enable and be able to cope with flexibility in assessment including the ability to manage the number of students who can access an online assessment environment at any one time.

5.1.1 Important questions to be asked

How would the system decide what 'when ready' means? And, who decides when a student is 'ready'? If the decision is made with and by the student, how will this be managed?

> This means an ongoing pedagogical focus on developing students' self-management and self-assessment skills (Harlen & Deakin Crick, 2003), which aligns with the focus in the NZC key competencies, the long-standing focus in New Zealand schools on formative assessment/assessment *for* learning, and current understandings of 21st century capabilities.

Would adding an extra assessment window exacerbate (or merely make more visible) equity issues associated with achievement gaps in that some students would progress very quickly to next steps? How would students who were otherwise less likely to progress quickly, be supported? (See Section 4.2.2).

A key question for implementation is how schools and instructors might manage groups of students who are at different points in the learning progression and level of readiness to proceed further?

> There are models of schooling that enable students to be more self-directive in terms of monitoring and progressing their own education within subjects. This shift away from a time-based learning model towards a more skills-based or competency achievement model has become a subject of discussion in professional education.

How can midyear assessment contribute to student self-assessment and be used to guide and scaffold progress towards a longer term and more expansive end goal?

What type of curriculum-based learning in each subject area is best summatively assessed early in the year or at the point of learning? For example, should students be first tested on their ability to demonstrate understanding of conceptual knowledge before they are required to research, investigate, evaluate or apply that knowledge in a new context?

How can learning goals be promoted over performance goals? How can assessments be developed that are broad enough to discourage schools' and students' micro focus on 'how to pass' exams (see Section 4.1.2).

> The negative impacts related to a performance focus can be reduced by promoting a learning goal orientation rather than performance orientation, avoiding or de-emphasising drills or practice tests, using a range of forms of assessment, and avoiding administering assessments to students who are unlikely to achieve (Harlen & Deakin Crick, 2003).

Does the same team of markers need to be responsible for marking and moderating a Standard each time is assessed? What are implications for cost/funding models? How do you create a process to attract, include and upskill new markers? What methods can be used to ensure teacher markers are available midyear? (See section 4.2.4)

5.2 Type of assessment

What is assessed and the mode of assessment influences teaching and learning – especially in high stakes environments (Baird et al., 2017). Hipkins and Cameron (2018) reporting to the Ministry of Education on trends in New Zealand assessment stated that, "what is assessed in high-stakes contexts continues to become the curriculum that is enacted in classrooms" (p. 25). This notion that assessment can drive curriculum direction and reform is prevalent. It has resulted in the exploration of how multiple types of assessment might lead to educative structures and to more integrative teaching that focuses on broader educational goals and reflects transferable skills and competencies. As an example, the ideas of authentic assessment, performance assessments or constructive alignment between 21st century curriculum goals are often considered key to a robust, valid assessment system (e.g., Biggs & Tang, 2011; Hillier, 2018). Currently, the rapid development of digital tools and technologies and increased

use of digital pedagogies provides a similar rationale for exploring methods of digital assessment as also key to authentic, valid assessment (e.g., Fluck, 2019; Frankl, 2018).

This section focuses on the aspects to be considered with respect to decisions regarding the implementation of different types of digital assessment.

Table 2: Considerations related to type of assessment

Learning affordances	Digital modes of assessment afford opportunities for different types of tasks within examination settings including those based on the software applications used in teaching and learning (Section 4.2.3). Digital modes of assessment afford opportunities for different modes of expression and demonstration of learning and have become a common framework in which students operate and communicate. Digital modes can potentially support formative - summative integration and collaborative work, for example, a series of open book, formative, collaborative tasks that strengthen individual summative performance (Section 4.1.3). There is scope for digital portfolios to support 'rich task' type assessment and curation of evidence in extended investigations (Section 4.2.4).
Learning challenges	Educators and education systems are unequally proficient in their expertise and the use of different types of assessment and also the integration of technology. It will be necessary to build teacher expertise in performance-based assessment and reporting strategies and in ways to use technology to support such assessment frameworks (Hipkins et al., 2018; Thille, 2016). It will be necessary to build teacher expertise enacting models of formative - summative assessment integration (Section 4.1.3). While students are generally very familiar with digital platforms, there remains significant variation in the associated skills (e.g., keyboarding speed and accuracy). While it is also true that there is variation in handwriting ability and fluency, students can feel that it is unfair if some have better-developed keyboard skills in a digital examination environment (Meacheam, 2018). Hillier et al. (2019) noted that a significant minority of students needed support in the adjustment of writing habits in the transition to online exams (Section 4.2.3).

Administrative advantages	Depending on the examination format, digitally assisted marking systems can free up time for teaching and learning (Section 4.2.3). BYOD policies mean devices are familiar to students, costs are lower and digital assessment is more easily scalable (Section 4.2.3). Costs associated with printing and transporting papers are reduced (Section 4.2.3). Based on the experiences of operational case studies, there is an initial cost to digitisation that is likely to be substantially higher than current annual costs for paper versions. This initial extra expenditure is likely to be recuperated over time. Further, digital platforms are much more responsive as highlighted by their ability to quickly resolve assessment errors. Improved security with automatic backup means exams are less likely to be lost (Sections 4.2.3, 4.3.1 and 4.3.3). Digital reticulation of scripts and marks means faster, more agile marking processes. This opens possibilities for flexible external marking and assessment i.e., multiple pairwise comparisons of a single piece of work executed using an online platform (Section 4.2.4). Digital marking of paper examinations could result in greater reliability and provide more question-by-question data to schools and qualification bodies. For example, each page of an examination has a barcode which enables papers to be machine-sorted by question before being scanned and uploaded to a marking platform/interface at a central processing centre. Markers meet and work in the online environment. Teams or pairs of markers are able to grade a single question. If in agreement, the grade is recorded, if not, the paper is automatically flagged and sent for moderation. Immediate feedback can then be sent to markers (Frost, 2010).
Pedagogical challenges	The examination mode and format need to be consistent with approaches used in teaching and learning. If digital devices and software applications are 'tools of the trade' in work or post-school learning environments, there needs to be alignment at the school level with the use of digital pedagogies and digital assessment. This means that digital assessment literacy will become increasingly important for students and teachers (Sections 4.1.3, 4.2.1 and 4.3.3). There are implications for those with relevant disabilities or varying levels of technological expertise and keyboard skills (Section 4.2.2). Digital examinations need to be mainstream at junior secondary level so that senior students are very familiar with the digital exam environment (Fluck, 2018). This has implications for development and costs for school systems and processes as junior examinations are managed at this level. As different subject areas use and apply different computer software applications and information systems, disciplinary digital literacy needs to become part of teacher pedagogical content knowledge in each field.

Administrative challenges	Markers would need to be comfortable with digital modes and be aware of cognitive bias or 'halo effects' associated with handwritten answers. Handwritten answers can look more substantial than typed answers. Markers can have higher expectations of typed answers due to the 'forgiveness factor' and the ability to read between the lines. Higher grades can be awarded for handwritten tests than typed papers (Fluck, 2018; Meacheam, 2018). On the other hand, typed answers are easier to read than written answers and can enhance marker objectivity (Műller & Bayer, 2007). Interestingly, e-Exam trials conducted by Austrian Alpen-Adra University reported changes in grading distribution with the introduction of digital exams. It was harder to gain very top marks in a digital exam than in a written exam (Frankl, 2018). Reasons for this were unknown but one could hypothesise that this could be related to marker expectations and bias?

5.2.1 Important questions to be asked

At present digital assessments tend to use traditional formats such as multiple-choice questions, short and long text answers. What potential is there in the system to explore new and innovative digital formats, knowing that the reliability and validity of such assessments have yet to be determined?

How can the different affordances and modes of representation available through digital assessment be used to expand the opportunities students have to demonstrate what they know and can do? In what ways can this expansion of representational tools be employed to address access and equity issues?

Given the focus on group work and collaboration as a key competency an important question to be asked is: Should assessments which by their design do not fully reflect the collective learning that occurs in a classroom, be a primary driver for pedagogical reform?

To what extent does digital/information literacy need to become part of the curriculum?

Can/how can school technological infrastructures be designed to support assessment security and multiple students to access different achievement standard assessment tasks simultaneously?

5.3 Scenarios

The scenarios below are intended to contextualise discussions about assessment timing and frequency (more external assessment opportunities) and types (options for digital or online tasks) of assessment. They focus on aspects of senior secondary learning and assessment under NCEA. The scenarios take into account the seven principles of the NCEA change package, review of achievement standards and digital assessment vision, as these policies and initiatives will act as mechanisms to support change (Hillier et al., 2019). Opportunities and challenges are presented from an individual student point of view and from a wider/school or system perspective.

