Trends in assessment:

An overview of themes in the literature

Rosemary Hipkins and Marie Cameron

NZCER
Rangahau Mātauranga o Aotearoa

Trends in assessment:

An overview of themes in the literature

Rosemary Hipkins and Marie Cameron

New Zealand Council for Educational Research

2018

NZCER
Rangahau Mātauranga o Aotearoa

New Zealand Council for Educational Research
Level 10, 178 Willis St
Wellington
New Zealand

www.nzcer.org.nz

ISBN 978-1-98-854249-2

Contents

1.
The context for this review

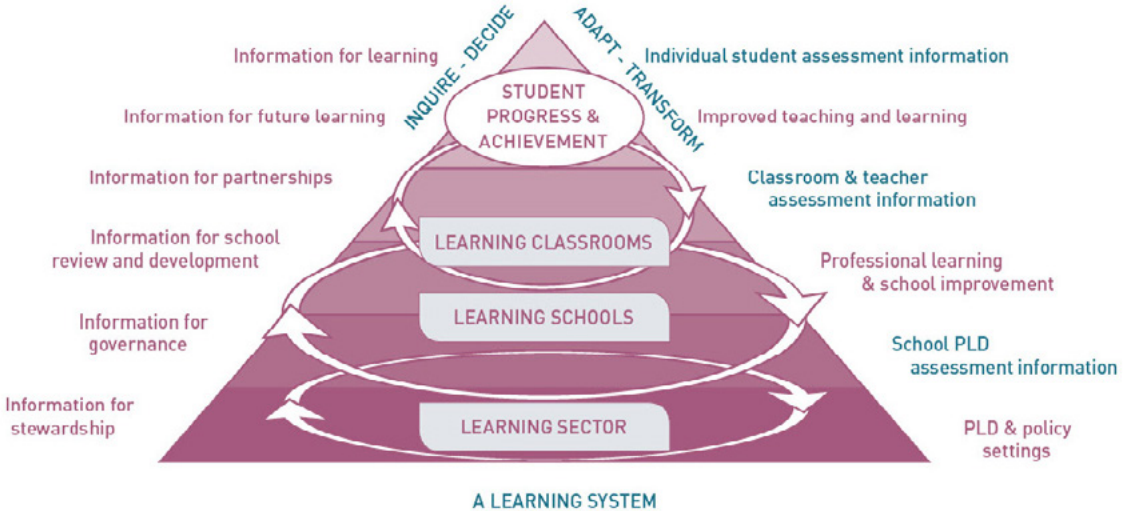
This report outlines findings from a literature review of trends in assessment policy and practice, carried out in June and July 2018. Funded by the Ministry of Education (the Ministry), the primary purpose for this work is to inform the Ministry as they are considering updating their position on assessment. This was last articulated in the position paper titled *Assessment (Schooling Sector) 2011* (Ministry of Education, 2011). This paper drew on a paper commissioned by the Ministry, titled *Directions for Assessment in New Zealand (DANZ): Developing students' assessment capabilities* (Absolum, Flockton, Hattie, Hipkins, & Reid, 2009). The central recommendation of the DANZ report, in brief, was that we should work to ensure the assessment system supports the growth in "assessment capability" of everyone with a stake in using assessment information, including teachers and students, but also parents, policy makers, and other stakeholders.

Given this recommendation, the idea of *assessment capability* has provided an important point of reference in the sections that follow. However the individual research papers tend to focus on one or at most several elements of the overall system. Therefore, the following brief discussion draws on the 2011 report to provide a system-level overview within which to locate the detail that follows.

The idea of an assessment system

Figure 1 on the next page is from the 2011 policy discussion (Ministry of Education, 2011, p. 16). This figure visually depicts the multi-layered nature of a national assessment system, where the ultimate goal of assessment is to provide support that enhances student learning. It also visually depicts the intertwined nature of assessment and *learning*. Purposeful assessment practices yield information that informs decision-making about learning at one or more levels of the system. In fact, the 2011 paper is unequivocal in stating that strengthened learning should be the goal, whatever the level at which assessment activity takes place: "Assessment not used formatively at some level of the system is not worth doing" (p. 15). A paper that reviews the overall assessment system in Scotland makes the same point in its title: "Assessment is learning: the preposition vanishes" (Hayward, 2015). The second part of this title refers to a distinction that is often made between assessment *for* learning, where the purpose is assumed to be formative, and assessment *of* learning, where the purpose is assumed to be summative.

FIGURE 1: **Assessment information in a schooling system that learns (Ministry of Education, 2011)**



It is unhelpful to make a binary distinction between formative and summative assessment purposes—learning, teaching and assessment are intertwined in a whole that needs to be carefully teased apart. In this way of thinking, information generated by *any* assessment can be used for formative or summative purposes, and those intended purposes must be heeded when evaluating any specific aspect of assessment practice. What does change according to the intended purpose is the nature of the information gathered. Rich data gathered at the classroom/individual level is appropriate for formative purposes, but will generally be aggregated and stripped of detail when used for summative purposes.

Keeping strong connections between the different layers of the assessment system depicted in Figure 1 has always been challenging and can be expected to become more so as curriculum thinking evolves. In a briefing paper produced for the GELP collective[1] Simon Breakspear warned against leaping too quickly to produce new measures (metrics in his words) of 21st century learning outcomes without at the same time thinking deeply about how to connect them into the assessment system as a whole (Breakspear, 2013):

> We should focus first on how to make accurate judgments about student growth and development on the new deep learning outcomes. Once we can do this, aggregating to the system level is simple next steps. In short, all the uses of metrics first depend on good assessment information about where students are in one or more aspects of their learning (Breakspear, 2013, p.2).

This advice is essentially the stance argued in the DANZ report, and the 2011 MOE assessment position paper. We kept it in mind when shaping the sections that follow.

1 Global Education Leaders' Partnership: https://www.gelponline.org/

2. Outline of research questions and approach

The 2011 position paper was founded on the following principles.

- The student is at the centre.
- The curriculum underpins assessment.
- Building assessment capability is crucial to achieving improvement.
- An assessment capable system is an accountable system.
- A range of evidence drawn from multiple sources potentially enables a more accurate response.
- Effective assessment is reliant on quality interactions and relationships.

The Ministry expressed an interest in any evidence, or arguments grounded in evidence, that these principles should be updated, including consideration of whether any new principles should be added, especially given the rapid evolution of digital technologies in the intervening years. Specific questions for the review included consideration of:

- how teachers are using assessment to support teaching and learning in New Zealand, with a particular focus on the compulsory schools sector
- any shifts in the ways educators are using assessment for learning
- the system conditions necessary to support effective assessment for learning practices
- the extent to which key competencies/skills for global citizenship are being assessed and how they are being assessed
- the extent to which digital technologies are being used to personalise learning and give students, teachers, parents, and whānau rich information about learning
- use of effective strategies that engage parents/whānau and the wider community about different approaches to assessing and credentialing achievement in order to support learning.

Our approach

The questions above were used to create a list of search terms that included combinations of the following: assessment for learning; personalised learning; individualised learning; key competencies; core competencies; digital technology; classroom; data literacy; global citizenship; assessment; evaluation; classroom; meta-analysis; systemic reviews; and teachers. The search prioritised papers published within the past 5 years. The online research databases used for this literature search were: ERIC; Education Research Complete; BEI; INNZ; NZResearch; A+Education; and Google Scholar.

In addition to searching the online research databases we used a snow-balling process to seek out future-focused commentary published on the blog sites of academic/educational organisations with a

strong interest in assessment (i.e., not blogs of random individuals). From these sites we also found a number of relevant policy reviews and commentaries such as those concerning the recent "Gonski 2" review in Australia. We added these to the database if they had something relevant and future-focused to say about the themes we were finding in the formal literature. While the search was ongoing, we also reached out to national assessment experts and others in our personal networks with expertise, and to others recommended to us by those people, or whose names came up as key informants in future-focused assessment work taking place elsewhere. The full list of people who sent papers, made suggestions, or otherwise helped us, is included as Appendix 1 to this report.

As soon as the first papers came in we began to build a Zotero database of the relevant items. Each of us initially tagged entries to fit with the review questions. Once the high-level findings began to emerge we co-ordinated and streamlined the tags to create a searchable database of papers pertinent to each of the 2011 principles. This process helped identify sparser areas where we needed to keep searching for further papers and areas where further efforts would likely not add significant new ideas or evidence to those we had already captured in the database. By the time the draft report was submitted, 205 references had been entered in the database and we expect we will continue to add more. The system of tags used is shown in Appendix 2.

The structure of this report

The next section addresses the research questions outlined above. It also serves as a gathering point for a brief summary of what we found out related to each of the principles. It has been designed to stand more or less alone as a short paper (we do not repeat the information about our approach, as just outlined).

The following five sections use the principles as a lens for discussing main themes. They are ordered a little differently from the 2011 review. No particular significance should be read into this. The order was determined by our discussion about the logical sequencing of ideas and dilemmas: "You need to know about this before you think about that…" No ordering is perfect of course. We could not avoid some recursion—we are discussing a complex whole—but we did our best to minimise it, so that the report would unfold as a smooth read.

3.
Short answers to big questions

We begin the report of our findings by answering the questions posed for the review. This overview is supported by the literature outlined in the sections that follow. In the interests of brevity and readability, no references are included in this section. The Ministry's additional questions have been slightly reordered so that the account flows logically and smoothly.

Are the assessment principles still fit for purpose?

The principles were published in 2011 and a lot has changed in the intervening years. Nevertheless, with a few qualifications, the principles are still fit for purpose. We found a great deal of up-to-date literature pertinent to each of them.

When discussing this project with the Ministry, we all wondered whether it would be necessary to add a principle that addresses the rapid evolution of digital technologies and tools used for assessment and reporting purposes. We now think a new principle is *not* needed because these technologies are so pervasive that they impact every area of assessment practice. They expand the scope of all the principles rather than requiring something separate to be added. The sections that follow model this reasoning—every one of them includes at least some discussion of digital technologies.

We wonder if one of the principles needs to be reworked. This principle is "An assessment capable system is an accountable system". The *relationship* between the idea of accountability and the idea of assessment capability is not especially clear in the wording of the principle. We took the principle to mean that system accountabilities should be designed in ways that enable a valid and fair account of learning to be given, for every student, in every school. At the moment, the principle is written the other way around so that the onus of responsibility for improving the system falls on teachers and schools. There is an extended discussion of this point in the final section of the report.

We also wonder whether there is indeed one missing principle, or perhaps two. One potential principle would describe the central importance of clear *progressions* to new assessment (and curriculum) initiatives. If there is a second missing principle, it would centre on the *consequences* of assessment and the *equity* challenges that arise. Each of these is now briefly elaborated.

Progressions

There is a clear international trend away from an emphasis in summative reporting of point-in-time achievement towards reporting of progress made over time. For this to happen, teachers need progressions to guide their assessment decision making. The development of *The New Zealand Curriculum*

(NZC) digital technologies curriculum reflects this trend, with "indicators of progress" replacing "outcomes" as the organisers of content at the different curriculum levels. The account of its development highlights the careful work that will be needed if reporting in other learning areas is expected to take this format. However, we know very little about how students actually make progress in developing more complex learning outcomes being signalled in recent competency-based curriculum development work (both in New Zealand and internationally). The Coherent Pathways tool makes a first attempt to address this dilemma but it is a work in progress (see Section 5).

The concept of progress is easy to understand but new and different resources are needed if its benefits are to be realised in practice. For example, involving students in building their assessment capabilities could be better supported if teachers had better access to likely trajectories of learning progress (Section 4). Reporting to parents exemplifies a different type of resourcing challenge. Current database systems capture achievement data as clusters of learning outcomes. They are not configured in ways that allow progress to be meaningfully captured. Models of progression are needed to underpin use of digital technologies for reporting meaningful learning gains (e.g., by use of dashboards etc.).

There is a discussion to be had about priorities. Would we want something similar to the level of detail in the Progress and Consistency Tool (PaCT) progressions for every learning area? Assuming it was even possible to do this, what would the pros and cons be and how should we balance these? Would a new assessment principle provide a useful gathering point for working on the challenges outlined, and associated questions?

Equity

The call for reporting on progress is predicated on the understanding that students learn at different rates, in different ways, etc. Reporting of progress allows the learning gains of every student to be acknowledged, with the implication that this will be more motivating.

There appears to be a growing conversation about the potential for traditional assessment practices to have a differential impact on different groups of students—indeed, to play a significant part in constructing the inequalities being reported. Debate about this is strongest in literature about international testing, which we have not included in the report because doing so would have required a considerable expansion of scope. We note it here as an issue to watch and perhaps investigate further. There are a number of papers on the topic of equity implications of international assessments in the Zotero database.

Equity concerns are central to NCEA debates, and to literature that discusses students with special learning needs. There are also strong equity concerns in relation to the design and use of data analysis systems, and also more cutting-edge technologies such as the leveraging of neuro-biology to enhance learning. These references are threaded throughout the various sections of the report.