Note: With regards to different timing and types of assessment, it needs to be noted that the existing structure and inherent flexibility of NCEA as a high stakes qualification (and prior to the NCEA change package) supports on-demand internal assessment of many different types and in many different formats. Clear guidance is given on the NZQA website as to when students should be assessed: when the teacher is confident that achievement of the standard is within a student's reach; or at the final deadline for the assessment, if there is one. Also, clear guidance is given about gathering evidence for assessment and how to minimise the need for further assessment (NZQA, n.d.-f). Thus, student learning and preparation of evidence for summative internal assessment typically involves cycles of formative assessment and feedback as part of permitted, level-appropriate guidance or supervision from teachers.

5.3.1 Exploring possibilities for revised Level 1 Science assessments

Students tend to consolidate both their learning and ideas about future plans as they progress through senior secondary school. The level structure of NCEA reflects this, with Level 1 typically offering a more expansive curriculum and teachers more opportunities to innovate learning approaches (New Zealand Government, 2019). The goal is to allow students to build a foundation for study at Levels 2 and 3 (Kōrero Mātauranga, 2020). More work is completed in class and authenticated by student and teacher before being sent to NZQA for external marking and or moderation. As students progress through NCEA Level 2, they are likely to be considering a number of different future possibilities and pathways. They are increasingly engaging with and developing specialised conceptual knowledge, capabilities and skills. In response there is a greater use of formal assessments, both internal and external, while also providing opportunities for students to experience an extended project or investigation to reinforce their growing independence as learners. Lastly, students progressing through NCEA Level 3 will begin to focus their learning, while keeping options open. The role of formal summative assessment increases and emphasis shifts as students focus on attaining university entrance and preparing for postsecondary school pathways, especially those pathways and careers that have entry requirements.

In the case of the Level 1 science, the new draft NCEA science achievement standards reflect the NCEA change principles of simplified structure with fewer, larger standards and a 50:50 split between internal and external assessments across four standards (Kōrero Mātauranga, 2020).

AS 1.1 Internal: Use a range of scientific investigative approaches

Students could complete a range of group, practical investigations with formative group feedback/evaluations as part of the learning process within each of the four disciplinary strands. A digital or paper-based portfolio would be used to curate individual write-ups of student experience and thinking over the course of the year with internal summative assessment taking place at the end of the year or when the teacher and/or student agrees the student is 'ready' based on the quality of evidence collated and the insight demonstrated by student annotation/discussion. Evidence could be represented in a range of modes including text, photographs, videos, simulations and so on.

AS 1.2 Internal: Engage with a socioscientific issue

Following foundational contextual and conceptual learning and development of capabilities, students might independently research a question related to the chosen socioscientific issue and produce a digital or paper-based report to be submitted for internal assessment midyear or when deemed ready.

AS 1.3 External common assessment activity: Describe attributes of science that contribute to the development of scientific ideas and processes

Students could curate a range of evidence that demonstrates learning about the development of scientific ideas and processes across a range of contexts. A digital or paper-based portfolio is submitted for external assessment midyear or when ready. Students could prepare a newspaper report or update a Wikipedia entry.

AS 1.4 - External common assessment activity: Interpret scientific claims in communicated information

A written examination is completed at the end of the year, which includes short and long answer tasks based on unfamiliar contexts (Kōrero Mātauranga, 2020).

Affordances:

- An external examination would address issues of assessment authenticity and reliability (Kōrero Mātauranga, 2020).
- Fewer standards, and externally marked standards reduce teacher workload.

- "Offering external assessment early in the year can provide students with an early opportunity to demonstrate their competence (for example, of an aspect of science capabilities that is built up across years 9-11) and builds assessment confidence. It will also allow Science courses to be split into semesters" (Kōrero Mātauranga, 2020).

Challenges:

- New standards and midyear external assessment will drive major pedagogical shifts - teachers will need to be part of conversations and 'on board'.
- For students to be able to curate a range of evidence that demonstrates learning about the development of scientific ideas and processes across a range of contexts, teachers need ensure that students engage with the same ideas and processes across a range of contexts.
- There is evidence that people need to meet a concept in three contexts to fully understand and appreciate it (Alton-Lee, 2003). Curriculum design will need to allow for this in order to prepare students for AS 1.4.
- The inclusion of unfamiliar contexts in examination questions has implications for equity and the cultural validity of assessments (Solano-Flores & Nelson-Barber, 2001).
- Overseeing portfolio evidence-gathering processes could increase teacher workload.
- Literacy demands of external examination standards in unfamiliar contexts - this may be an issue even if students are familiar with relevant conceptual knowledge. For example, low-literacy students may study and become familiar with the language necessary to understand and interpret scientific claims communicated in the context of climate change with in-class support from a teacher, but when presented with a different context (for example, challenges associated with non-communicable diseases) in an examination task, they may struggle.
- 'Funds of knowledge' demands of external examination standards in unfamiliar contexts may be an issue even if students are familiar with relevant conceptual knowledge (Solano-Flores, & Nelson-Barber, 2001).
- Currently NCEA external examination and folio marking begins in November, once seniors have left. Markers generally work into the holiday period until around December 22nd.
- . If external assessments were submitted midyear, markers would need to be available.
- Under the current NCEA structure, students commonly enter more than one external to "have extra chances at getting an excellence," i.e., to maximise their opportunities for attaining a merit or excellence grade. What are the implications of midyear assessment and fewer standards for NCEA certificate and level endorsements?
- If achievement standards were only available midyear (rather than available on demand or mid- and end of year), this would reduce flexibility and could interfere with schools' ability to develop and implement local curriculum, meaning a loss of local flavour as promoted in NZC.

5.3.2 Flexible pathways for students: Midyear summative assessment

The new NCEA achievement standards will be developed to reflect the NCEA change principles of clearer pathways to further education and employment, including opportunities for students to follow mātauranga Māori pathways. Adding a midyear external assessment window coupled with changes to the achievement standards could conceivably revolutionise how senior students plan and approach their senior schooling years by supporting diverse pathways from secondary to postsecondary learning and employment.

Midyear summative external assessment could open opportunities for semesterisation of the school year and curriculum.

Affordances:

- Assist students who want to change pathways: a student taking L2 general science might switch midyear to taking L2 health or biological science as they prioritise an option to continue with a

health-focused tertiary qualification. A student taking L2 physics might complete midyear assessments and switch focus to join the Trades Academy as their thinking about securing a mechanics apprenticeship becomes more definite.

- Help to ensure students leave school with qualifications: a student is leaving school part-way through Year 12 to enter a building apprenticeship but needs to complete literacy/numeracy requirements. They prepare for and enter the midyear examinations. The student also has been working on L2 Hard Materials and would like their learning and achievement in this recognised, and so submits a portfolio midyear for assessment.
- Future-proof student options and pathways by supporting a broad foundational focus at Year 11. For example, a Year 11 student decides to take one semester each of Drama and one of Music, keeping their options open for senior years.
- Future-proof student options and pathways at Year 12/13. For example, a Year 12 student who is unsure about their future path returns to complete Year 13. During that year, an opportunity presents to secure an entry-level job in the horticulture industry with good prospects for progression. The student submits work for internal and external assessment before they leave, enabling them to keep options open for university study.
- Support increased secondary-tertiary connections and mobility: a student who is part of a secondary school accelerated learning pathway completed L2 mathematics, chemistry and physics in Year 11 and now in Year 12, is working on L3 calculus, chemistry and physics. They plan to study science or engineering at university. In Year 13 they will enter the three scholarship examinations at the end of the year. Additionally, they will take stage one university chemistry and biology papers in the first semester of their Year 13.
- Under the current system, a Year 12 student studying L2 history, classical studies and French as well as L3 chemistry, calculus and physics is under constant pressure to complete internals during the year and will prepare to sit six external examinations (and possibly two scholarship examinations) at the end of the year. Midyear assessment and fewer standards could spread this demanding academic load more evenly throughout the year.
- Benefit students who change schools during the year: a student who arrives at a new school in July could/would have submitted work and gained credits from the first semester.

Challenges:

- While midyear assessment would take the pressure off the end of the year, a question arises about the impact of assessment pressure midyear.
- While there could be increased opportunities for personalised programmes, issues associated with equity of opportunity for diverse student groups to learn and progress would need to be carefully thought through and actions taken to avoid negative or unintended consequences.
- A factor to consider is the impact of semesterisation and extra external assessment opportunities on the current differences in status and the achievement gap between internal and external (standardised) assessments.
- Will there be an opportunity for re-assessment of the midyear external assessment in the end-of-year assessment window? This would require multiple (at least two) versions of the same assessment, thus introducing challenges for task equivalence and reliability as well as increasing marking/moderation demands.
- External portfolio resubmission would require additional work by the student and remarking, adding to external marking/moderation workloads.