The additional questions

What evidence exists of shifts in the ways educators use assessment for learning?

This is an optimistic question. Even though the building of assessment capability across the sector is central to the 2011 assessment policy, we did not find strong evidence that shifts have occurred. If anything, the evidence suggests that assessment reform has remained static. Nor did we find evidence of gains in associated knowledge and skills—for example, in the area of increasing data literacy. Teacher educators are concerned about this issue and much of the relevant research we found was generated by them (see Section 4 for details).

In what ways are other jurisdictions managing AfL effectively?

New Zealand is not alone in struggling to introduce and then sustain assessment for learning (AfL) practices. We found several systematic reviews that suggest other nations are facing the same challenges as we are. There are various reasons why this might be so. The international literature is clear that AfL initiatives will falter if teachers perceive that assessment undertaken for accountability purposes does not fit with AfL practices. Another reason is that adoption of AfL practices requires considerable teacher knowledge and skill, and therefore sustained professional learning support.

What system conditions are necessary to support effective assessment for learning?

Pockets of progress in New Zealand schools suggest that building assessment capabilities to support AfL practices is challenging but achievable. The literature suggests that supportive conditions include:

- strong, knowledgeable, committed leaders who model described practices
- opportunities for teachers to work collegially as they build their assessment capabilities; for example, via learning about how to use data more effectively; social moderation of students' work; uptake and implementation of AfL pedagogies
- access to supportive curriculum materials that include clear indications of likely progress (i.e., progressions)
- amelioration of perceived accountability pressures; for example, by showing how achievement data can be used both formatively and summatively, and by leveraging moderation as a professional learning opportunity
- supported introduction of digital assessment resources that can generate rich formative feedback.

What evidence exists about modern curriculum design?

The expanded version of this question asks how other jurisdictions are meeting the challenge of an ever-expanding view of what constitutes valued learning outcomes, such as key competencies / skills for global citizenship.

The short answer is that there seems to be plenty of rhetoric but little in the way of modern curriculum design that successfully integrates traditional knowledge structures and new types of learning outcomes. The OECD's most up-to-date version of its "2030 learner compass" includes a vague zig-zag structure between the separate elements and the intended outcomes. New Zealand's fledgling "weaving" approaches do not appear to have international equivalents.

To what extent are key competencies / skills for global citizenship being assessed and how are they being assessed?

How best to assess key competencies is an unresolved question internationally. Their overall purpose (which we assume to include fostering new types of curriculum thinking) appears unresolved. It is clear that there are no easy answers. Many nations or groups of nations (e.g., the European Union) are struggling to find ways forward. In the literature, there is recognition that rich open-ended tasks with accompanying performance-based assessments provide students with opportunities to demonstrate their competencies. But such tasks create a series of challenges:

- Outcomes can be unpredictable—ideally, assessment would be co-constructed as learning unfolds, whereas rubrics are traditionally created ahead of the learning (and in NCEA, for example, are peer moderated before learning begins and cannot then be changed).
- Many tasks cross curriculum boundaries, creating a mismatch with traditional siloed assessment practices and adding to unpredictability challenges.

- There is limited evidence of students' actual capabilities on which to draw when planning assessments—building progressions is a complex area where ongoing research is needed.
- Traditional knowledge components are not yet necessarily assessed in rigorous ways. Outcomes in scope include: conceptual understanding (not just recall or understanding of pieces of knowledge); epistemic knowledge (how knowledge building practices work in different disciplines); and transfer of knowledge to new contexts.

Internationally, research efforts are being made to investigate ways of assessing collaboration in digital learning environments. This particular competency appears to have been singled out because of its importance for work and life in a globally networked world. This work is still in a development phase and the reports we read suggest there are obstacles to be overcome before such tools are made available at scale.

Microcredentials offer new possibilities for assessing contextually located performances such as specific competencies or work-related professional skills (in the case of teachers). Ideally, microcredentials will be designed to "stack"—the small grain size of each one contributes to the planned and sequenced building of a greater whole (see Section 9).

To what extent are digital technologies being used to personalise learning and give students, teachers, parents, and whānau rich information about learning?

There is considerable advocacy for personalisation of assessment, especially given new affordances of digital technologies. This can be viewed in two ways. A technological view of personalisation centres on artificial intelligence (AI) built into learning and assessment resources. Computer adaptive testing sits at one end of a spectrum here. At the other end, "stealth assessments" analyse students' interactions within the learning environment, without them being aware that they are being assessed. A clear message in the literature is that robust assessments of this type are not easy to design and hence expensive to build and trial. New validity and reliability challenges are raised by these sorts of assessments, which are clearly still in their infancy (see Section 6).

A broader view of personalisation involves tailoring assessment to promote the best learning outcomes for each child. This involves selecting and adjusting assessment methods, adapting content, personalising feedback and being inclusive. Digital technologies are included where appropriate but the teacher's decision making remains central to the learning action.

We did not find a strong evidence base demonstrating impacts of personalised assessment. When evidence of impacts is reported, the research tends to focus on personalised learning, accompanied by more traditional standardised assessments.

Is there use of effective strategies that engage parents/whānau and the wider community about different approaches to assessing and credentialing achievement?

The short answer to this question is again no. While the challenge of rapidly evolving curriculum and assessment practices is very clear, we found no papers that directly engaged in *how* others should be apprised of the changes and their implications. *That* they should be engaged has been advocated by one large American group as a principle for designing competency-based systems that are more equitable. The need is obvious but solutions, if they have been trialled, are not forthcoming.

There are a number of sources of evidence that describe the nature of interactions between school and parents and whānau. In recent years, schools appear to have put considerable effort into these communication efforts, at least in terms of reporting achievement. Two-way communication, in which schools draw on family and community expertise to strengthen and support learning, is not as common.

Thinking laterally, we could address this question via more active involvement of local communities in addressing some of the future-focused *design* challenges facing all education systems. The following examples illustrate this potential. They are elaborated in upcoming sections of the report:

- Active involvement of local communities is seen as a desirable principle for building a future-focused competency-based curriculum.
- Researchers in the field of data analytics advocate for wider public input into the design and building of databases of all types, including those that organise and store education data. In a world of AI and big data, the uses to which such data might be put can be far-reaching and it is important to get multiple perspectives at the design stage.
- Related arguments are made about the design of systems to award microcredentials, where credibility depends on recognition and a shared understanding of the robustness of the different parts of the overall "ecology" that generates the awards.

4.
Themes related to assessment capability

Principle: Building assessment capability is crucial to achieving improvement

This section of the paper focuses on the vision of assessment capability first described in the DANZ report (Absolum et al., 2009). In the DANZ report, building assessment capability across the whole education sector was seen as central to lifting overall achievement of students. The wording of the principle in the Ministry of Education's assessment position paper (Ministry of Education, 2011) implies that the DANZ argument was supported by the Ministry. It therefore provided a logical starting point for our own report. In this section we ask:

- Does this principle continue to be centrally important to assessment policy more generally?
- Is there evidence that the assessment capability and motivation of teachers and students has improved?
- To what extent has our overall system become more "assessment capable"?

The New Zealand studies that informed this section are mostly small-scale and qualitative. They are complemented by a number of larger international studies, with some including New Zealand as a context. We also briefly report on recent technology-enabled developments pertinent to building assessment capability (these will be covered more fully in other sections of the report).

The nature of assessment capability and its importance

Internationally, the term "assessment for learning" (AfL) is used to differentiate between formative assessment more generally and those pedagogies that involve *students as active meaning-makers* when considering their own progress and achievement (Booth, Dixon, & Hill, 2016). AfL pedagogies have proved demanding to adopt wherever they have been introduced (for example, see Hayward, 2015). Teachers need to build their own and their students' assessment capabilities so that they can confidently bring AfL to life in their classrooms.

Building the specific assessment capabilities needed for assessment meaning-making (for both teachers and learners) entails building the knowledge, skills, and dispositions to understand the learning implications of evidence generated by specific assessment feedback, to then be able to formulate next learning steps, and to be willing and able to act on these insights. It is also important for teachers to be data literate, so that they can support students to understand what assessment feedback is saying about

their learning and achievement. Booth and her colleagues (2016) describe building assessment capability as challenging but possible.

Since 2011, research in New Zealand and elsewhere has continued to support the principle that building assessment capability is important (Hill et al., 2017). AfL practices, which create opportunities for assessment capabilities to be deployed, have long been demonstrated to have a strong positive impact on achievement (Black & Wiliam, 1998). One international research collaboration (Baird, Andrich, Hopfenbeck, & Stobart, 2017) has recently cautioned that the impacts of AfL may be of a more modest nature than is often assumed, given the lack of robust quantitative evidence for AfL impact. This commentary raises an interesting dilemma about the sorts of evidence used to demonstrate impact. Robust quantitative evidence implies use of standardised assessment tools, which are typically available for a limited range of learning outcomes and hence cannot reflect the breadth of the curriculum (see Section 5). Nor can they measure other impacts seen as desirable—such as sustained changes in teachers' practice.

Teachers need to conceptualise AfL as part of everyday classroom practice, so that both they and their students notice and respond to evidence of learning (Darr, 2018). But high-stakes accountability pressures have been shown to act against widespread uptake of AfL, even when this has been a focus for systematic professional learning (Hayward, 2015; Jonsson, Lundahl, & Holmgren, 2015). One reason is that the "washback" from summative assessment can result in a narrowed curriculum which is less amenable to AfL practices (Baird et al., 2017). In the New Zealand context, Hill (2011) notes that school-level factors make it challenging for secondary teachers to change their assessment practices so that they are weighted towards AfL. One multi-national study, led by a Canadian team, identified National Standards as the "greatest threat" to using assessment formatively in New Zealand primary classrooms (Birenbaum et al., 2015).

Have teachers become more assessment capable?

Our review located limited substantive evidence that assessment capability has improved since the Education Review Office (ERO) reported (2007) that just over half of New Zealand primary schools and fewer secondary schools were demonstrating effective assessment practices. In its most recent national evaluation report on assessment, ERO highlighted that:

> … although considerable improvements have occurred in the collection and use of assessment in primary schools over the past decade, some schools continued to face challenges in improving the quality of their assessment practices. In the schools where leaders and teachers understood and valued the place of assessment, they introduced useful and manageable systems that benefited teaching and learning. At the other extreme, teachers collected assessments that were not well administered, analysed, moderated or used for improvement. This variability reduces opportunities for system-wide improvements in New Zealand schools. (Education Review Office, 2018b, p. 49)

ERO acknowledged "the considerable improvements many primary schools have made in their use of assessment over the past decade" but recommended "further work … to make sure all schools collect and use assessment data effectively to benefit all students" (2018b, p. 50).

Other studies of classroom assessment practices also report inadequacies in New Zealand teachers' assessment knowledge and skills (Hill & Eyers, 2016). Xu and Brown (2016) comment that teachers' ability to collect, analyse and plan using both formal and informal assessment data is generally weak. A major Teaching and Learning Research Initiative (TLRI) study reported that prolonged and intensive professional development is required to generate assessment capable teachers (Smith, Hill, Cowie, & Gilmore, 2014).

The Teacher-led Innovation Fund (TLIF) was designed to support collaborative teacher inquiries with the aim of developing innovative practices that could improve teaching and learning outcomes for students, especially those in underserved groups. However, a recent evaluation of TLIF projects concluded that,

despite some notable successes, there are ongoing issues with teacher confidence and capability in collecting, analysing, and acting on data. Consequently, many TLIF projects have not demonstrated substantive improvements for learners (Sinnema, Alansari, & Turner, 2018). Despite funding to provide time for teachers to work together on their inquiries, and for external support (which was not always accessed), teachers' uncertainties in using data to understand their practice worked against their efforts to effect meaningful improvement in outcomes.

One recently completed TLIF project directly addressed practices consistent with AfL in six junior primary classes in one school. Teachers used a range of explicit literacy and assessment practices, including assisting learners to self-assess. Supported by their project adviser, the teachers collaboratively developed and trialled a visual self-assessment tool for encoding text. These changes in pedagogy resulted in closer monitoring of student learning and responsive action by their teachers. The final project report (Sunnybrae School, 2018, unpublished) reports that teachers are now much more responsive to learning from collaborative data analysis, and "they can see the impact of responding to data, adjusting their teaching and observing the effect of this on student achievement". The views of the project team are supported by the external evaluation (Sinnema et al., 2018) which provides a case study of the project (pp. 67–73).

We found one study that probed New Zealand primary and secondary *students'* experiences of assessment (Harris, Brown, & Harnett, 2014). Nearly 200 students responded to a survey and drew images of their experiences of feedback. The majority of students drew images of teacher-led feedback practices, dominated by written comments or grades. They generally depicted and described this feedback as positive and constructive but there was little in this study to suggest that AfL was being enacted in their classrooms.

Data from NZCER's 2015 National Survey of Secondary Schools report modest shifts in students' opportunities to be actively involved in assessment decision making since 2012:

> The report of the 2012 survey results suggested that involving students in decisions about their learning was "still on the horizon for many teachers" (p. 25). In 2015, there was evidence of teachers making some small shifts towards that horizon, with changes for two items. Fewer teachers reported that their students never/almost never helped to set expected outcomes/standards for assigned work (47% in 2012, compared with 33% in 2015). Those reporting that students never/almost never co-created their own NCEA plan related to their career/academic goals decreased from 46% in 2012 to 34% in 2015. (Wylie & Bonne, 2016, p. 14)

The most recent data from NZCER's National Survey of Primary and Intermediate Schools provide an interesting indication that younger children might be seen by many teachers as not yet able to be active participants in the various learning and assessment activities described. Around a third of Years 1–2 teachers, compared with around two-thirds of Years 7–8 teachers, said their students quite often, or most of the time, had opportunities to: assess each other's work and give each other feedback; identify their own learning needs; and document their own learning achievements (Bonne & Wylie, 2017).

What might help? Learning from studies of practice

Are there instances where AfL is being implemented effectively and/or teachers' assessment capabilities are being strengthened in other ways? In this section we outline studies which suggest that such instances are likely to be characterised by teachers working on aspects of assessment practice in collaborative learning groups.

One recently completed TLIF project (Ramsay, Vetelino, Dewar, & Barker, 2018), highlighted the support that teachers need to change towards AfL practices. Working together as a collaborative learning

community provided the support for teachers and students to build their understanding and practice of AfL, including student self and peer assessment. Regular, supportive, structured peer observations assisted teachers to make steady progress toward implementing a wider range of teaching and assessment strategies. Success factors were identified as:

- knowledgeable within-school facilitation
- facilitators who were also supported in their roles
- commitment from school leaders
- provision of time for teachers to think critically about what they were learning from their inquiry to enable them to build new practices.

Several studies have suggested that teachers' assessment capability can be enhanced by sharing and moderating their judgements of student work with their colleagues (Hipkins & Robertson, 2012; Smaill, 2018). The introduction of National Standards for primary schools focused schools' attention on moderation. Principal survey data suggest that use of in-school moderation processes increased between 2010 and 2013 (Ward & Thomas, 2016). One qualitative study described the many hours teachers spent in social moderation of students' achievements against National Standards (Smaill, 2018). However, we have not found empirical research that links moderation, growth in teachers' assessment capability, and improved student outcomes. Smaill proposes further research that provides more information on actual shifts in reliability and assessment capability resulting from moderation, as well as research on how and what teachers in Kāhui Ako learn about moderation through their involvement in moderation conversations.

Teacher responses to NZCER's 2012 National Survey of Secondary Schools generated a factor that the researchers called "growing student assessment capability" (Hipkins, 2015). The factor comprised five items from a set of 12 items that asked about students' active involvement in a range of assessment practices. Exploratory analysis reported indications of relationships between this factor and teachers' perceptions of the professional learning support and the working ethos of their school. When teachers indicated there was a culture of learning together, a sense of valuing the curriculum vision and values of the school, and a view that the individual teacher had become better at meeting the needs of Māori students, they were more likely to have said that they valued and implemented practices that could build students' assessment capabilities.

Thorpe, Gilmour, and Walton-Roy (2017) described teacher practices when supporting students being assessed for composing music in collaborative groups, using NCEA achievement standards. In this high-stakes but innovative context (students are usually assessed individually) the practices used included self, peer, and teacher feedback on the creative process. The development of a clear conceptual model of what is involved in group composing was a key step in supporting teachers to adopt new approaches.

A very small qualitative study of three teachers of English for Speakers of Other Languages (ESOL) reported that classroom teachers used a restricted range of assessment tools to assess oral language skills (S. Edwards, 2017). The teachers were uncertain about what the English Language Learning Progressions (ELLP) meant. The study concludes that supporting mainstream teachers to use these progressions was helpful in terms of their knowledge of how to support ESOL learners.

Use of standardised test results for AfL purposes

AfL literature emphasises that summative assessments can be used in ways that support student learning—if teachers are assessment capable and hence know how to do so. We looked for evidence of such practice in New Zealand classrooms. Again, the studies are small and mostly qualitative.

Although most New Zealand schools use standardised tests, there is not a lot of evidence about how they use them. Caldwell and Hawe (2016) investigated how six teachers of Years 4–8 students in two schools analysed, interpreted, and used information gained from the Progressive Achievement Test: Mathematics (PAT: Mathematics) assessment tool. Although their schools had a longstanding commitment to PAT: Mathematics, once teachers passed students' stanine scores to school leaders they were not required to further analyse and use the information. Four of the six teachers placed little value on the data beyond accessing stanine scores. Teachers reported that they relied on colleagues and past experience rather than reading and using the assessment manual. The researchers concluded that a systematic and planned approach to the analysis, interpretation, and use of data is needed if students, teachers, schools, and other stakeholders are to get full value from the PAT: Mathematics tool.

Cowie and Cooper (2017) provide support for this view and point to insufficient grounding in basic mathematics and statistics as factors that impact on teachers' ability to use data in meaningful ways.

Building assessment capabilities during initial teacher education (ITE)

The evidence outlined above points to challenges in supporting teachers to change their assessment practices once these are established. Is it possible to work from the ITE level, building assessment capabilities while learning to be a teacher? Again, the evidence we located points to challenges, and suggests a range of reasons why ITE is not likely to generate sustainable change.

Typically, new teachers describe their pre-service preparation as less than adequate when it comes to using assessment in ways that support learning (Xu & Brown, 2016). Students entering programmes of ITE typically enter with low levels of confidence and negative views about assessment (L. Smith et al., 2014). Teacher educators are faced with student teachers who hold summative, formal views of assessment (Smith et al., 2014) and who bring negative attitudes from their previous experiences of being assessed (Hill et al., 2017). For these reasons it is challenging for teacher educators to shift the ideas of students in ITE courses towards assessment practices that are more educative and empowering.

Nevertheless, there is some evidence of success when teacher educators focus on building the assessment knowledge of student teachers (Grudnoff et al., 2017; Hill & Eyers, 2016). Hill and her colleagues (2017) described a one-year Master of Teaching programme in which assessment principles and approaches, as well as a focus on teaching for equitable outcomes, were embedded in most of the pre-service courses and were included in authentic teaching experiences in classrooms. Data from the 27 participants demonstrated that the pre-service assessment courses significantly impacted on student teachers' attitudes and skills. However, the teacher educators did not know how these ITE students were supported to continue to develop their assessment capabilities as practising teachers in the schools they were employed in.

In 2017, ERO reported that nearly one-third of new graduated primary teachers (NGTs) who completed its online survey were only somewhat confident or not confident at all to use data to inform their planning and teaching. Secondary teachers rated themselves as more confident than their primary colleagues (Education Review Office, 2017b).

Do professional teaching standards support AfL pedagogies?

New Zealand is unique in its provision of a 0.2 component of additional salary in the first year of provisional registration for the provision of induction and mentoring programmes to support the ongoing professional learning of newly graduated teachers. At the end of the 2-year induction period, the school principal attests that the beginning teacher has met standards for full teacher registration (Cameron &

Baker, 2004). The continuing development of their assessment capabilities after they begin teaching is dependent on the quality of the support provided to them by other teachers, school leadership, and the education system. ERO noted that, because of the variation in NGTs' confidence and preparedness, they need different types and levels of support as they work towards full certification. According to ERO, "this support is especially important in relation to assessment for learning, responding to diverse learners and working collaboratively with parents and whānau" (Education Review Office, 2017b, report summary, no page number).

When gaining and maintaining full registration, teachers are required to demonstrate evidence that they have met New Zealand's teaching standards. Aitken, Sinnema, and Meyer (2013) identified a number of inadequacies with the New Zealand standards, including their focus on knowledge acquisition separated from the act of teaching, and their non-active, non-applied nature. They suggested that a shift toward conceptualising teaching as inquiry could go some way towards addressing their concerns. This recommendation appears to have been acted on—responsibility for professional inquiry is woven through several of the updated standards, which have been fully implemented in 2018 (Education Council, 2017). However, teaching as inquiry rests upon assessment capability, and this has recently been identified as problematic (Sinnema et al., 2018). We also noted that aspects of assessment practice are distributed across the standards rather than forming a specific group of their own (Education Council, 2017). Some research raises the question of how much impact professional standards do actually have on practice. Call (2018) reported that an Australian Institute for Teaching and School Leadership (AITSL) Survey in 2013 found that, although 83% of teacher respondents thought that the Australian Standards would improve the profession, only 54% stated that they used them to improve their own teaching.

In short, there is little published evidence that the professional teaching standards are impacting assessment capabilities, either for early career teachers or those who support them. There is one promising international innovation—when microcredentials are used to assess and document specific professional capabilities, teachers can be motivated to demonstrate evidence of their professional learning and growth (Kuriacose & Warn, 2018). We return to this development when we discuss the principle of system-level accountability.

Supporting AfL with digital assessment tools

Digital assessment tools can be designed to personalise feedback on learning. These affordances are discussed more comprehensively in the section covering the principle that the students should be at the centre of learning (Section 6). In this section we review the small amount of evidence we found that explores whether students might be taking up these opportunities and, in particular, building their assessment capabilities by interacting with formative feedback generated digitally.

In a New Zealand study, Harper and Brown (2017) investigated tertiary students' use of online feedback, and the impact on summative assessment, in a large first-year biology course. Students were encouraged to access online feedback on their achievement throughout the 12-week course, and to access tutorials to build their understandings of course content. The few students who accessed the tutorials showed a small improvement in their results but, overall, students did not engage with the tutorials. The researchers comment that qualitative information would be needed to understand why, but consider that it was likely there was insufficient time to engage with feedback in such an intensive course. In their view, effective formative assessment opportunities can be ignored by students in the context of high pressure and summative assessment.

Recent Canadian studies have explored the impact of iPads on formative assessment in Y7–9 classrooms. Searle, Elrofaie, Kirkpatrick, Sauder, and Brown (2017) found that teachers provided with iPads as part of a

District initiative reported using them for diagnostic assessment before teaching a unit. They used a range of apps (Showbie, Schoology, Kahoot) to record, sort, organise data, keep track of learning, and to provide feedback throughout an 8-week blended learning unit of work. Learners were involved in assessment (e.g., via video analysis in Physical Education, and in using self-evaluation checklists). However, these teachers also reported frustration at getting students to read and use the feedback. Consequently, these researchers suggest that coaches are essential to help teachers and students utilise the affordances that technology provides and to sort out technological and practical difficulties.

System-level support for strengthening assessment capabilities

Shortly after the publication of the Ministry of Education position paper on assessment, one of the DANZ authors (Flockton, 2012) published a commentary which stressed that "strengthening assessment capability will require considerable work to address curricular issues and ambiguity around descriptions of learning experiences and progressions" (p. 143). We return to the challenging issue of describing clear progressions in Section 5 of the report. Flockton also noted the need for alignment at every level of the system, and suggested that teachers would need sustained professional learning opportunities.

A recent international study tracked the actions of positional leaders with responsibility for implementing AfL. Initiatives lasted between 3–10 years and the leaders met with the research team at dedicated symposia. The key finding of this large study was that system shift is achieved when leaders engage deeply with, and model, AfL themselves. They need to continue to do so over time, supporting system and/or teacher learning with AfL approaches, and engaging with a wide range of stakeholders to provide feedback about what is working and what might need to evolve or change. They value both qualitative and quantitative data and develop systems for triangulation. When there are external pressures for summative data to be provided, successful leaders find ways to maintain the focus on AfL as the means of generating the data needed (Davies, Busick, Herbst, & Sherman, 2014).

With these challenges in mind, we looked for recent New Zealand initiatives with the potential to support and strengthen assessment capabilities at all levels of the education system. The development of Kāhui Ako is the most recent of such initiatives. ERO has published the first of a planned series of iterative reports drawing together what it has learnt about Kāhui Ako as they move from the establishing phase to implementing their agreed inquiries. The report stresses the importance of strong leadership, including building the collective capacity for data-informed inquiry that leads to sustained improvement. ERO recommends targeted professional development for those appointed to leadership roles (Education Review Office, 2017a). In a working paper for the Education Council, Bendikson (2015) indicated that, in addition to skills such as ability to engender trust, and confidence, leaders of Kāhui Ako require capabilities in leading people through a process of problem analysis, and a willingness to measure "intermediate outcomes", to enable participants to learn how they are progressing towards their goals. This, she says, requires "a reasonable amount of data literacy" (p. 3). It is also highly likely that the assessment capability issues identified in the TLIF evaluation (Sinnema et al., 2018) will create challenges for Kāhui Ako communities.