- Some students may be 'ready' to sit the literacy/numeracy assessments in Year 7 (New Zealand Government, 2019). Others may not progress to this stage until Year 11 or beyond, which would require schools to provide extra targeted support to these students (e.g., by making learning goals explicit and showing them how to direct effort in learning) (Harlen & Deakin Crick, 2003, p. 202).

5.4 Possible steps towards creating more flexibility in the system

This section explores possible steps towards the introduction of on-demand or when ready digital assessment. To achieve the necessary system changes when instigating a move towards offering on-demand or when ready digital assessment, separate technical, procedural, and pedagogical conditions, contexts and issues need to be identified, allowed for and resolved. A staged process could usefully involve planned iterations of devising, designing, piloting, refining, systematising and finally, scaling (see Hillier, 2019).

Possible starting points that take into account contexts and directions of the NCEA change package and review of achievement standards could include:

- Work towards trialling and offering two online L1 literacy/numeracy assessment windows per year (midyear and end of year), with on-demand and when ready assessment available as future investment in digital administrative and security systems permits.
- Develop and extend existing online formative-summative tools for NCEA on-demand internal (school-based) assessment, for example, digital tools to support student self-assessment and student-led evidence collection and curation towards digital portfolio submission.
- Trial midyear external summative assessment via moderation of school-based, teacher administered, teacher authenticated portfolio work. Work towards externally assessed midyear submission of school-based work or digital portfolio.
- Explore online exams that require lower levels of security, for example, open book or restricted open book digital exams. Use time-limited format, plagiarism checkers, and task/question design that limits opportunity for plagiarism.
- Conduct pairwise comparison marking trials supported by online digital tools.
- Audit/ensure retailers are selling devices that are 'examination ready'.
- Audit/ensure equity of access to devices - all students need access to 'examination ready' devices over a period of time so that they are familiar with their operation and how to express ideas on these.

5.5 Ideas for investigation

Possible areas for future research are identified below:

- What shifts in student learning attitudes, confidence and motivation, and achievement are apparent with digital, when ready or midyear external assessment formats?
- What pedagogical shifts will be necessary in new assessment environments (digital assessments and NCEA review)?
- Collecting and analysing teacher and student voice using surveys and focus groups on the advantages/disadvantages and learning implications of offering more than one assessment opportunity per year will be important.

- Stakeholder expectations: What are the expectations of students, whānau and community for flexibility and mobility of qualification? What are the expectations of alignment with pedagogy and current tertiary/business sector use of digital technologies for timing and type of digital assessment (Rinehart, 2019)?
- Conducting when ready e-assessment trials e.g., for performance or portfolio work including development of tools for producing and curation of evidence, when ready submission processes and marking/moderation trials.
- There are effective systems in place to ensure the consistency, quality and effectiveness of the way schools administer internal NCEA assessment (NZQA, n.d.-g). However, there are factors that contribute to higher rates of achievement in internal assessments. These include: the ability to give specified, level-appropriate guidance or direction to students; task and marking variability in internal achievement standards due to the ability to select an assessment task and mode of evidence collection that best suits any one group of learners; the ability to offer when ready assessment, i.e., once a student has developed the relevant disciplinary literacy skills to enable them to cope with the literacy demands of the standard or has completed the assessment; and reassessment opportunities. Does increasing the number of opportunities for external standards close the gap between internal and external standards?

References

Absolum, M., Flockton, L., Hattie, J., Hipkins, R., & Reid, I. (2009). *Directions for assessment in New Zealand: Developing students' assessment capabilities*. Ministry of Education. https://assessment.tki.org.nz/Media/Files/Directions-for-Assessment-in-New-Zealand

Alberta Education. (n.d.). *Using technology to administer provincial achievement tests and diploma exams*. https://tinyurl.com/y76er87r

Aldowah, H., Al-Samarraie, H., & Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telematics* and *Informatics*, *37*, 13–49. https://doi.org/10.1016/j.tele.2019.01.007

American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (USA). (2014). *The standards for educational and psychological testing.* AERA. https://www.apa.org/science/programs/testing/standards

Angus, S. D., & Watson, J. (2009). Does regular online testing enhance student learning in the numerical sciences? Robust evidence from a large data set. *British Journal of Educational Technology*, *40*(2), 255–272. https://doi.org/10.1111/j.1467-8535.2008.00916.x

Baak, M., Miller, E., Sullivan, A., & Heugh, K. (2020). Tensions between policy aspirations and enactment: Assessment and inclusion for refugee background students. *Journal of Education Policy*, 1–19. https://doi.org/10.1080/02680939.2020.1739339

Backes, B., & Cowan, J. (2019). Is the pen mightier than the keyboard? The effect of online testing on measured student achievement. *Economics of Education Review*, *68*, 89–103. https://doi.org/10.1016/j.econedurev.2018.12.007

Baird, J. A., Meadows, M., Leckie, G., & Caro, D. (2017). Rater accuracy and training group effects in expert- and supervisor-based monitoring systems. *Assessment in Education: Principles, Policy & Practice*, *24*(1), 44–59. https://doi.org/10.1080/0969594X.2015.1108283

Barana, A., Conte, A., Fissore, C., Marchisio, M., & Rabellino, S. (2019). Learning analytics to improve formative assessment strategies. *Journal of E-Learning and Knowledge Society*, *15*(3), 75–88. https://doi.org/10.20368/1971-8829/1135057

Bargh, S. (2011). The future of secondary school examinations: The use of technology in the New Zealand context. In M. Hodis, & S. Kaiser (Eds.), *The future of secondary school examinations: The use of technology in the New Zealand context* (pp. 41–56). Victoria University. https://www.wgtn.ac.nz/education/pdf/jhc-symposium/Proceedings-of-the-Symposium-on-Assessment-and-Learner-Outcomes.pdf

Başol, G., & Johanson, G. (2009). Effectiveness of frequent testing over achievement: A meta-analysis study. *Journal of Human Sciences*, *6*(2), 99–121.

Beagley, J. E., & Capaldi, M. (2016). The effect of cumulative tests on the final exam. *PRIMUS*, *26*(9), 878–888. https://doi.org/10.1080/10511970.2016.1194343

Becker, K. A., & Bergstrom, B. A. (2013). Test administration models. *Practical Assessment, Research and Evaluation*, *18*(14), 1–7. https://doi.org/10.7275/pntr-yz21

Begnum, M. E. N., & Foss-Pedersen, R. J. (2018). Digital assessment in higher education. *Universal Access in the Information Society*, *17*(4), 791–810. https://doi.org/10.1007/s10209-016-0513-9

Bell, B., & Cowie, B. (2001). The characteristics of formative assessment in science education. *Science Education*, *85*(5), 536–553. https://doi.org/10.1002/sce.1022

Bergmann, E. (2014). *An examination of the relationship between the frequency of standardized testing and academic achievement* [Doctoral dissertation. University of Oregon]. Scholarbank. http://hdl.handle.net/1794/18351

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201–210. https://doi.org/10.3758/BF03193441

Biggs, J. & Tang, C. (2011). *Teaching for quality learning at university*. Berkshire, England: McGraw-Hill.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Springer. https://doi.org/10.1007/978-94-007-2324-5_2

Bird, S. (2010). Effects of distributed practice on the acquisition of second language English syntax. *Applied Psycholinguistics*, *31*(4), 635–650. https://doi.org/10.1017/S0142716410000172

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Bolstad, R., & Gilbert, J. (2012). *Supporting future-oriented learning and teaching: A New Zealand perspective.* Ministry of Education. https://www.educationcounts.govt.nz/__data/assets/pdf_file/0003/109317/994_Future-oriented-07062012.pdf

Boyle, A. (2010). *Regulatory research into on-demand testing* (p. 29). Office of Qualifications and Examinations Regulation. https://dera.ioe.ac.uk/1066/1/Ofqual-10-4725-Regulatory-research-into-on-demand_testing-2010-03-08.pdf

British Columbia Ministry of Education. (2019). *Grade 10 graduation numeracy assessment: Specifications. English language version*. https://curriculum.gov.bc.ca/sites/curriculum.gov.bc.ca/files/pdf/Sample_GLA10_Assessment_Key_and_Rubrics_and_Scoring_Guides.pdf

Broadbent, J., Panadero, E., & Boud, D. (2018). Implementing summative assessment with a formative flavour: A case study in a large class. *Assessment & Evaluation in Higher Education*, *43*(2), 307–322. https://doi.org/10.1080/02602938.2017.1343455

Brown, G. T. L. (2019). Technologies and infrastructure: Costs and obstacles in developing large-scale computer–based testing. *Education Inquiry*, *10*(1), 4–20. https://doi.org/10.1080/20004508.2018.1529528

Buchan, J. F., & Swann, M. (2007). A bridge too far or a bridge to the future? A case study in online assessment at Charles Sturt University. *Australasian Journal of Educational Technology*, *23*(3), 408–434. https://doi.org/10.14742/ajet.1260

Campbell, J. R., Robinson, W., Neelands, J., Hewston, R., & Mazzoli, L. (2007). Personalised learning: Ambiguities in theory and practice. *British Journal of Educational Studies*, *55*(2), 135–154. https://www.jstor.org/stable/4620550

Carpenter, R., & Alloway, T. (2018). Computer versus paper-based testing: Are they equivalent when it comes to working memory? *Journal of Psychoeducational Assessment*, *37*(3), 382–394. https://doi.org/10.1177/0734282918761496

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review*, *24*(3), 369–378. https://doi.org/10.1007/s10648-012-9205-z

Chang, C. C., Tseng, K. H., & Lou, S. J. (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a Web-based portfolio assessment environment for high school students. *Computers & Education*, *58*(1), 303–320. https://doi.org/10.1016/j.compedu.2011.08.005

Cheliminsky, E., & York, R. L. (1994). *Educational testing: The Canadian experience with standards, examinations, and assessments.* General Accounting Office.