The Ministry has invested in new tools to help Kāhui Ako to better understand the learning needs of students and adults. These are very recent, and Kāhui Ako leaders need to apply to the Ministry to access the resources.[2] It is too soon to know what impact they may have.

2 https://curriculumtool.education.govt.nz/

In summary

Building assessment capability across the whole education system continues to be an important goal. However, there is little evidence to suggest that, since 2011, widespread progress has been made on achieving this goal, either for individual teachers or the system as a whole.

Pockets of progress suggest that building assessment capability is challenging but achievable. These challenges could be addressed by the provision of more supportive conditions. The literature suggests that supportive conditions include:

- strong, knowledgeable, committed leaders who model described practices
- sharing of cases where schools have built strong AfL cultures, including evidence of the positive impact for students and teachers
- opportunities for teachers to work collegially as they build their assessment capabilities; for example, via learning about how to use data more effectively; social moderation of students' work; uptake and implementation of AfL pedagogies
- access to supportive curriculum materials that include clear indications of likely progress (i.e., progressions), and that tease out the conceptual demands of novel task types
- amelioration of perceived accountability pressures; for example, by showing how achievement data can be used both formatively and summatively, by leveraging activities such as moderation or the making of comparative judgements (see Section 5) as professional learning opportunities
- supported introduction of digital assessment resources that can generate rich formative feedback.

We expand on these opportunities in the following sections of the report, progressively building up a rich picture of interconnections between the six principles.

5. Themes related to alignment of curriculum and assessment

Principle: The curriculum underpins assessment

This section focuses on the relationship between curriculum and assessment. Fundamentally, assessment should focus on what is valued in the curriculum, and what is important. Alignment between *The New Zealand Curriculum* (*NZC*) and both formative and summative assessment is in scope. More informal relationships between teachers' enacted curriculum and assessment practices are also in scope. In this section we ask:

- Does this principle continue to be centrally important to assessment policy more generally?
- Is there evidence that alignment between higher stakes (summative) assessments and *NZC* is problematic?
- What new alignment challenges have arisen since 2011 and how might we respond to these?

Section 6 drew on a range of already existing evidence. In contrast, the research discussed in this section tends to be focused towards what *could* happen rather than what is already happening—at least at levels of scale. Some exploratory studies do open up new possibilities.

Evolving relationships between curriculum and assessment

Ideally, assessment should reflect what is valued in the curriculum—as this principle implies. The reality, suggested by a considerable body of research, is that this is often not the case. Instead, what is assessed in high-stakes contexts *becomes* the curriculum that is enacted in classrooms. Section 4 noted that the "washback" from summative assessment can result in a narrowed curriculum (Baird et al., 2017) and that these effects apply in New Zealand, just as much as they do in other nations (Birenbaum et al., 2015; M. Hill, 2011; Thrupp & White, 2013). Across a decade of NZCER national surveys a clear majority of secondary school teachers continue to believe that NCEA drives the curriculum in the senior secondary school (Hipkins, 2013).

Apart from the narrowing effect, this washback would not necessarily be a problem if high-stakes assessments did fully reflect the intent of the curriculum. But, as we now outline, there are complex reasons why this is unlikely to be the case. Given the rapid evolution of curriculum thinking, commonly used assessment tools and practices are lagging behind.

The need to expand and adapt assessment practices to gather evidence of "21st century" outcomes lies at the heart of the challenges discussed in this section. Internationally, there is a focus on ways that learning demands are changing in the face of rapid technological change and continuing globalisation, with associated environmental and social challenges (OECD, 2018). A common theme of these commentaries is that students need to be more knowledgeable than ever, yet also more adaptable, agentic, and resilient (Bereiter & Scardamalia, n.d.). Competencies now needed by every student are complex and multifaceted, and hence require new assessment thinking. Given these challenges, the final report of the Gordon Commission on the Future of Assessment in Education (2013)[3] concluded that familiar assessment traditions are now "dysfunctional to the needs of education in the 21st century" (p. 9). The Commission's final report also argues that artificial intelligence and big data analytics will exacerbate the need for some competencies that we currently know less about. These include "pattern recognition and generation of patterns; rationalization of contradictions; the adjudication of relational paradoxes; and the capacity for virtual problem-solving" (p. 16).

Even the traditional "knowledge" component of the curriculum cannot be taken for granted. We found one international group of researchers who question whether traditional assessments actually do capture the deep conceptual understanding that is now needed (Bisson, Gilmore, Inglis, & Jones, 2016).

Traditional pen and paper assessments measure the acquisition and understanding of traditional curriculum content. Over many years, robust processes have evolved to ensure validity and reliability of these traditional assessments. The evolution of new digital assessment tools intensifies these challenges. New ways of thinking about establishing the validity and reliability of evidence are now needed (Shute & Rahimi, 2017).

Digital tools also enable ideals such as "just in time" assessment to become achievable in principle, if not yet in practice (Murgatroyd, 2018a) with associated implications for how curriculum time is managed. Digital assessment tools can also shift the balance between formative and summative assessment, opening up access to different types of evidence of learning and achievement. At the more radical end of a continuum of change, these shifts potentially make it possible to fully embed assessment in curriculum, so that no learning time is diverted to stand-alone assessment events (Groff, 2018).

The challenge of assessing key competencies

NZC was designed as a curriculum framework that would be responsive to the rapid changes taking place in the 21st century. The key competencies, derived from the OECD DeSeCo project (OECD, 2005), were an important new addition. OECD's own thinking has continued to evolve beyond the originating DeSeCo conception (OECD, 2018). Assuming that our national curriculum thinking should also continue to evolve, the challenges of ensuring that assessment reflects the high-level intent of *NZC* have actually intensified since 2011. Similar challenges are being faced all around the world, including by the nations of the European Union (Siarova, Sternadel, & Mašidlauskaitė, 2017).

The full potential of curriculum changes signalled by the key competencies has taken a decade to be understood, even by curriculum researchers (McDowall & Hipkins, 2018). In brief, recent thinking argues for a *weaving* of key competencies and traditional content, such that different types of outcomes are envisaged. This approach is predicated on the assumption that key competencies were intended to change the curriculum, not just add additional layers. They should do so in ways that change students' encounters with knowledge, allowing them to *demonstrate what they can do* with their learning (Hipkins, 2017). This argument implies a need for at least some assessments that are performance-based. We return to this challenge when we discuss the principle of using a range of evidence drawn from multiple sources.

3 The paper by Bereiter and Scardamalia cited in this paragraph is one of around 30 papers written by assessment experts who contributed to the work of the Commission.

Researchers from the Centre for Assessment Reform and Innovation (CARI) at the Australian Council for Educational Research (ACER) have recently described a "chicken and egg holding pattern" (Scoular & Heard, 2018, p. 1). Teachers are reluctant to take risks with integrating, teaching, and assessing general capabilities in the face of uncertainties about the sorts of outcomes toward which they should aim, and in the absence of effective approaches to developing new sorts of outcomes. Meanwhile, researchers can't generate evidence of achievement of different types of outcomes unless teachers work with them (Scoular & Heard, 2018). CARI is currently undertaking a national trial of specific assessment resources to support the capabilities in the Australian national curriculum. These resources model problem-based learning tasks, and have been designed with transfer in mind—they should be able to be adapted to many learning contexts. Complex problem solving requires students to demonstrate capabilities of collaboration, critical thinking, creativity, research, and communication skills. The intention is to assess these as clusters, as they are used together in action. The researchers then hope to underpin the task exemplars with learning progressions and some broad frameworks for these have been developed. However, it is not clear to us at this stage, if these will be generic or task specific (Scoular, 2018, outlines plans in an as yet unpublished conference paper).

The CARI researchers acknowledge the limitations of their chosen solution, noting that 21st century skills and capabilities "manifest themselves in an enormous range of expressions, contexts and applications that are beyond the scope of a small suite of classroom tasks to definitely assess" (Scoular & Heard, 2018, p. 3). Nevertheless, they say, there is a need to start somewhere and to directly address the critique that the capabilities are too nebulous to be assessed reliably. The intention is that the trial will lead to wider adoption of the approaches modelled so that the curriculum capabilities do become more "comprehensively" embedded in classroom practice.

How new/emergent outcomes create assessment challenges

In order to build their competencies/capabilities, students need opportunities to grapple with rich tasks set in contexts that are meaningful for them (McDowall & Hipkins, 2018). This curriculum requirement surfaces a range of assessment challenges. Collectively, the Gordon Commission papers:

> … challenge the testing industry to develop assessment systems that can capture evidence of student learning at multiple time points, from different sources (i.e. inside and outside school settings), different types (e.g. quantitative and qualitative), and that allow for demonstration of student learning in different ways. (The Gordon Commission on the Future of Assessment in Education, 2013, pp. 7–8)

This challenge implies the need to: expand assessment practice beyond well-established methodologies; include performance assessments in a broad repertoire of data-gathering strategies; and think more laterally about where and when learning can take place. All of these points have already been raised in this section.

New curriculum thinking surfaces another important implication that has not yet been raised. Rich tasks and learning carried out in different locations raise the distinct possibility that students will demonstrate important learning that was not anticipated in curriculum and assessment plans. Bolden and DeLuca (2016) highlight the limitations of measuring only predetermined learning outcomes, and focusing only on the learning destination. They say teachers should make space for unintended (emergent) outcomes. The idea of emergence is drawn from complexity theory and they use this theory to advocate for assessment of capabilities that students demonstrate when using pedagogies that are known to yield emergent learning. Such assessment would be focused on evidence of: imaginative, non-mastery learning; learning with and from others; self- organisation and autonomous learning; and revising ideas in new ways (i.e., recursive elaboration) (Bolden & DeLuca, 2016). Elements of this assessment manifesto are rather similar to student capabilities highlighted in the OECD's most recent "2030" curriculum thinking (OECD, 2018) and

overlap strongly with the *NZC* principle of *learning to learn*. Elements of self and peer assessment are also implied but, first, students need support to know what they might focus on. The next small exploratory study illustrates one potential way to at least partially address this challenge.

One recent TLIF project explored practical ways to capture unanticipated outcomes of learning from an innovative integrated Year 9 programme (White & Hipkins, 2017). The school curriculum was integrated around rich inquiry topics, affording opportunities for learners to simultaneously build knowledge and key competencies. For assessment, students presented narrative "learning stories" to teacher and peers, describing what they perceived to be their most powerful learning during the term. As part of their iterative inquiry process the teachers devised "outcome cards" based on the school's learning-to-learn values (which predated the key competencies but are similar in scope). These cards were used in a second stage of the assessment process. They prompted students to recognise, reflect, and report on ways in which they had developed capabilities embedded in their initial narrative (but not necessarily explicitly recognised until prompted).

We also found a recent blog post that addressed the dilemma of how to assess for emergent outcomes from open-ended tasks (White, 2017). White recommended that teachers co-develop criteria with the students—but not until both they and the students have been immersed in the open-ended learning activity to build practical experience and insights to bring to the setting of criteria and personal goals.

Open-ended (rich) tasks often require curriculum integration because they cross the subject divisions of the school curriculum. In the senior secondary school they also demand selection of NCEA standards from different subject domains. This is another context that opens up the possibility of emergent outcomes. A recent case study of curriculum integration in a New Zealand secondary school highlighted challenges for meaningful integration of different subject areas. The paper makes the case that outcomes of integration should be *different* from anything that can be achieved by learning within the individual subjects (McPhail, 2018). The critical focusing question for assessment then becomes: What learning opportunities emerge that could not be achieved in stand-alone subjects? This question is essentially a variant of the key competencies question outlined above. How are new curriculum possibilities opened up by meaningful weaving? Can new types of outcomes be described and shown to combine knowledge in new ways and/or combine knowledge with strategically selected competency-related learning goals?

As we have already indicated, the assessment of curriculum knowledge per se cannot rest only on traditional assessment of declarative knowledge. An emergent thread of discussion relates to the importance of building awareness of the nature of disciplines—labelled as "epistemic knowledge" in the OECD's 2030 learning compass (OECD, 2018). One small case study illustrates how this type of knowledge is woven into assessments in some NCEA achievement standards but not others (Johnston, Hipkins, & Sheehan, 2017). Drawing on longitudinal NCEA results, the researchers argue that learning about how disciplines build knowledge provides a stronger foundation for subsequent study in the same subject area. Elsewhere, the same team has argued that knowledge about how the disciplines work is also important when learning areas are integrated, so that their integrity as knowledge-building systems does not become lost in the integration process (Hipkins, Johnston, & Sheehan, 2016).

Practice-related research with a focus on building epistemic knowledge is not easy to find.[4] One small case study, set in the context of biotechnology learning in a primary classroom, argues that teachers must model disciplinary thinking as part of their formative assessment practice if they want to support students to become more autonomous learners. This type of modelling should help students "appreciate and experience how knowledge is generated and warranted within biotechnology" (Cowie & Moreland, 2015). This sentence clearly implies a focus on epistemic knowledge although the term is not used.

4 There is considerable philosophical discussion of this challenge but that is beyond the scope of this paper.

Progressions: A necessary link between curriculum and assessment?

Increasingly, policy makers (and politicians) are calling for evidence of progress to be used as the preferred accountability measure of whether students are learning as they should. In Australia, the recently released "Gonski 2" report advocates strongly for this (Gonski et al., 2018). Measuring *progress* provides a deliberate counterpoint to the traditional practice of measuring *achievement* at specific time points. Problems with this traditional model of point-in-time achievement include the different starting points of different students. In a recent blog post, Masters (2017) asserts that "in any year of school, the gap between the most advanced 10% of students and the least advanced 10% is the equivalent of at least five to six years of school". He also notes that students who consistently receive low grades do not get any sense that they are making progress, while the most able students are seldom stretched when they consistently receive high grades. By contrast, a focus on progress is underpinned by a belief that every student is capable of learning if they can be engaged and motivated to make the effort. It is also possible to personalise and target learning and assessment at an appropriate level for each student (Masters, 2017). We come back to the relationship between personalisation and progression in Section 6.

While the argument for benefits of assessing and reporting progressions appears straightforward, identifying what these should look like is not. A group of American assessment experts have described three different ways of understanding the nature of progression: (1) increasingly sophisticated ways of thinking about or understanding a topic; (2) a framework for formative classroom practice that reflects how students learn within a domain; (3) building blocks to mastery of knowledge and skills addressed in college- and career-ready standards (Sturgis, 2015). Given the potential for confusion about which understanding is meant in any specific context, this group recommended caution in proceeding with the development and use of progressions intended to support assessment for learning.

This group made two other arguments in their critique. First, research has yet to definitely establish whether mapping out and following learning pathways based on progressions does actually lift achievement. The same point has been made by critics of the Gonski 2 report (Buckingham & Joseph, 2018). Second, progressions research to date has been carried out inside subject silos. Little is known about whether students will progress at different rates in interdisciplinary contexts (Sturgis, 2015). This question compounds the challenge of clearly specifying emergent outcomes from interdisciplinary learning, as outlined above.

What might progress look like when the curriculum aims to develop new types of outcomes? This question is not easy to answer, in part because of the "chicken and egg" situation described above (Scoular & Heard, 2018). Only when students are offered rich opportunities to demonstrate their capabilities will we know what they are actually capable of. Nor is this dilemma restricted to newer and emergent types of outcomes. A systematic review of research-based progressions in mathematics and science (Mosher, 2011) identifies the possible underestimation of students' potential progress as an issue within traditional subject learning. The progressions in question were of the second type described above (i.e., frameworks for supporting formative assessment in the classroom). Even within that one type, Mosher identified two approaches to formative practice. Some progressions were based on a "fix it" model of practice while others were based on a "work with it" model that gave educative guidance to teachers about why students might not be making expected progress.

This is clearly a complex area where ongoing research is needed on a number of fronts.

The impact of high-stakes assessments on the enacted curriculum

The relationship between curriculum and assessment in the senior secondary school is seen as particularly problematic. Given the review of NCEA now underway, we did not focus too much on this theme but it was clear in the papers that we did read. The modular nature of NCEA is seen to provide

opportunities to build new types of learning experiences but, as in the case of AfL, teachers' NCEA decision making is sensitive to accountability demands. This is a problem when accountability measures create perverse incentives to lift achievement results in ways that short-change students' learning opportunities (Hipkins et al., 2016).

In a national context in which curriculum thinking continues to evolve, many teachers are holding back from making substantive curriculum changes, or are continuing to rely on traditional curriculum content, as reflected in high-stakes assessments. But there are pitfalls in relying on NCEA in this way. Hipkins et al. (2016) make the case that, given the modular flexibility of NCEA, teachers now have to be active *curriculum builders*. They also suggest that curriculum coherence is best achieved by deliberate selection of achievement standards that support a clear purpose for a course, regardless of whether they come from within the same subject.

The flexibility and freedom to select standards can also be used in ways that are ultimately not in students' best interests. Assumptions about students' abilities and interests can act to limit their opportunities to experience challenging learning and a coherent curriculum, especially if their teachers mainly offer assessment via standards that they think will be easy for students to achieve. We return to equity challenges in Section 6.

As Groff (2018) has done for digital resources within in-built assessment functions, Hipkins et al. (2016) envision a future for NCEA where assessment is so enmeshed in the everyday work of students that they are not necessarily aware of being assessed. All the work they undertake during a course potentially provides evidence that could contribute to NCEA credits and challenges such as fragmentation, over-assessment, credit-counting, and selective disengagement from assessment would all be positively addressed. They say that achieving this vision requires confident teachers who have strong disciplinary knowledge and the full support of their colleagues and school leaders. Earlier parts of this section also suggest that teachers would need to build expertise in a repertoire of performance-based assessment and reporting strategies.

Case study research on the impact of National Standards in a selection of primary schools reported a narrowing of the curriculum when schools devoted more of each day to literacy and numeracy learning:

> … it was increasingly difficult to fit in 'topic work', the 'big idea' or 'concept' and the attention to science, social science, environmental studies and arts they represented. Such material was often only covered in the last block of the day when children were getting tired and less focused" (Thrupp & White, 2013, p. 20).

They said that the scope of the curriculum in reading, writing, and mathematics also became narrower, as teachers focused on technical points needed to lift achievement to specific curriculum levels.

Digital assessment tools and curriculum

Digital technologies are impacting on the curriculum/assessment interface in a range of ways. One example is the way the dynamic between curriculum and assessment changes when students are learning in Innovative Learning Environments (ILEs). The impact on assessment of digitally-enabled innovative ways of using space was an interesting but sparse theme in the literature we found (Crisp, 2014). Another example is the use of computing power to support new ways of making comparative judgements about more complex demonstrations of learning. This is outlined in some detail because of its potential to contribute new insights about assessment challenges such as describing new types of progressions.

One international research team argues that learning in traditional academic subjects such as mathematics and language is increasingly seen to require complex ill-defined abilities such as creativity, sustained reasoning, and communication skills (Bisson et al., 2016; Jones & Henderson, 2016). Such abilities can be evidenced using open-ended tests that are designed to prompt a varied and

unpredictable range of student responses: for example, the prompt "What are the differences between the mean, the mode and the median? Give examples of when they are appropriate for summarising data." It is difficult to design corresponding scoring rubrics, which traditionally assume predictable and uniform responses so that a team of markers can apply points objectively. Recent developments in comparative judgement have enabled this barrier to be overcome by doing away with rubrics and instead collating experts' pairwise judgements of student responses. It doesn't matter if different experts look for slightly different markers of conceptual understanding so long as each example is assessed a sufficient number of times (these researchers advocate that each example is compared in 10 different combinations) (Bisson et al., 2016).

This research team's systematic programme of evaluation of comparative judgement shows promise. Comparisons with traditional assessment measures indicate high levels of validity and reliability (Bisson et al., 2016). Jones and Henderson (2016) say that three components were critical to the success of the studies: (1) a robust and usable online comparative judgement engine; (2) the recruitment of judges who are experts in their fields and undertake their assessments sincerely and thoughtfully; (3) the design of test questions that are truly open-ended, and that are perceived to have high face validity by end users. They also noted that the tests proved popular with many teachers, because their students' responses provided valuable, intuitive snapshots on deep learning. One blog post on the team's website[5] gives examples of how very large samples of student writing, generated in different contexts (e.g., different time points, different ages, different places) can be validly and reliably compared. The post describes several purposes for which such large-scale comparisons might be made—to check progress over time; to compare cohorts in consecutive years, etc. (Christodoulou, 2018).

A working paper recently released by the Ministry illustrates the potential for using comparative judgement technology to develop curriculum progressions in areas that are breaking new curriculum ground (Ministry of Education, 2017). This paper describes the process followed to develop two sets of progressions for the new Digital Technologies curriculum, using a comparative judgement process. Teacher volunteers with expertise in the Digital Technologies curriculum judged a range of student responses to rich tasks. They were asked to determine which of each pair was the more sophisticated. Psychometric analysis was used to determine the location of each exemplar on a common scale, and to check for variation in the judgements of different individuals. Use of actual student work ensured that the in-principle level for each task, as designed, was tested against what actual students could do.

Game-Based Assessment (GBA) opens up new possibilities for embedding assessment within the learning context. This is sometimes called "stealth assessment" (Shute & Rahimi, 2017). Games designed for learning now typically offer constructivist environments where "learning is situated in context, with individual and collaborative problem-solving and real-time data collection and where feedback guides the leaners towards deeper, more meaningful learning" (Groff, 2018, p. 189).

> Imagine a future where learners have seamless, integrated experiences across domains and their education experience and where data are constantly collected, modelled and fed back to learners in order to personalise their learning experience, not just in game-based environments, but across all their learning experiences, from science labs, to writing tasks, to maths practice, whilst never interrupting their experience to implement a formal test. Understanding how to assess and design formative experiences in game-based environments is a critical step in pursuit of this reality. (Groff, 2018, p. 197)

5 The website is called No More Marking (www.nomoremarking.com). This name reinforces the emphasis on the speed and practicality of this type of assessment, but also draws attention to the potential commercial opportunities. As with some other studies reported in this section, there are vested interests at play. This is no doubt why this team, and others like the EPQ team in the UK which is discussed later in the report, have been careful to underpin their claims with evidence-based research.

In common with other researchers in this space (Baker, 2018; DiCerbo, Shute, & Kim, 2017), Groff advocates use of Evidence Centred Design (ECD) processes for constructing valid assessments. ECD processes are rigorous and require deep expertise on the part of the assessment designers (DiCerbo et al., 2017; Shute & Rahimi, 2017). Unanticipated outcomes can occur, as has been demonstrated when new assessment tools are evaluated using video and audio of learners interacting in the environment. Horwitz (2018) provides an interesting case study in the context of a physics resource designed to assess students' knowledge of Ohm's Law as well as their ability to collaborate in problem solving. One student initially appeared to understand Ohm's Law but had trouble persuading his team mates to listen to him, and did not himself try to explain to them what he seemed to know. Once they figured out how to work together, this triad made some progress in round 2 but a misconception that arose early in round 3 derailed the learning of all of them, including the student who had initially seemed to have mastered Ohm's Law. Horwitz concludes that complex learning tasks such as this are best analysed by humans rather than data analytics until more is understood about the dynamics of the learning pathways that a resource such as this can open up.

The demand for evidence-centred design at the outset of the creation of a digital resource with embedded assessment stands in contrast to the process of comparative judgement (CJ) outlined above. The CJ researchers stress the ease of constructing suitable assessment tasks, with the rigour coming at the judgement stage (Bisson et al., 2016). Both CJ and ECD approaches offer the potential to exemplify and support new curriculum thinking. A mix of them would allow their differing benefits to be achieved.

In summary

This principle continues to be very important. Curriculum needs to underpin assessment so that we assess what we value (most). What this section has highlighted is that what we value is changing, and that we need to grow our understanding of how what we value might be assessed. We have also highlighted that what is assessed in high-stakes contexts continues to become the curriculum that is enacted in classrooms. There is evidence of this effect in New Zealand—both from NCEA and from the now disestablished National Standards.

There are complex reasons why assessments are challenged to reflect the full intent of *NZC*. Given the rapid evolution of curriculum thinking internationally, commonly used assessment tools and practices are lagging behind. The need to expand and adapt assessment practices to gather evidence of "21st century" outcomes lies at the heart of alignment challenges.

How best to assess key competencies is an unresolved question internationally. It is acknowledged that opportunities to demonstrate these are provided by rich open-ended tasks and by performance-based assessments. But such tasks create a series of challenges:

- Outcomes can be unpredictable—ideally assessment would be co-constructed as learning unfolds, whereas rubrics are traditionally created ahead of the learning.
- Many tasks cross curriculum boundaries, creating a mismatch with traditional siloed assessment practices and adding to unpredictability challenges.
- There is limited evidence of students' actual capabilities on which to draw when planning assessments—building progressions is a complex area where ongoing research is needed.
- Traditional knowledge components need to be assessed in more rigorous ways—outcomes in scope include: conceptual understanding (not just recall or understanding of pieces of knowledge); epistemic knowledge (how knowledge-building practices work in different disciplines); and transfer of knowledge to new contexts.

Digital assessment tools add to curriculum/assessment alignment challenges by:

- enabling "just in time" assessment to take place close to the learning
- allowing access to different types of evidence of learning and achievement
- shifting the balance from summative to formative assessment
- requiring new thinking about validity and reliability challenges
- providing new methods for analysing and assessing "progress"
- blurring the boundaries between curriculum and assessment when the latter is embedded within a digital learning environment.