Chen, O., Castro-Alonso, J. C., Paas, F., & Sweller, J. (2018). Extending cognitive load theory to incorporate working memory resource depletion: Evidence from the spacing effect. *Educational Psychology Review*, *30*(2), 483–501. https://doi.org/10.1007/s10648-017-9426-2

Chian, M. M., Bridges, S. M., & Lo, E. C. M. (2019). The triple jump in problem-based learning: Unpacking principles and practices in designing assessment for curriculum alignment.

Interdisciplinary Journal of Problem-Based Learning, *13*(2). https://doi.org/10.7771/1541-5015.1813

Chua, Y. P., & Don, Z. M. (2013). Effects of computer-based educational achievement test on test performance and test takers' motivation. *Computers in Human Behavior*, *29*(5), 1889–1895. https://doi.org/10.1016/j.chb.2013.03.008

Cowie, B., Hipkins, R., Keown, P., & Boyd, S. (2011). *The shape of curriculum change.* New Zealand Council for Educational Research. https://www-nzcer-org-nz.ezproxy.waikato.ac.nz/research/publications/shape-curriculum-change

Cowie, B., & Penney, D. (2016). Challenges, tensions and possibilities: An analysis of assessment policy and practice in New Zealand. In S. Scott, D. Scott, & C. Webber (Eds.), *Leadership of assessment, inclusion, and learning* (pp. 287–304). Springer. http://link.springer.com/chapter/10.1007/978-3-319-23347-5_12

Cramp, J., Medlin, J. F., Lake, P., & Sharp, C. (2019). Lessons learned from implementing remotely invigilated online exams. *Journal of University Teaching & Learning Practice*, *16*(1). https://ro.uow.edu.au/jutlp/vol16/iss1/10

Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*(4), 438–481. https://doi.org/10.3102/00346543058004438

Crooks, T. (1993). *Principles to guide assessment practice*. Higher Education Development Centre, University of Otago.

Crooks, T. (2011). Assessment for learning in the accountability era: New Zealand. *Studies in Educational Evaluation*, *37*(1), 71–77. https://doi.org/10.1016/j.stueduc.2011.03.002

Csapó, B., & Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in Psychology*, *10*, 1522. https://doi.org/10.3389/fpsyg.2019.01522

Darr, C. (2019). Assessment news. *Set: Research Information for Teachers*, *3*, 57–61. https://doi.org/10.18296/set.0153

Dawson, P. (2016). Five ways to hack and cheat with bring-your-own-device electronic examinations: Five ways to hack and cheat with BYOD e-exams. *British Journal of Educational Technology*, *47*(4), 592–600. https://doi.org/10.1111/bjet.12246

Dawson, S., & Siemens, G. (2014). Analytics to literacies: The development of a learning analytics framework for multiliteracies assessment. *International Review of Research in Open and Distance Learning*, *15*(4), 284–305. https://eric.ed.gov/?id=EJ1039824

Day, I. N. Z., van Blankenstein, F. M., Westenberg, M., & Admiraal, W. (2018). A review of the characteristics of intermediate assessment and their relationship with student grades. *Assessment & Evaluation in Higher Education*, *43*(6), 908–929. https://doi.org/10.1080/02602938.2017.1417974

Deed, C., Blake, D., Henriksen, J., Mooney, A., Prain, V., Tytler, R., Zitzlaff, T., Edwards, M., Emery, S., Muir, T., Swabey, K., Thomas, D., Farrelly, C., Lovejoy, V., Meyers, N., & Fingland, D. (2019). Teacher adaptation to flexible learning environments. *Learning Environments Research*. https://doi.org/10.1007/s10984-019-09302-0

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. https://educationaltechnology.net/ijet/index.php/ijet/article/view/47

Dosch, M. P. (2012). Practice in computer-based testing improves scores on the national certification examination for nurse anaesthetists. *AANA Journal*, *80*(4), S60–S66. https://www.aana.com/docs/default-source/aana-journal-web-documents-1/pract-comp-test-scores-nat-cert-ex-na-0812-ps60-s66.pdf?sfvrsn=e8894bb1_4

East, M. (2014). Working for positive outcomes? The standards–curriculum alignment for learning languages, and its reception by teachers. *Assessment Matters*, *6*, 65–85. https://www-nzcer-org-nz.ezproxy.waikato.ac.nz/nzcerpress/assessment-matters/articles/working-positive-outcomes-standards-curriculum-alignment

Education Review Office. (2016). *The collection and use of assessment information: Good practice in secondary schools*. Education Review Office. https://www.ero.govt.nz/publications/the-collection-and-use-of-assessment-information-good-practice-in-secondary-schools/

Educational Policy Improvement Center, & Conley, D. (2015). *A new era for educational assessment*. Education Policy Analysis Archives. https://doi.org/10.14507/epaa.v23.1983

Facer, K. (2011). *Learning futures: Education, technology and social change*. Routledge. https://doi.org/10.4324/9780203817308

Fensham, P., & Rennie, L. (2013). Towards an authentically assessed science curriculum. In D. Corrigan, R, Gunstone, & A. Jones (Eds.), *Valuing assessment in science education: Pedagogy, curriculum, policy* (pp. 69–100). Springer. https://doi.org/10.1007/978-94-007-6668-6

Figueroa-Cañas, J., & Sancho-Vinuesa, T. (2018). Investigating the relationship between optional quizzes and final exam performance in a fully asynchronous online calculus module. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2018.1559864

Flórez Petour, M., Assael, T., Gysling, J., & Astorga, J. (2018). The consequences of metrics for social justice: tensions, pending issues, and questions. *Oxford Review of Education, 44*(5), 651-667. https://doi.org/10.1080/03054985.2018.1500356

Fluck, A. (2019). An international review of eExam technologies and impact. *Computers & Education*, *132*(April), 1–15. https://doi.org/10.1016/j.compedu.2018.12.008

Fluck, A., & Hillier, M. (2016). *Innovative assessment with eExams* [paper presentation]. Australian Council for Computers in Education Conference, Brisbane, Australia. https://www.researchgate.net/publication/314352356_Innovative_assessment_with_eExams

Frankl, G. (2018, November). *The Austrian experience with e-Exams*. Paper presented at the Transforming assessment e-Exam Symposium, Melbourne, Australia. http://transformingassessment.com/e-exam_symposium_2018.php

Friyatmi, Mardapi, D., Haryanto, & Rahmi, E. (2020). The development of computerized economics item banking for classroom and school-based assessment. *European Journal of Educational Research*, *9*(1), 293–303. https://www.eu-jer.com/EU-JER_9_1_293.pdf

Frost, P. (2010). *Administration of the Victorian Certificate of Education. Victorian Auditor-General's Report*. https://www.parliament.vic.gov.au/papers/govpub/VPARL2006-10No319.pdf

Gharib, A., Phillips, W., & Mathew, N. (2012). Cheat sheet or open-book? A comparison of the effects of exam types on performance, retention, and anxiety. *Psychology Research, 2*(8), 469–478. https://doi.org/10.17265/2159-5542/2012.08.004

Gierl, M. J., Lai, H., & Li, J. (2013). Identifying differential item functioning in multi-stage computer adaptive testing. *Educational Research and Evaluation*, *19*(2–3), 188–203. https://doi.org/10.1080/13803611.2013.767622

Gokcora, D., & DePaulo, D. (2018). Frequent quizzes and student improvement of reading: A pilot study in a community college setting: *SAGE Open, April-July*, pp.1–9. https://doi.org/10.1177/2158244018782580

Gonski, D., Arcus, T., Boston, K., Gould, V., Johnson, W., O'Brien, L., & Roberts, M. (2018). *Through growth to achievement: Report of the review to achieve educational excellence in Australian schools.* Canberra: Commonwealth of Australia. https://docs.education.gov.au/node/50516

Government of Alberta. (2019). *General information bulletin: Diploma examinations program 2019–2020*. https://open.alberta.ca/publications/0846-7250