6.
Themes related to student-centred assessment practices

Principle: The student is at the centre

NZC includes a set of curriculum principles that predate the assessment principles. One of the *NZC* set "put[s] students at the centre of teaching and learning, asserting that they should experience a curriculum that engages and challenges them, and is forward-looking and inclusive, and affirms New Zealand's unique identity" (Ministry of Education, 2007, p. 9). There is clearly congruence between this intention and the assessment principle we examine in this section. We ask:

- Does this principle continue to be centrally important to assessment policy more generally?
- Is there evidence that student-centred assessment practices are being used?
- What new challenges might the Ministry consider when reviewing this policy?

The New Zealand studies we report in this section are again mostly smaller qualitative studies, some undertaken by teacher practitioners. The international studies tend to be larger in scale.

The idea of personalised assessment

The Effective Pedagogy section of *NZC* lists the *active involvement of students* as a characteristic of effective assessment (Ministry of Education, 2007, p. 40). Assuming this means they are at the centre of assessment practices, this assessment principle has clear overlaps with the "assessment capability" principle already discussed. If students are to be active participants in their own assessment, then logically the teacher must have a focus on their achievements, progress, and ongoing needs. What this principle adds is the concept of *personalised* learning and assessment. Personalised assessment provides learners, teachers, parents, and whānau with information about individual learning progress. As we noted in Section 5, the literature we reviewed displays a clear policy trend towards assessing and reporting progress, in contrast to the more traditional practice of assessment and reporting on point-in-time achievement (see, for example, Gonski et al., 2018).

In theory, personalised assessment allows for the different starting points and rates of progress to be made by different students—achievement is relative to where they started. As Masters (2017) asserts, "one of the best ways to build students' confidence as learners is to help them to see the progress that they are making over extended periods of time" (n.p.). Tools such as AsTTle and the Literacy Learning Progressions

(LLP) were developed to enable teachers to target teaching more directly to learners' strengths and needs, and to reinforce the idea that assessment is an ongoing and integral part of teaching (Ministry of Education, 2018). However, teachers require significant content knowledge about the characteristics of progression in different curriculum areas to target their teaching towards individual learning needs. And as we noted in Section 5, appropriate, evidence-informed progressions do not yet exist across the breadth of the curriculum. Building them is not straightforward but is now a critical system activity.

Four of the eight *NZC* curriculum principles—*high expectations*; *inclusion*; *cultural diversity*; and *Treaty of Waitangi* (Ministry of Education, 2007, p. 9)—focus attention on equity challenges, which are also in scope in this section.

We found an interesting instance of international recognition that personalisation of competency-based learning (and assessment) might be key to enacting this cluster of principles. The final report from a summit of educators with an interest in competency-based education in the US argues that a competency-based system has the best chance of achieving equity goals for all students (Sturgis & Casey, 2018). A number of reasons are set out to support this argument. One is that competency-based education replaces an emphasis on summative assessment by embedding a "personalized learning cycle" (p. 12) that aligns with intended outcomes in a more transparent way. Similarly, traditional grading practices are replaced by communication of learning progress. Sturgis and Casey say that the best way of ensuring shifts in curriculum and assessment practice are understood and accepted is to engage the local community in conversations about the competencies students need to develop, as well as ensuring that assessment processes are as transparent as possible.

Evidence of student-centred assessment practices

There is considerable advocacy for personalisation of assessment but we did not locate a strong evidence base demonstrating the impact of such practice. In fact, with the exception of assessments embedded in digital resources, much of the research we did find focuses on personalised *learning*, accompanied by more traditional assessments. To the extent that data from these assessments are used to make teaching and learning decisions, we could say that some personalised assessment *practices* are in evidence but the *enactment* of assessment is still largely standardised. In research projects with a traditional design, this is inevitable—if students in a control group are to be compared with students who received more personalised learning opportunities, then obviously all of them must be assessed in the same way. The next example illustrates this point.

One American research project (Pane, Steiner, Baird, & Hamilton, 2015) demonstrated that students in schools using personalised learning practices made greater progress over the course of two school years, compared with matched groups of students in others of the 32 schools in their study. Personalised assessment practices reported in this study included using data to understand student progress and make instructional decisions, along with time for individual academic support, and the use of technology for personalisation of learning. Students who started out behind made accelerated progress and could perform at or above national averages after 2 years. Pane et al. (2015) also reported that some strategies, such as competency-based progression, were less commonly used and more challenging to implement.

What does personalised assessment look like?

The New Zealand evidence for personalisation of an aspect of assessment practice is sparse. Parr (2016) examined the practices of teachers who were deemed to be exemplary because of their track record of successfully supporting students to make accelerated gains in writing. This study identified seven dimensions of personalised practices that accelerated progress in writing:

- acquiring and applying deep knowledge of students as writers
- making connections with, and validating, relevant cultural and linguistic funds of knowledge
- aligning learning goals in writing with appropriately designed writing tasks and ensuring that students understand what they are learning and why
- providing quality feedback
- scaffolding self-regulation in writers
- differentiating instruction (while maintaining high expectations)
- providing targeted and direct instruction at the point of need.

This combination of practices reveals the complexity and deep practical knowledge that teachers need to engineer learning environments that produce personalised and self-regulated learning (Hill, Ell, & Eyers, 2017).

Price, Smith, and Berg (2017) conducted a randomised experimental study in two Years 9 and 10 secondary English classes. Students in these classes indicated that they had not received useful formative feedback previously. Over two school terms, the study compared the relative effectiveness and efficiency of personalised feedback versus annotated exemplars as ways to improve student writing. Both approaches supported substantial growth in writing performance. Although students greatly preferred personalised feedback, there was no significant difference in achievement whether the feedback was given individually or via annotated examples. Providing personalised feedback took between two to eight times as much teacher time as giving annotated examples. Therefore, the study suggests that working with students using annotated examples may greatly reduce teacher workload without negative impacts on student achievement. A similar study (Perry, 2015) examined the impact of students receiving instant feedback in an economics class. This teacher reported positive impacts on student motivation and engagement. She also noted that, although worthwhile, the practice was time consuming for learners and the teacher.

Tolmie (2016) explored personalisation of learning and assessment practices in the context of an ILE. She notes that ILEs are all about personalised learning (and hence also assessment). The teachers in her thesis study believed that personalisation had increased achievement: “Kids who were kind of meeting N S and just above are now well above. And it’s because they got to self-direct and self-manage and take a personalised approach to their learning that they are succeeding much more” (teacher comment, p. 63).

Some evidence highlights ways that accountability pressures might count against personalising learning and assessment practices. For example, one recent small New Zealand study (Aitken, Villers, & Gaffney, 2018) focused on the teaching practices of three new entrant teachers. These teachers engaged their new learners in guided reading activities as soon as they started school, without any assessment or analysis of their existing competencies in oral language. They explained that they were responding to perceived pressure to have all children reading at Green Level after 1 year at school, in order to meet the National Standards. Observations of their practice showed that the same teaching was aimed at all children in the group regardless of their competencies and time at school, and feedback was not provided to individual learners.

Supporting self and peer assessment

Actively involving students provides another lens for thinking about ways to put students in the centre of assessment practices and decision making. The evidence is unequivocal that self and peer assessment practices are strongly associated with student achievement. Booth et al. (2016) have linked self and peer assessment with developing the skills of self-regulated learning. Gadd (2014) identified that engaging students in co-constructing goals and success criteria, then supporting students to self-monitor and self-regulate against those goals, was strongly associated with accelerated progress in writing.

Nevertheless, international and local research shows that students are more typically not enabled to assess their own work. Researchers in Sweden found that teachers still tended to take most of the responsibility for assessment practices, resulting in very high workloads, while the opportunity was missed to include students in assessment decision making and taking responsibility for their learning (Jonsson et al., 2015). Data from NZCER's national surveys suggest that teachers also take most of the responsibility for assessment practices in New Zealand classrooms (Bonne & Wylie, 2017; Wylie & Bonne, 2016). We have already reported that students themselves may be resistant to seeking and using formative feedback to support their learning (Harper & Brown, 2017; Searle et al., 2017).

While time pressures may inhibit opportunities that teachers offer for self and peer assessment, teachers' subject knowledge may also be a factor. Hawe and Parr (2014) noted that the primary teachers in their study were able to share broad goals for the writing expected of students, but they struggled with explaining the specific features and quality they were looking for. This could be why few teachers provided opportunities for student peer review and feedback on writing.

Several authors have identified aspects of teacher knowledge and thinking needed to build self and peer assessment capabilities in students (Casey, 2013; Hill et al., 2017; Katie White, 2017; Willis & Klenowski, 2018). These aspects include:

- a belief that the practice is likely to benefit learners
- a deep practical knowledge of quality outcomes
- willingness to allocate time to designing assessment criteria with students
- knowing how to support students to apply this knowledge to their works in progress, to have peer-to-peer assessment conversations, and to respond appropriately to feedback from teacher and peers.

In their major review of assessment of key competencies in Europe, Siarova et al. (2017) stress the role of peer and self-assessments when seeking evidence of non-traditional outcomes such as: initiative and entrepreneurship; learning to learn and social competence; critical thinking; creativity; problem solving; risk assessment; decision taking; and constructive management of feelings. However, they also note that criteria to judge these types of performances need to be better clarified and illustrated at the national level. Here, too, building teacher capabilities is a necessary precursor to shifting practice.

Affordances of technology for personalising assessment

Social tools such as blogs, groups and discussion forums can be used to involve students in peer assessment conversations. Casey (2013) notes that these tools can facilitate interactions between learners and the teacher, as well as with peers. They can also be used for reporting to parents. However, careful teacher structuring of both a project and its instructions is important to avoid confusion and information overload.

A review of studies of computer-based assessment for learning (CBaFL) in primary and secondary schools described three phases in the development of CBaFL resources and practices (Shute & Rahimi, 2017). In phase 1, classroom teachers draw on computer-based resources to give timely feedback to students. Studies have suggested that this feedback should be designed so students need to use rather than ignore it; and be elaborated rather than simple verification feedback. In phase 2 classrooms, teachers use web-delivered assessment for learning. In these online settings, assessment still tends to be *of* learning rather than *for* learning but CBaFL tools enhance learning by: keeping students engaged; providing the means for students to monitor their own progress; and developing their self-regulatory skills. Phase 3 classrooms are characterised by data-driven and continuous CBaFL. Shute and Rahimi say that research is on-track to deliver on the "great promise" of "high-quality, ongoing, data-driven assessment for learning" (p. 15).

They expect the innovative techniques being developed to move beyond the laboratory to the mainstream and, when that happens, the system will "no longer have to rely solely on high-stake tests for assessing students' knowledge and skills" (p. 15).

The concern has been expressed that large-scale commercial interest in personalisation of learning (and assessment) using an algorithmic business-minded model of personalisation will actually work against the equity intent of personalisation (Kucirkova, 2018). This will happen if commercial companies take "a product-centric approach to children's education, with no agency or reciprocity for the learner" (p. 22) where learning is seen to be essentially about "absorbing facts in nice packaging" (p. 22).

Taking personalisation of assessment a step further again, blockchain technology can act as a data management tool that: (1) makes sensitive data simultaneously more shareable and more secure; and (2) puts sensitive data into the hands of its users, allowing learners to take control of their own data (Briggs, 2018). This is potentially very disruptive to traditional assessment practices because it will also allow students to legitimise informal learning achievements and streamline knowledge transfer and (at least for older students) the job application process.

Some cautions are also expressed about blockchain technology. One blogger stresses that educational applications are still in their infancy and a lot of research and development is needed, including at the system level:

> While blockchain will make it easier to share credentials, it leaves wide open the question of who creates and grants certifications. Employers and educational institutions will need to decide what knowledge and skills are important and how those are developed and assessed. In some cases machine scoring will be able to verify certain skill claims, but in many cases with important and multidimensional skills, human judgement observation will remain important. (Vander Ark, 2017).

Every section of this report so far has discussed the challenges inherent in building teachers' assessment capabilities. Here, too, digital technologies can help. Technological developments make it feasible to assess students' work automatically when students are working in rich digital environments (Tondeur, van Braak, Siddiq, & Scherer, 2016). This use of technology not only frees up time, it also offers opportunities for teachers to gain immediate access to the samples of work from large groups of students working on open response tasks, categorising them to fit relevant characteristics for the teacher to attend to, and also freeing up teachers' attention for student solutions that are out of the ordinary in different ways. In this way, educative uses of assessment technologies can build teacher assessment capabilities.

Assessment of students with special learning needs

Many assessment practices fail to account for the learning and progress of students with specific learning needs. These students are frequently invisible in assessment data. According to McIlroy (2017), assessment practices in some New Zealand primary schools fail to recognise and value the learning of some disabled children who become marginalised within the curriculum available to their peers. Guerin (2015) similarly states that the "New Zealand education research remains silent in its considerations of the experiences of disabled students and their families within the assessment practices of New Zealand secondary schools" (p. 5).

Factors contributing to this situation could include:

- a lack of understanding of what should be assessed
- a lack of teacher knowledge of suitable assessment pedagogies to use
- the purposes for which assessment of students with learning support needs is currently carried out.