Griffin, P., & Care, E. (2015). *Assessment and teaching of 21st century* skills*: Methods and approach*. Springer. https://doi.org/10.1007/978-94-017-9395-7

Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 1–15). Springer. https://doi.org/10.1007/978-94-007-2324-5

Guha, R., Wagner, T., Darling-Hammond, L., Taylor, T., & Curtis, D. (2018). *The promise of performance assessments: Innovations in high school learning and college admission*. New York Performance Standards Consortium. http://www.performanceassessment.org/research

Hagen-Zanker, J., & Mallet, R. (2013). *How to do a rigorous, evidence-focused literature review in international development.* https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/8572.pdf

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Hamhuis, E., Glas, C., & Meelissen, M. (2020). Tablet assessment in primary education: Are there performance differences between TIMSS' paper-and-pencil test and tablet test among Dutch grade-four students? *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.12914

Harju, A. (2015). *"I trust the old way": Opinions and attitudes towards digitalizing the matriculation examination of English*. https://jyx.jyu.fi/handle/123456789/46096

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*(3), 245–270. https://doi.org/10.1080/02671520500193744

Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice, 10*(2), 169-207. https://doi.org/10.1080/0969594032000121270

Harris, L. R., Dargusch, J. (2020). Catering for diversity in the digital age: Reconsidering equity in assessment practices. In M. Bearman, P. Dawson, R. Ajjawi, J. Tai, & D. Boud (Eds.), *Re-imagining university assessment in a digital world. The Enabling Power of Assessment*, (vol. 7, pp. 95-110). Springer. https://doi.org/10.1007/978-3-030-41956-1_8

He, Q. (2012). On-demand testing and maintaining standards for general qualifications in the UK using item response theory: Possibilities and challenges. *Educational Research*, *54*(1), 89–112. http://dx.doi.org.ezproxy.waikato.ac.nz/10.1080/00131881.2012.658201

Herrmann-Abell, C. F., Hardcastle, J., & DeBoer, G. E. (2018). *Comparability of computer-based and paper-based science assessments* [Conference paper] NARST Annual International Conference, Atlanta, GA. https://files.eric.ed.gov/fulltext/ED581626.pdf

Hewson, C., & Charlton, J. P. (2019). An investigation of the validity of course-based online assessment methods: The role of computer-related attitudes and assessment mode preferences. *Journal of Computer Assisted Learning*, *35*(1), 51–60. https://doi.org/10.1111/jcal.12310

Hillier, M. (2018, November). *Hands on! Technology for moving from paper to authentic e-assessment.* Paper presented at the Transforming assessment e-Exam Symposium, Melbourne, Australia. http://transformingassessment.com/e-exam_symposium_2018.php

Hillier, M., Bower, M., Cowling, M., Fluck, A., Geer, R., Grant, S., Harris, B., Howah, K., Meacheam, D., McGrath, D., Pagram, J., & White, B. (2019). *Transforming exams across Australia: Processes and platform for e-exams in high stakes, supervised environments.* http://search.informit.com.au/documentSummary;res=AEIPT;dn=223302

Hipkins, R. (2005). The NCEA in the context of the knowledge society and national policy expectations. *New Zealand Annual Review of Education*, *14*, 27–38. https://www.nzcer.org.nz/research/publications/ncea-context-knowledge-society-and-national-policy-expectations

Hipkins, R. (2015). *Learning to learn in secondary classrooms*. New Zealand Council for Educational Research. https://www.nzcer.org.nz/research/publications/learning-learn-secondary-classrooms

Hipkins, R., Johnston, M., & Sheehan, M. (2016). *NCEA in context*. NZCER Press. https://www.nzcer.org.nz/nzcerpress/ncea-context

Hipkins, R., & Cameron, M. (2018). *Trends in assessment: An overview of themes in the literature.* New Zealand Council for Educational Research. https://www.nzcer.org.nz/research/publications/trends-assessment-overview-themes-literature

Hopfenbeck, T. N. (2015). Formative assessment, grading and teacher judgement in times of change. *Assessment in Education: Principles, Policy & Practice, 22*(3), 299-301. https://doi.org/10.1080/0969594X.2015.1050261

Houston, D., & Thompson, J. N. (2017). Blending formative and summative assessment in a capstone subject: 'It's not your tools, it's how you use them'. *Journal of University Teaching and Learning Practice*, *14*(3). https://files.eric.ed.gov/fulltext/EJ1170183.pdf

Hume, A., & Coll, R. (2010). Authentic student inquiry: The mismatch between the intended curriculum and the student-experienced curriculum. *Research in Science & Technological Education, 28*(1), 43–62. https://doi.org/10.1080/02635140903513565

Hung, S-T. A. (2012). A washback study on e-portfolio assessment in an English as a foreign language teacher preparation program. *Computer Assisted Language Learning*, *25*(1), 21–36. https://doi.org/10.1080/09588221.2010.551756

Ilomäki, L., & Lakkala, M. (2018). Digital technology and practices for school improvement: Innovative digital school model. *Research and Practice in Technology Enhanced Learning*, *13*(25), 1–32. https://doi.org/10.1186/s41039-018-0094-8

Ingram, J., Elliott, V., Morin, C., Randhawa, A., & Brown, C. (2018). Playing the system: incentives to 'game' and educational ethics in school examination entry policies in England. *Oxford Review of Education*, *44*(5), 545-562. https//doi.org/10.1080/03054985.2018.1496906

Irwin, B., & Hepplestone, S. (2012). Examining increased flexibility in assessment formats. *Assessment & Evaluation in Higher Education, 37*(12), 773–785. http://www.tandfonline.com/doi/abs/10.1080/02602938.2011.573842

Istiyono, E., Dwandaru, W. S. B., Setiawan, R., & Megawati, I. (2020). Developing of computerized adaptive testing to measure physics higher order thinking skills of senior high school students and its feasibility of use. *European Journal of Educational Research*, *9*(1), 91–101. https://www.eu-jer.com/EU-JER_9_1_91.pdf

Jackel, B. (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, & A. Delaney (Eds.), *Diagrammatic representation and inference* (vol. 8578, pp. 71–77). Springer. https://doi.org/10.1007/978-3-662-44043-8_12

Jeong, H. (2012). A comparative study of scores on computer-based tests and paper-based tests. *Behaviour & Information Technology*, *33*(4), 410–422. *http*://www.tandfonline.com/doi/abs/10.1080/0144929X.2012.710647

Johnson, M., Maguire, J., & Wood, A. (2017). *Digital technologies-in-schools-2016-17*. 2020 Trust. https://2020.org.nz/wp-content/uploads/2014/05/Digital-Technologies-in-Schools-2016-17-04-05-2017-FINAL.pdf

Johnston, M., Hipkins, R., & Sheehan, M. (2017). Building epistemic thinking through disciplinary inquiry: Contrasting lessons from history and biology. *Curriculum Matters*, *13*, 80–102. https://doi.org/10.18296/cm.0020

Kable, A. K., Pich, J., & Maslin-Prothero, S. E. (2012). A structured approach to documenting a search strategy for publication: A 12 step guideline for authors. *Nurse Education Today*, *32*(8), 878–886. https://doi.org/10.1016/j.nedt.2012.02.022

Kennedy, A. B., & Fiester, S. E. (2020). Reduced performance on examinations following untimed short assessments. *Studies in Educational Evaluation*, *65(*100835). https://doi.org/10.1016/j.stueduc.2019.100835

Klinger, D. A., Deluca, C. & Miller, S. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration and Policy*, *76*, 1-34. http://www.umanitoba.ca/publications/cjeap/pdf_files/klinger.pdf

Kimura, T. (2017). The impacts of computer adaptive testing from a variety of perspectives. *Journal of Educational Evaluation for Health Professions*, *14*(12). https://doi.org/10.3352/jeehp.2017.14.12

Klinger, D. A., McDivitt, P., Howard, B., Rogers, T., Munoz, M., & Wylie, C. (2015). *Classroom assessment standards for PreK-12 teachers.* Amazon Digital Services.

Kōrero Mātauranga. (2020). *Review of Achievement Standards (RAS) - Pilot phase: Science.* https://consultation.education.govt.nz/ncea/sector-feedback-science/

Kupiainen, S., Marjanen, J., & Hautamäki, J. (2016). The problem posed by exam choice on the comparability of results in the Finnish matriculation examination. *Journal for Educational Research Online*, *8*(2), 87–106. http://hdl.handle.net/10138/232733

Ladyshewsky, R. K. (2015). Post-graduate student performance in 'supervised in-class' vs. 'unsupervised online' multiple choice tests: Implications for cheating and test security. *Assessment & Evaluation in Higher Education*, *40*(7), 883–897. https://doi.org/10.1080/02602938.2014.956683

Leadbeater, C. (n.d.). *The future of public services: Personalised learning*. OECD. https://tinyurl.com/yavnlokv

Leithner, A. (2011). Do student learning styles translate to different "testing styles"? *Journal of Political Science Education*, *7*(4), 416–433. https://doi.org/10.1080/15512169.2011.615195

Ling, G. (2016). Does it matter whether one takes a test on an iPad or a desktop computer? *International Journal of Testing*, *16*(4), 352–377. https://doi.org/10.1080/15305058.2016.1160097

Liu, C., Han, K. T., & Li, J. (2019). Compromised item detection for computerized adaptive testing. *Frontiers in Psychology*, *10*(829). https://doi.org/10.3389/fpsyg.2019.00829

Lovett, B. J., Lewandowski, L. J., Berger, C., & Gathje, R. A. (2010). Effects of response mode and time allotment on college students' writing. *Journal of College Reading and Learning*, *40*(2), 64–79. https://doi.org/10.1080/10790195.2010.10850331

Maclean, G., & McKeown, P. (2013). Comparing online quizzes and take-home assignments as formative assessments in a 100-level economics course. *New Zealand Economic Papers*, *47*(3), 245–256. https://doi.org/10.1080/00779954.2012.707530

Marden, N. Y., Ulman, L. G., Wilson, F. S., & Velan, G. M. (2013). Online feedback assessments in physiology: Effects on students' learning experiences and outcomes. *Advances in Physiology Education*, *37*(2), 192–200. https://doi.org/10.1152/advan.00092.2012

Masters, G. N. (2013). *Reforming educational assessment: Imperatives, principles and challenges*. Australian Education Review. https://research.acer.edu.au/aer/12

Masters, G. N. (2017). *Assessment online: Informing teaching and learning*. ACER. https://research.acer.edu.au/columnists/28

May, S. (n.d.). *Assessment: what are the cultural issues in relation to Pasifika, Asian, ESOL, immigrant and refugee learners?* TKI. https://assessment.tki.org.nz/Media/Files/May-S.-Assessment-what-are-the-cultural-issues-in-relation-to-Pasifika-Asian-ESOL-immigrant-and-refugee-learners-University-of-Waikato

McEwen, N. (1995). Accountability in education in Canada. *Canadian Journal of Education, 20,* 1-17.

Mc Tiernan, K., Smith, M., & Walsh, I. (2007). The 'triple jump' assessment in problem based learning: An evaluative method used in the appraisal of both knowledge acquisition and problem solving skills. In G. O'Neill, S. Huntley-Moore, & P. Race (Eds.), *Case studies of good practice in assessment of student earning in* higher *education* (pp. 116– 119). AISHE. http://www.tara.tcd.ie/handle/2262/60450

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger III, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*(2), 399–414. https://doi.org/10.1037/a0021782

McMahon, S., & Jones, I. (2014). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice, 22*(3), 368–389. http://www.tandfonline.com/doi/abs/10.1080/0969594X.2014.978839

McManus, R. (2016). Assessment timing: Student preferences and its impact on performance. *Practitioner Research in Higher Education*, *10*(1), 203–216. https://files.eric.ed.gov/fulltext/EJ1130071.pdf

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher 18(*2), 5–11. https://doi.org/10.3102/0013189X018002005

Michael, T. B., & Williams, M. A. (2013). Student equity: Discouraging cheating in online courses. *Administrative Issues Journal: Education, Practice & Research*, *3*(2), 30–41. https://doi.org/10.5929/2013.3.2.8

Ministry of Education. (n.d.). *Assessment online: Reliability and validity*. TKI. https://assessment.tki.org.nz/Using-evidence-for-learning/Working-with-data/Concepts/Reliability-and-validity

Ministry of Education. (1993). *The New Zealand curriculum framework*. Ministry of Education.

Ministry of Education. (1994). *Assessment: Policy to practice*. Learning Media.

Ministry of Education. (2007). *The New Zealand curriculum*. TKI. http://nzcurriculum.tki.org.nz/The-New-Zealand-Curriculum

Ministry of Education. (2011). *Ministry of Education position paper: Assessment*. TKI. https://assessment.tki.org.nz/Media/Files/Ministry-of-Education-Position-Paper-Assessment-Schooling-Sector-2011

Ministry of Education and Culture, Finland. (n.d.). *Finnish matriculation examination*. Opetus- Ja Kulttuuriministeriö. https://minedu.fi/en/finnish-matriculation-examination

Moeed, A. (2010). Teaching to investigate in Year 11 science, constrained by assessment. *The New Zealand Annual Review of Education*, *20*, 74–101. https://doi.org/10.26686/nzaroe.v0i20.1571

Moulton, C. A., Dubrowski, A., Macrae, H., Graham, B., Grober, E., & Reznick, R. (2006). Teaching surgical skills: what kind of practice makes perfect?: randomized, controlled trial. *Annals of surgery*, *244*(3), 400–409. https://doi.org/10.1097/01.sla.0000234808.85789.6a

Murchan, D., & Oldham, E. (2017). Exploring the role of computer-based assessment in diagnosing children's mathematical errors in primary education in Ireland. *Irish Educational Studies 36*(4), 489–510. https://doi.org/10.1080/03323315.2017.1393765

Murgatroyd, S. (2018). New approaches to the assessment of learning: New possibilities for business education. In A. Khare, & D. Hurst (Eds.), *On the line* (pp, 141–155). Springer. https://doi.org/10.1007/978-3-319-62776-2_12

Mutch, C. (2012). Assessment for, of and as learning: Developing a sustainable assessment culture in New Zealand schools. *Policy Futures in Education*, *10*(4), 374–385. https://doi.org/10.2304/pfie.2012.10.4.374

Nardi, A., & Ranieri, M. (2019). Comparing paper-based and electronic multiple-choice examinations with personal devices: Impact on students' performance, self-efficacy and satisfaction. *British Journal of Educational Technology*, *50*(3), 1495–1506. https://doi.org/10.1111/bjet.12644

New Zealand Government. (n.d.-a). *NCEA review*. Kōrero Matauranga: NCEA Review. https://conversation.education.govt.nz/conversations/ncea-review/

New Zealand Government. (n.d.-b). *Review of Achievement Standards (RAS)*. https://conversation.education.govt.nz/conversations/ncea-review/review-of-achievement-standards/

New Zealand Government. (2019). *NCEA change package 2019 overview* (p. 27). https://conversation.education.govt.nz/assets/Uploads/NCEA-Change-Package-2019-Web.pdf

New Zealand Qualifications Authority (NZQA). (n.d.-a). *Digital assessment vision: A design principles approach*. https://www.nzqa.govt.nz/about-us/future-state/digital-assessment-vision/

New Zealand Qualifications Authority (NZQA). (n.d.-b). *How NCEA works*. https://www.nzqa.govt.nz/ncea/understanding-ncea/how-ncea-works/
New Zealand Qualifications Authority (NZQA). (n.d.-c). *Managing national assessment in schools*. https://www.nzqa.govt.nz/providers-partners/assessment-and-moderation-of-standards/managing-national-assessment-in-schools/
New Zealand Qualifications Authority (NZQA). (n.d.-d). *NCEA endorsements*. https://www.nzqa.govt.nz/ncea/understanding-ncea/how-ncea-works/endorsements/
New Zealand Qualifications Authority (NZQA). (n.d.-e). *Reviews and reconsiderations*. https://www.nzqa.govt.nz/ncea/ncea-results/reviews-and-reconsiderations/
New Zealand Qualifications Authority (NZQA). (n.d.-f). *Assessment opportunities in schools*. https://tinyurl.com/yaszk7tg
New Zealand Qualifications Authority (NZQA). (n.d.-g). University entrance. https://www.nzqa.govt.nz/qualifications-standards/awards/university-entrance/
New Zealand Qualifications Authority (NZQA). (2018). Kia noho takatū ki tō Āmua Ao–Qualify for the future world. https://www.nzqa.govt.nz/assets/About-us/Future-State/Innovation-at-NZQA/Qualify-for-the-Future-World.pdf
New Zealand Qualifications Authority (NZQA). (2019). Annual report: NCEA, university entrance, and NZ scholarship data and statistics. https://www.nzqa.govt.nz/assets/About-us/Publications/stats-reports/ncea-annual-report-2018.pdf
Newhouse, C. P. (2013). Computer-based practical exams in an applied information technology course. *Journal of Research on Technology in Education*, *45*(3), 263–286. https://doi.org/10.1080/15391523.2013.10782606
Newhouse, C. P. (2014). *Digital portfolios for summative assessment* [Paper presentation]. ACEC 2014: Australian Computers in Education Conference: Now It, Adelaide, Australia. https://preview.tinyurl.com/y7y56zgb
Newhouse, C. P. (2016). Digital forms of assessment in schools: Supporting the processes to improve outcomes. In M. Pector, B. Lockee, & M. Childress (Eds.), *Learning, design, and technology* (pp 1–29). Springer. https://doi.org/10.1007/978-3-319-17727-4_41-1
Newhouse, C. P., & Tarricone, P. (2014). Digitizing practical production work for high-stakes assessments. *Canadian Journal of Learning and Technology*, *40*(2), 1–17. https://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1650&context=ecuworkspost2013
Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Galbraith, R., Hays, R., Kent, A., Perrott, V., & Roberts, T. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, *33*(3), 206–214. https://www.tandfonline.com/doi/abs/10.3109/0142159X.2011.551559
Nyland, R., Davies, R. S., Chapman, J., & Allen, G. (2017). Transaction-level learning analytics in online authentic assessments. *Journal of Computing in Higher Education*, *29*(2), 201–217. https://doi.org/10.1007/s12528-016-9122-0
Oduntan, O. E., Ojuawo, O. O., & Oduntan, E. A. (2015). *A comparative analysis of student performance in Paper Pencil Test (PPT) and Computer Based Test (CBT) examination system*. https://www.semanticscholar.org/paper/A-Comparative-Analysis-of-Student-Performance-in-(-O.E-O.O/f17af46973871d596f54d6a6a578ef6d7d7a7c3c
OECD. (2015). *Students, computers and learning: Making the connection*. https://doi.org/10.1787/9789264239555-en
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
Opposs, D., Baird, J., Chankseliani, M., Stobart, G., Kaushik, A., McManus, H., & Johnson, D. (2020). Governance structure and standard setting in educational assessment. *Assessment in Education: Principles, Policy & Practice*, *27*(2), 192–214. https://doi.org/10.1080/0969594X.2020.1730766