Each of these is now briefly elaborated.

A key recommendation in the literature is that *all* students should be taught and assessed in ways that allow them to access the full breadth of the curriculum (Burgon, Roberts, Darr, & Hipkins, 2017; Morton, 2014). Specifically, students with learning support needs should not have assessment of their progress restricted to reporting on the development of social and life skills. Despite steady progress towards inclusion of students with special educational needs, ERO (2015) found that monitoring of their progress tended to be focused on participation rather than on analysis of student learning or the identification of teaching strategies that had been effective for particular groups of students. The ERO report made several recommendations focusing on developing detailed goals, particularly for those students who learn within level 1 of *NZC*, as well as building teacher capability in using and sharing differentiated teaching strategies.

A subsequent exploratory project (Burgon, Eyre, & Stevens, 2017) scoped ways that schools might develop their capacity to recognise and improve the progress of this group of students. The researchers recommended the development of finer grained descriptions of learning achievements in the Learning Progressions Framework. Since then, the Mathematics Learning Progressions framework has been elaborated, and work on a framework for communication progressions is ongoing. Burgon, Eyre et al. see potential for these progressions to be validated and included as part of the PaCT Tool, which would support engagement with parents and whānau by sharing evidence of progress in the new and elaborated frameworks, utilising information from educational and home contexts.

The use of learning stories (Bourke & Mentis, 2014) or narrative assessment (Guerin, 2015) provides a pedagogical approach to capturing learners' development in a way that is child-centred, non-standardised, and embedded in authentic contexts. Narrative assessments can utilise data that include observation, photos, videos, assessment tools that show progress over time rather than relying on a single source of information. These researchers say that narrative assessment recognises and describes authentic learning and does not seek to quantify children's learning against each other or against predetermined measures.

Although these approaches are not new, there is little research on their utilisation in New Zealand primary and secondary schools. In 2007, Bourke and Mentis (2014) undertook a survey of 964 New Zealand primary and secondary teachers with special needs students in their classes. Results revealed a preference for using evidence-gathering methods that the researchers describe as criterion-referenced approaches (in contrast to standardised tests). Methods used by most surveyed teachers included: making observations (94%); collecting examples of work (94%); keeping anecdotal records (88%); and organising portfolios (80%). Guerin's research in a secondary school focused on two Ongoing Resourcing Scheme (ORS)-funded students (Guerin, 2015). While narrative assessment did support their learning, she also identified tensions when assessments were carried out by some non-education professionals if these "fulfilled contractual agreements but did not reflect a knowledge of the learner other than in single contexts" (p. 213). She said that, in some cases, the students disengaged from the assessment process and were not able to show the range of their capabilities.

Morton also draws attention to the purposes for which assessment is carried out when students have special learning needs (Morton, 2014). She comments that this purpose has been to decide which students get access to which types of education. Even today, assessment of children with learning support needs tends to emphasise what they can't do, so as to gain access to additional funding. However, the concluding comment of an unpublished report for NZQA makes it clear that technologies that aid (learning and) assessment offer the prospect that these attitudes will begin to change. Very recently, Burgon (2018) investigated use of Special Assessment Conditions (SAC) to support NCEA assessments. She noted the findings of a Ministry of Education SAC review (2014), and a more recent paper by Munro (2016). These papers point to a future where

> the long-term outcome would be to reduce or eliminate most SAC, by supporting schools with: multiple assessment options; the knowledge of technologies that would achieve that goal; and resources to enable use of those technologies. This would achieve the goal of making special assessment conditions for some the standard assessment conditions for all (Burgon, 2018, p. 36).

This argument draws attention to the possibility of using Universal Design for Learning (UDL) approaches to meet the needs of all learners by differentiating learning opportunities. UDL practices increase student ownership and engagement and contribute to the "learner at the centre" principle by allowing for individual differences. Use of UDL for assessment has been advocated and modelled for some years (see, for example, Rose, Meyer, Strangman, & Rappolt, 2002) but we have not found empirical evaluations of its use in New Zealand. In any case, Rao, Ok, and Bryant (2014) reported that teachers use UDL principles in varied ways, which does not allow for ready comparisons across studies.

How high-stakes assessments contribute to equity challenges

Finally in this section we turn the equity question on its head to report on arguments that assessment practices can actually *construct* inequalities in achievement. This concern is most often expressed in the context of international assessments but the same arguments have been made about local assessment contexts as well. Sellar, Rutkowski, and Thompson (2017) claim that PISA tests how well young people can complete tests such as PISA, when a range of factors can impact overall results. These include gender (boys tend to be advantaged); socioeconomic status (poverty is a factor in poor performance); and level of school preparation. In one study they cite, being of Chinese origin was associated with higher achievement, regardless of whether a student lived in China, Australia, or New Zealand.

International test results typically report disproportionate numbers of Māori and Pasifika students as underachieving, as do NCEA analyses. May (n.d.) notes that there has been very limited research internationally that critiques the "technicist, objectivist" assumptions that mask the "cultural locatedness" of traditional assessment practice (p. 2). He notes that context, concepts, and the language of the assessment questions are all likely to be biased towards the cultural norms of the dominant social group. He explains in considerable detail how traditional practices can help construct inequalities in both summative and formative assessment practice.

May says that such critique as exists is mostly aimed at summative assessment practices, but AfL is also implicated in constructing cultural norms that disadvantage minority groups. As one example, fostering a degree of learner independence is problematic when minority students are more likely to feel disempowered because the locus of control lies with the majority culture. As another example, meaningful performance assessments, which in theory allow students to express their learning in culturally relevant ways, may not be appropriately supported by teacher feedback. Teachers themselves are bound by cultural norms, with the result that even in performance tasks, minority students are likely to go through the motions of metacognitive reflection etc. Rather than actually reflecting on ways in which their own thinking and learning might be different from dominant groups, they reproduce shallow versions of what they think is expected. (May cites NCEA as a context that illustrates this problem.) These assessment issues are compounded when, on the basis of test results that have already disadvantaged them, minority students are assigned to lower ability groups that do not receive the same rigorous learning opportunities as those assumed to be more able. Any narrowing of the curriculum in response to perceived pressures of high-stakes testing can further disadvantage minority students because the subjects that are relegated are more likely to be those in which they might thrive.

In summary

This principle addresses personalised learning and assessment, which provides learners, teachers, parents, and whānau with information about individual learning progress. Personalised assessment focuses on tailoring assessment to promote the best learning outcomes for each child. This involves selecting and adjusting assessment methods, adapting content, personalising feedback, and being inclusive. It also includes being mindful of how individual students are affected by assessment practices.

Personalised assessment allows for the different starting points and rates of progress to be made by different students—achievement is relative to where they started. Clear progressions should provide a necessary starting point for making such judgements.

There is considerable advocacy for personalisation of the act of assessment but we did not locate a strong evidence base demonstrating ways this can be achieved in non-digital contexts. When evidence of impacts is reported, the research tends to focus on personalised learning, accompanied by more traditional assessments.

Self and peer assessment practices are associated with accelerated achievement gains. These practices do not appear to be widespread (just as AfL pedagogy in general is not widespread). Aspects of teacher knowledge and thinking needed to build self and peer assessment capabilities in students include: a belief that the practice is likely to benefit learners; a deep practical knowledge of quality outcomes; willingness to allocate time to designing assessment criteria with students; and knowing how to support students to apply this knowledge to their works in progress, to have peer-to-peer assessment conversations, and to respond appropriately to feedback from teacher and peers.

Computer-based assessment for learning (CBaFL) can provide timely personalised feedback to students and allow them to monitor their own progress and develop their self-regulatory skills. These technological developments make it feasible to assess students' work automatically, freeing up time for both student and teacher learning. Educative uses of assessment technologies can build teacher assessment capabilities.

All students should be taught and assessed in ways that allow them to access the full breadth of the curriculum. However, students with specific learning needs are frequently invisible in assessment data. Factors contributing to this situation could include: a lack of understanding of what should be assessed; a lack of teacher knowledge of suitable assessment pedagogies to use; and the limited purposes for which assessment of is carried out (e.g., primarily for additional funding rather than to open up access to the curriculum). Narrative assessment is seen as a useful pedagogical approach to assessment of and with students with specific learning needs.

Universal Design for Learning (UDL) practices increase student ownership and engagement and contribute to the "learner at the centre" principle by allowing for individual differences. Despite advocacy for and modelling of UDL, we did not find any empirical evaluation of its use in New Zealand. Doubtless this reflects a lack of funding for such research, and indeed a lack of opportunities for systematic studies of classroom assessment practices in general.

7.

Themes related to drawing on a range of evidence

Principle: A range of evidence drawn from multiple sources potentially enables a more accurate response

This section focuses on the challenges of bringing together varied evidence sources to make an overall judgement about the learning that has been demonstrated in a specific assessment context. Previous sections have discussed the importance of expanding beyond traditional test-based sources of evidence, so that newer curriculum elements such as key competencies can be meaningfully assessed. Because this already implies that a range of evidence is needed, we do not begin by justifying the importance of this principle. Taking that as a given, we ask:

- Are there ways of gathering evidence that should be considered but have not yet been discussed?
- What is known about the challenges to be expected when bringing multiple sources of evidence together?

Performance-based assessment

There have already been indications of an answer to the first of these questions. The discussion of the curriculum principle noted that at least some assessments need to be performance-based if the *NZC* key competencies are to be meaningfully assessed. We now outline what is meant by performance-based assessment and the nature of outcomes that might be documented using this type of assessment pedagogy.

The literature yielded some examples of performance-based assessment tasks and/or their characteristics. Collectively, these examples imply that any rich learning task might qualify—if what students *actually do* is the assessment focus. One survey of a range of performance-based assessment initiatives (Darling-Hammond & Falk, 2013) lists the following as suitable task types:

- extended problem solving
- framing and conducting investigations
- analysing and synthesising data
- applying learning to new situations
- explaining and defending thinking in relation to an issue or argument.

The Alberta Assessment Consortium[6] provides a range of performance-based tasks for schools in the province, and advice about using them effectively, including a rubric template. These materials are behind a paywall but the public part of the site identifies the following as characteristics of effective performance assessments. They:

- engage students in their learning
- develop critical thinking skills
- encourage innovation and problem solving
- provide quality assessment evidence
- are based on the [Alberta] curriculum.

The "in-depth" assessment tasks used in the National Monitoring Study of Student Achievement (NMSSA) are broadly performance-based (Educational Assessment Research Unit, n.d.). It could be interesting to compare and contrast the characteristics of these tasks, as used in different learning areas for the in-depth component of the current NMSSA programme, to exemplify and elaborate on the use of performance assessments in the New Zealand context.

We also found several international examples of performance-based assessments for school exit qualifications:

- UK: The Extended Project Qualification (EPQ) is assessed in four categories: managing a project; use of resources; realising the project; and review.[7]
- South Australia: The South Australian Certificate of Education (SACE) offers final-year students opportunities to undertake two extended research projects. In each project, students are required to demonstrate one or more of the ACARA general curriculum capabilities.[8] Interestingly, it is up to students to interpret what an excellent demonstration of the named capability could look like, and then provide evidence based on that understanding.[9]
- Performance-based assessments for qualifications are also possible in some subjects in New South Wales[10] and for some subjects they take the form of an extended Personal Interest Project (PIP).

Some NCEA assessments are also based on performances of learning. This is most obvious in the arts (for example, Thorpe et al., 2017), but also arguably applies in other learning areas, where assessment is based on students *doing something specific* to demonstrate their learning. For example, the Learning Languages achievement standards include some that require students to "*interact" as they demonstrate their language capabilities.* A recent literature survey of research in the impact of NCEA on language learning (Stevens & Hipkins, 2016) outlined research that showed impacts on pedagogy. These "interact" standards have moved assessment away from summative tests towards the collection of ongoing evidence of authentic and unrehearsed spoken interactions. A teacher survey confirmed that authenticity was the main advantage of the new standards. Some said the assessment was less stressful for students because it was less test-like. The main disadvantage was the impracticality—specifically, technical challenges of recording interactions.

Collectively, the diverse examples just outlined arguably demand demonstrations of capabilities similar to those that have been recently developed as potential focuses for assessment when *NZC* key competencies are woven together with curriculum area content in rich tasks (Hipkins, 2017; McDowall & Hipkins, 2018). Therefore, support for building teachers' capabilities in carrying out robust performance assessments might be one way of supporting more widespread curriculum weaving.