O'Sullivan, A. J., Harris, P., Hughes, C. S., Toohey, S. M., Balasooriya, C., Velan, G., Kumar, R. K., & McNeil, H. P. (2012). Linking assessment to undergraduate student capabilities through portfolio examination. *Assessment & Evaluation in Higher Education*, *37*(3), 379–391. https://doi.org/10.1080/02602938.2010.534766

Öz, H., & Özturan, T. (2018). Computer-based and paper-based testing: does the test administration mode influence the reliability and validity of achievement tests? *Journal of Language and Linguistic Studies*, *14*(1), 67–85. https://tinyurl.com/y8egrqcp

Paiva, R. C., Ferreira, M. S., & Frade, M. M. (2017). Intelligent tutorial system based on personalized system of instruction to teach or remind mathematical concepts. *Journal of Computer Assisted Learning*, *33*(4), 370–381. https://doi.org/10.1111/jcal.12186

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing*, *12*(1), 21–43. https://doi.org/10.1080/15305058.2011.602920

Polesel, J., Rice, S., & Dulfer, N. (2014). The impact of high-stakes testing on curriculum and pedagogy: A teacher perspective from Australia. *Journal of Education Policy 29*(5), 640–657. http://www.tandfonline.com/doi/abs/10.1080/02680939.2013.865082

Pollari, P. (2017). *(Dis)empowering assessment? : Assessment as experienced by students in their upper secondary school EFL studies*. University of Jyväskylä. https://jyx.jyu.fi/handle/123456789/55478

Government of Alberta. (2020). *Population statistics*. https://www.alberta.ca/population-statistics.aspx

Penney, D., Jones, A., Newhouse, P., & Campbell, A. (2012). Developing a digital assessment in senior secondary physical education. *Physical Education and Sport Pedagogy*, *17*(4), 383-410. https://doi.org/10.1080/17408989.2011.582490

Penney, D., Newhouse, P., Jones, A., & Campbell, A. (2012). Digital technologies: Enhancing pedagogy and extending opportunities for learning in senior secondary physical education? In T. Le & Q. Le (Eds.), *Technologies for enhancing pedagogy, engagement and empowerment in education: Creating learner-friendly environments* (pp. 15-26). Hershey, P.A.: CGI Global.

Pretorius, L., van Mourik, G. P., & Barratt, C. (2017). Student choice and higher-order thinking: Using a novel flexible assessment regime combined with critical thinking activities to encourage the development of higher order thinking. *International Journal of Teaching and Learning in Higher Education*, *29*(2), 389–401. https://files.eric.ed.gov/fulltext/EJ1146270.pdf

Price, T., Lynn, N., Coombes, L., Roberts, M., Gale, T., Bere, S. R. de, & Archer, J. (2018). The international landscape of medical licensing examinations: A typology derived from a systematic review. *International Journal of Health Policy and Management*, *7*(9), 782–790. https://doi.org/10.15171/ijhpm.2018.32

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

Puentedura, R. R. (2013). *The SAMR model explained*. https://www.youtube.com/watch?v=_QOsz4AaZ2k

Quest A+. (n.d.). *Practice test instructions.* https://questaplus.alberta.ca/help/Practice%20Test%20Instructions.pdf

Redecker, C., & Johannessen, Ø. (2013). Changing assessment - Towards a new Assessment paradigm using ICT. *European Journal of Education*, *48*(1), 79–96. https://doi.org/10.1111/ejed.12018

Reinertsen, N. (2018). Why can't it mark this one? A qualitative analysis of student writing rejected by an automated essay scoring system. *English in Australia 53*(1), 52–60. https://eric.ed.gov/?id=EJ1183097

Rezaei, A. R. (2015). Frequent collaborative quiz taking and conceptual learning: *Active Learning in Higher Education, 16*(3), 187–196. https://doi.org/10.1177/1469787415589627

Rideout, C. A. (2018). Students' choices and achievement in large undergraduate classes using a novel flexible assessment approach. *Assessment & Evaluation in Higher Education, 43*(1), 68–78. https://doi.org/10.1080/02602938.2017.1294144

Ripley, M. (2007). *E-assessment–an update on research, policy and practice*. Futurelab. https://www.immagic.com/eLibrary/ARCHIVES/GENERAL/FUTRLBUK/Assessment_Review_update.pdf

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*(6), 481–498. https://doi.org/10.1007/s11251-007-9015-8

SACE Board South Australia. (n.d.). *Electronic exams—South Australian Certificate of Education*. https://www.sace.sa.edu.au/about/sace-improvement/electronic-assessment/exams

SACE Board South Australia. (2018). *Special provisions in curriculum and assessment policy*. https://tinyurl.com/ya2slxjg

Savolainen, J. (2017). *Digitalize it! : Upper secondary school students' views on the digitalized matriculation examination* [Doctoral thesis, University of Jyväskylä]. JYX Digital Repository. https://jyx.jyu.fi/handle/123456789/53979

Scalise, K., Irvin, P. S., Alresheed, F., Zvoch, K., Yim-Dockery, H., Park, S., Landis, B., Meng, P., Kleinfelder, B., Halladay, L., & Partsafas, A. (2018). Accommodations in digital interactive STEM assessment tasks. *Journal of Special Education Technology*, *33*(4), 219–236. https://doi.org/10.1177/0162643418759340

Scott, E. P. (2012). Short-term gain at long-term cost? How resit policy can affect student learning. *Assessment in Education: Principles, Policy & Practice*, *19*(4), 431–449. https://doi.org/10.1080/0969594X.2012.714741

Sindre, G., & Vegendla, A. (2015). *E-exams versus paper exams: A comparative analysis of cheating-related security threats and countermeasure* [Paper presentation]. Norwegian Information Security Conference (NISK 2015), Norway. https://ojs.bibsys.no/index.php/NISK/article/view/298

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, *33*(1), 1–19. https://doi.org/10.1111/jcal.12172

Smaill, E. (2013). Moderating New Zealand's national standards: Teacher learning and assessment outcomes. *Assessment in Education: Principles, Policy & Practice*, *20*(3), 250–265. https://doi.org/10.1080/0969594X.2012.696241

Smaill, E, (2020). Using involvement in moderation to strengthen teachers' assessment for learning capability. *Assessment in Education: Principles, Policy & Practice*. https://doi.org/10.1080/0969594X.2020.1777087

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*(5), 763–767. https://doi.org/10.1002/acp.1747

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 553–573.

Sormunen, E. (2019). *Upper secondary school students' views on the new digitalized matriculation examination*. University of Jyväskylä. https://jyx.jyu.fi/handle/123456789/62848

Stödberg, U. (2012). A research review of e-assessment. *Assessment & Evaluation in Higher Education*, *37*(5), 591–604. https://doi.org/10.1080/02602938.2011.557496

Stotter, J., & Gillon, K. (2011). Inquiry learning with senior secondary students: Yes it can be done. *Access*, *25*(3), 14–19. https://search.informit.org/documentSummary;dn=341090467276495;res=IELHSS

Sulun, E., Nalbantoglu, E., & Oztug, E. K. (2018). The effect of exam frequency on academic success of undergraduate music students and comparison of students performance anxiety levels. *Quality & Quantity*, *52*(1), 737–752. https://doi.org/10.1007/s11135-017-0653-x

Tarricone, P., & Newhouse, C. P. (2016). A study of the use of pairwise comparison in the context of social online moderation. *The Australian Educational Researcher*, *43*(3), 273–288. https://doi.org/10.1007/s13384-015-0194-z

Tate, T. P., & Warschauer, M. (2019). Keypresses and mouse clicks: Analysis of the first national computer-based writing assessment. *Technology, Knowledge and Learning*, *24*(4), 523–543. https://doi.org/10.1007/s10758-019-09412-x

Terrell, J. (2016). *Getting it right: Guidelines for online assessment in New Zealand tertiary contexts*. Ako Aotearoa National Centre for Tertiary Teaching Excellence. https://ako.ac.nz/knowledge-centre/guidelines-for-online-assessment/

Tertiary Education Commission Te Amorangi Mātauranga Matua. (n.d.). About tertiary education organisations. https://www.tec.govt.nz/teo/working-with-teos/about-teos/

The Gordon Commission on the Future of Assessment in Education. (2013). *To assess, to teach, to learn: A vision for the future of assessment.* Princeton, NJ. http://www.gordoncommission.org/rsc/pdfs/gordon_commission_technical_report.pdf

Thille, C. M. (2016). *Bridging learning research and teaching practice for the public good: The Learning Engineer*. New York, NY: TIAA Institute. https://www.tiaa.org/public/pdf/bridging_learning_research_and_teaching_practice.pdf

Thomas, M. (2011*). Deconstructing digital natives: Young people, technology, and the new literacies.* New York, NY: Routledge.

Thompson, G. (2017). Computer adaptive testing, big data and algorithmic approaches to education. *British Journal of Sociology of Education*, *38*(6), 827–840. https://doi.org/10.1080/01425692.2016.1158640

Thorpe, V. (2012). Assessment rocks? The assessment of group composing for qualification. *Music Education Research*, *145*(4), 417–429. http://www.tandfonline.com/doi/abs/10.1080/14613808.2012.699957

Tian, T., DeMara, R. F., & Gao, S. (2019). Efficacy and perceptions of assessment digitization within a large-enrollment mechanical and aerospace engineering course. *Computer Applications in Engineering Education*, *27*(2), 419–429. https://doi.org/10.1002/cae.22086

Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2016). Rethinking assessment in a digital age: Opportunities, challenges and risks. *British Educational Research Journal*, *42*(3), 454–476. https://doi.org/10.1002/berj.3215

Timms, M. J. (2017). Assessment of online learning. In A. Marcus-Quinn & T. Hourigan (Eds.), *Handbook on digital learning for K-12 schools* (pp. 217–231). Springer. https://doi.org/10.1007/978-3-319-33808-8_13

Trask, S. (2019). *Repositioning teachers and learners in science assessment for 21st century learning* (Unpublished doctoral dissertation). The University of Waikato, Hamilton, New Zealand. https://researchcommons.waikato.ac.nz/handle/10289/12826

Vaessen, B. E., Beemt, A., van den Watering, G., van de Meeuwen, L. W., van Lemmens, L., & den Brok, P. (2017). Students' perception of frequent assessments and its relation to motivation and grades in a statistics course: A pilot study. *Assessment & Evaluation in Higher Education*, *42*(6), 872–886. https://doi.org/10.1080/02602938.2016.1204532

van der Vleuten, C. P. M. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education: Theory and Practice*, *11*, 41–67. https://doi.org/10.1007/BF00596229

van Groen, M. M., & Eggen, T. J. H. M. (2020). Educational test approaches: The suitability of computer-based test types for assessment and evaluation in formative and summative contexts. *Journal of Applied Testing Technology*, *21*(1), 12–24. http://www.jattjournal.com/index.php/atp/article/view/146484/103188

van Lent, G. (2009). Risks and benefits of CBT versus PBT in high-stakes testing. In F. Scheuermann & J. Björnsson (Eds.), *The transition to computer-based assessment: New approaches to skills assessment and implications for large-scale testing* (pp. 83–91). OPOCE.

Vikburg, T. (2018). *A national roll out of e-exams for high stakes Matriculation* [Oral presentation]: E-exam symposium 2018, Melbourne, Australia. https://www.youtube.com/watch?v=QLvGIdDeSmE

Vista, A., Care, E., & Griffin, P. (2015). A new approach towards marking large-scale complex assessments: Developing a distributed marking system that uses an automatically scaffolding and rubric-targeted interface for guided peer-review. *Assessing Writing*, *24*, 1–15. https://doi.org/10.1016/j.asw.2014.11.001

Vlach, H. A., & Sandhofer, C. M. (2012). Distributing learning over time: The spacing effect in children's acquisition and generalization of science concepts. *Child Development*, *83*(4), 1137–1144. https://doi.org/10.1111/j.1467-8624.2012.01781.x

Volante, L., & Ben Jaafar, S. (2008). Educational assessment in Canada. *Assessment in Education: Principles, Policy & Practice, 15*(2), 201-210, https://doi.org/10.1080/09695940802164226

von Heyking, A. (2019). *Alberta, Canada: How curriculum and assessment work in a plural school system*. John Hopkins School of Education. https://edpolicy.education.jhu.edu/wp-content/uploads/2019/06/Alberta-Brief.pdf

von Zansen, A. (2016). *Finnish matriculation examination computer-based: The impact of pictures and video in assessing listening comprehension*. University of Jyväskylä. https://jyx.jyu.fi/handle/123456789/60443

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

Wallace, C. S., & Priestley, M. R. (2017). Secondary science teachers as curriculum makers: Mapping and designing Scotland's new Curriculum for Excellence. *Journal of Research in Science Teaching*, *54*(3), 324–349. https://doi.org/10.1002/tea.21346

Wallace, P., & Clariana, R. B. (2005). Gender differences in computer-administered versus paper-based tests. *International Journal of Instructional Media*, *32*(2), 171-180.

Wei, H., & Lin, J. (2015). Using out-of-level items in computerized adaptive testing. *International Journal of Testing*, *15*(1), 50–70. https://doi.org/10.1080/15305058.2014.979492

Weiner, J. A., & Hurtz, G. M. (2017). A comparative study of online remote proctored versus onsite proctored high-stakes exams. *Journal of Applied Testing Technology*, *18*(1), 13–20. http://www.jattjournal.com/index.php/atp/article/view/113061

Weiss, D. J. (2011). Better data from better measurements using computerized adaptive testing. *Journal of Methods and Measurement in the Social Sciences*, *2*(1), 1–23. https://www.semanticscholar.org/paper/Better-Data-From-Better-Measurements-Using-Adaptive-Weiss/4638fa3d23e2375d795c94c7f440fe98d820ab20

Wesolowski, B. (2014). Documenting student learning in music performance: A framework. *Music Educators Journal*, *101*(1), 77–-85. https://doi.org/10.1177/0027432114540475

Wheadon, C., Whitehouse, C., Spalding, V., Tremain, K., & Charman, M. (2009). *Principles and practice of on-demand testing*. https://dera.ioe.ac.uk/9464/1/2009-01-principles-practice-on-demand-testing.pdf

Williams, P. J., & Penney, D. (2011). Authentic assessment in performance—based subjects. *Teachers and Curriculum*, *12*(1), 31–39. https://doi.org/10.15663/tandc.v12i1.28

Wilson, A., & McNaughton, S. (2014). Using selected NCEA standards to profile senior students' subject-area literacy. *SET, 4*, 61. https://www.nzcer.org.nz/nzcerpress/set/articles/using-selected-ncea-standards-profile-senior-students-subject-area-literacy

Wyatt–Smith, C., & Castleton, G. (2007). Examining how teachers judge student writing: An Australian case study. *Journal of Curriculum Studies*, *37*(2), 131–154. https://doi.org/10.1080/0022027032000242887

Wylie, C., & Bonne, L. (2016). *Secondary schools in 2015: Findings from the NZCER national survey*. New Zealand Council for Educational Research. https://doi.org/10.18296/rep.0001

Yamaguchi, R., & Hall, A. (2017). *Compendium of education technology research funded by NCER and NCSER: 2002–2014.* National Center for Education Research. https://eric.ed.gov/?id=ED573369

Yerushalmy, M., Nagari-Haddif, G., & Olsher, S. (2017). Design of tasks for online assessment that supports understanding of students' conceptions. *ZDM*, *49*(5), 701–716. https://doi.org/10.1007/s11858-017-0871-7

Ylioppilastutkintolautakunta Studentexamensnamden. (n.d.). *Digital Matriculation Examination*. https://www.ylioppilastutkinto.fi/en/matriculation-examination/digital-matriculation-examination

Zhang, Y., Wang, D., Gao, X., Cai, Y., & Tu, D. (2019). Development of a computerized adaptive testing for internet addiction. *Frontiers in Psychology*, *10*, 1–12. https://doi.org/10.3389/fpsyg.2019.01010