6 https://aac.ab.ca/materials/assessment-materials/
7 https://www.aqa.org.uk/subjects/projects/aqa-certificate/EPQ-7993
8 See http://www.acara.edu.au/curriculum/general-capabilities. For all Australian states these capabilities are named as: Literacy; Numeracy; Information and communication technology capability; Critical and creative thinking; Personal and social capability; Ethical understanding; Intercultural understanding.
9 See https://sites.google.com/site/saceresearchproject/the-folio/match-the-research-question-to-a-capability
10 http://educationstandards.nsw.edu.au/wps/portal/nesa/11-12/hsc/rules-and-processes/practical-performance-exams

Evaluative research suggests other potential benefits of making more use of performance assessments. Both the EPQ and the SACE websites include some evaluative information about the impacts of learning that are assessed as a performance, albeit assembled by the providers of each qualification. A summary of four research studies of the EPQ (Drummond, 2017) asserts that undertaking an EPQ is associated with: higher performance in traditional senior secondary assessment (A levels); better degree-level performance in subsequent university studies; enhanced self-management of learning; and greater motivation to learn. The most recent SACE report from the Chief Examinations Officer comments on the increasing levels of sophistication students are now showing in all aspects of these research projects (SACE Board of South Australia, 2018).

The survey of initiatives to implement performance-based assessments in different American states also reported on benefits of this type of assessment (Darling-Hammond & Falk, 2013). They said that performance-based assessment had a powerful impact on practice because teachers needed to become engaged in content-specific moderation activities. These became, in essence, effective forms of professional learning that subsequently resulted in significant gains in student achievement because they kept teachers focused on providing the learning support that students needed.

We also found one study of students' perceptions of performance tasks that had been designed for the former National Education Monitoring Project (NEMP) assessments (Smith, Gilmore, Berg, Smith, & Jameson-Charles, 2012). Analysis of achievement in a range of tasks, from several different curriculum areas, was statistically related to students' perceptions of those tasks. The researchers found a clear relationship between liking a task and doing well at it, but said the relationship between motivation and achievement is not straightforward to elucidate, and that more research is needed.

Moderation as a professional learning opportunity

Performances of any sort are complex acts and hence making judgements about them is not straightforward. Moderation processes are primarily designed to support greater consistency in making such judgements but, as we have already noted, they can also support powerful teacher learning (Darling-Hammond & Falk, 2013). The literature we now outline points to the difficulties in doing moderation well, but also reinforces the possibility of rich professional learning opportunities when moderation is conducted in conditions conducive to that.

A small study of the moderation of writing in one Auckland primary school illustrates why moderation is "by no means a straightforward process" (Hipkins & Robertson, 2012, p. 42). The researchers observed the teachers discussing multiple dilemmas when making sense of the writing samples. These dilemmas included:

- the relative amount of attention to pay to surface and deep features of writing
- what to do when one work sample was of variable quality
- how to balance form and flair (caution and experimentation)
- whether the criteria adequately captured variations in achievement.

The teachers also worried about fairness and the impact of their decisions on students' motivation. Dilemmas here included: whether effort should be recognised as well as achievement; whether they should allow for recent teaching of some aspects in some classes, but not others; perceptions of the ability of different students; and what was fair when judging the work of ESOL students.

A comparative study of moderation in three different schools (Smaill, 2018) reported that teachers were able to use their participation in social moderation to improve their understandings of AfL principles and practices. The participating teachers learnt about factors that affect the dependability of student assessment information and believed that involvement in social moderation had contributed positively to

their assessment capability. However, they also experienced qualitatively different professional learning opportunities, linked to a series of school-specific conditions. These conditions included: the amount of time schools committed to moderation; the types of moderation activity teachers engaged in; and the nature of the rationale that teachers developed to sustain their involvement in moderation. Hipkins and Robertson (2012) similarly described features that support professional learning opportunity:

- participation is carefully structured and facilitated
- taking learning risks together is a habitual way of being for the teachers in the school
- staff turnover is low
- post-moderation actions share decision making with students (and follow AfL principles)
- the overall combination of school structures and processes (not just specific moderation meetings) allows multiple accountabilities to be balanced.

In the NCEA context, Hipkins et al. (2016) describe changes to moderation processes over the years since NCEA was introduced. They chart a shift from an initial emphasis on using moderation to support teacher learning towards a strong accountability focus that arguably acts against open-minded learning and exploration of curriculum and assessment possibilities.

There appears to be a paucity of New Zealand research with a specific focus on moderation as a form of teacher meaning-making and learning when considering individual pieces of evidence. As we next outline, greater attention appears to have been directed towards ways in which teachers combine assessment information from different evidence sources.

Making overall teacher judgements (OTJs)

Teachers have always needed to bring diverse pieces of evidence together to make overall judgements of students' progress, at least informally (e.g., when writing report comments about different aspects of learning). How they do so has recently become a focus for research attention for several interconnected reasons:

- The stakes are higher when OTJs generate data that will be used for accountability purposes (as with the National Standards in reading, writing, and mathematics). Consistency and fairness become matters of more apparent concern in these contexts.
- Making sound OTJs is more complex when evidence from multiple specific learning progressions must be brought together (again as in the case of the National Standards). By comparison, generating traditional measures was more straightforward, if less informative (e.g., generating an average mark or grade from a series of point-in-time assessment events).

Different research projects paint a similar picture about how, and how well, teachers make OTJs. All four studies that follow are set in the context of National Standards, where OTJs were needed to judge whether students were at, above, or below a specified standard.

Bonne (2016) draws on primary teacher responses from the 2016 NZCER national survey. Most teachers said they used classroom observations, classroom work, and one or more assessment tools (some standardised, some not) to make OTJs. By contrast, a quarter or less used student self and peer assessment. In 2016, just 9% were using PACT to make OTJs. Seventy-four percent thought the moderation work they did to make OTJs had given them useful insights into their practice.

Via semi-structured interviews with 30 teachers in 10 primary schools, Poskitt and Mitchell (2012) investigated understandings of OTJs and the processes used to make them. The teachers in this study had all been part of an AfL initiative. Nevertheless, they varied in how they conceptualised OTJs and they employed a range of approaches when making them. Use of moderation processes was limited. Across two cycles of data gathering they described these teachers as "surrounded by uncertainty and confusion" which meant they were unlikely to generate "highly valid and reliable data" (p. 72).

Ward and Thomas (2016) conducted a 5-year study of implementation of National Standards in New Zealand primary and intermediate schools. They drew on a variety of evidence sources to conclude that OTJs lacked consistency in many (but not all) instances. However, they did also report an increase in use of moderation processes between 2010 and 2013, based on principals' survey responses.

Using a large New Zealand database, another team explored the relation between psychometrically designed standardised achievement results and teacher judgements in reading (*N* = 4,771 students) and writing (*N* = 11,765 students) using hierarchical linear modelling (Meissel, Meyer, Yao, & Rubie-Davies, 2017). This team reported that priority learners were being systematically assigned lower OTJs, even when their standardised achievement scores were the same as other students. They found that the teacher judgements were inversely related to the classroom and school achievement profiles. They expressed concern about these discrepancies because they have important implications for equitable educational opportunities, particularly as ability groups are an "entrenched" practice in New Zealand schools.

Collectively, these reports point to the difficulty that teachers experience in making consistent and fair judgements of student progress. Robust moderation processes should help, but again the picture painted here is of limited use of these. Bonne notes that experts such as John Hattie and Vince Wright have pointed out how difficult it can be for experts to reach consensus judgement over multiple pieces of evidence. Therefore, they say, it should not be surprising that teachers have found this hard to do (Bonne, 2016).

Digital technologies: New possibilities for diversifying collection and judgement of evidence

Smart technologies allow online, anytime, anywhere, and on-demand assessment (Murgatroyd, 2018a). These can include multimedia-based assessments, where learners share video or audio (or other media) as the basis for peer and instructor assessment. Evidence collected in this way opens up new possibilities for collaborative assessment conversations between the learner and their teacher. The tertiary sector is also beginning to use simulation (including 3D simulators) for assessment. An example is the way in which pilots are assessed for competency. We did not find any equivalent examples in the school sector.

As well as innovating ways to gather evidence, new technologies are opening up new ways to manage the complexities of assessment decision making. In a comprehensive survey of new developments, Murgatroyd notes that automated marking is now possible for all forms of assessment, including video, audio, essays, multiple choice, and short-form writing, driven by artificial intelligence (AI) (Murgatroyd, 2018a, 2018b). These developments are changing the balance between formative and summative assessment because online learning systems have adaptive assessment functionality that is being used increasingly as a basis for continuous assessment and support of learning (Murgatroyd, 2018b).

Other recent research has described an extension of comparative judgement processes, made by using fuzzy logic mathematical models (Ayca & Hasan, 2017). These models can create multi-criteria hierarchies that can be applied to assessment of complex projects using AI. Experts have an active role in establishing the hierarchies of criteria but once the model is built AI takes over the decision making.

Capturing data about learning in learning management systems (SMS/LMS)

AI researchers argue that attention needs to be given to the student/learning management systems used to collate and process data (Gulson, Murphie, Sellar, & Taylor, 2018). These systems have the potential to present outputs in readily readable forms (e.g., dashboards). How meaningfully they do so will obviously depend on the nature and quality of the data captured in the system being used.

We could find little systematically collected New Zealand evidence about use of SMS for purposes other than recording summative data. One article described responses from secondary teachers and principals to a small number of questions about SMS included in the 2012 NZCER National Survey (Hipkins & Dingle, 2013). The overall need identified was for systems that are easier to use, and for professional learning about their efficient use. More recently, the 2016 NZCER National Survey of Primary and Intermediate Schools identified an additional need for SMS for longitudinal tracking capacity across the schools in a Kāhui Ako, so that they can share information that travels with the student as they transition between learning organisations (Bonne & Wylie, 2017). They note that students in low decile schools are more likely to be impacted when data cannot be readily shared between systems because transience rates are higher in these schools.

A small study (Edmunds & Hartnett, 2014) investigated the potential of the one school's SMS to be used to personalise learning for students. For the three teachers in the study, doing so effectively was interrelated with effective use of AfL pedagogy and seeing the SMS as a tool that could support learning.

Nor did we find any systematic analysis of the ways in which data are organised in relation to the broader curriculum goals signalled by *NZC*. The Linc-Ed NZ system claims to offer a facility for "progressions-based reporting" which offers "powerful features to automate the process of goal setting, reporting and tracking progress" (Linc-Ed website).[11] As far as we can tell from the promotional videos, this facility is activated by selecting individual achievement objectives, which are then treated as discrete entities for data collation and reporting purposes. This linear, additive model is at odds with the *NZC* weaving approach described in earlier sections of the report. It does not appear to reflect a good understanding of the intent of *NZC*. Nor is it clear how "progression" can be extracted from the data recorded from individual assessments.

We found some commentary about what it takes to build fit-for-purpose SMS, and their likely impact on reporting practices in schools (Heard & Hollingsworth, 2018). In brief, the potential now exists for continuous reporting rather than waiting until a fixed time in the year, which will make it possible to analyse and report on actual progress over time. However, many school reports currently misuse the term "progress" or "progression" to report "performance" (i.e., a moment-in-time level of achievement rather than a meaningful measure of change over time). While some SMS used in Australia are moving to embed features that potentially allow schools to analyse and report progress, Heard and Hollingsworth's research (as yet unpublished) suggests that these newer features are not being widely used in Australian schools.

In summary

Bringing together varied evidence sources to make an overall judgement about the learning is important for several reasons. The wording of the principle implies that the most important purpose is to allow *more accurate* judgements to be made. While this might be so in principle, the evidence suggests that actually making consistent overall judgements based on multiple pieces of evidence is challenging, and is often not carried out consistently. Teachers need support to do this well, including opportunities to participate in moderation conversations that are structured and facilitated in ways that support their professional learning and expand their curriculum thinking.

Another way of thinking about the importance of this principle is to focus on the "multiple sources" per se. As one example, this principle opens up new opportunities for making greater use of performance assessments, which require students to show how they *put their learning to work* to address specific challenges. The literature describes a number of characteristics of tasks suitable for performance assessment. Such tasks would not be out of place in many classrooms right now, and they readily lend

11 https://linc-ed.zendesk.com/hc/en-us/articles/360000078496-Progressions-based-reporting

themselves to demonstrations of competencies. There is some evidence that performance assessments can confer additional learning benefits for students—and for teachers—when they are well moderated.

Digital technologies are rapidly expanding the potential repertoire of evidence sources that might be collated. The manner in which data from multiple evidence sources are collated and stored is important—traditional learning management systems tend to be designed in ways that support linear and additive ways of thinking about curriculum progress (or, more accurately, point-in-time achievement). Both structures and uses of SMS need to change to keep up with new curriculum thinking, and to make the most of rapid developments in other aspects of AI, including machine-marking and judgement-making.

8.
Themes related to quality interactions and relationships

Principle: Effective assessment is reliant on quality interactions and relationships

This principle is potentially very wide in scope. The word "effective" draws attention to the ends for which quality relationships are the means. Many matters pertinent to this principle have been raised in previous sections so the importance of the principle has already been established. In this section, we briefly reframe ideas already covered, and add some further material, to ask:

- Do further insights emerge when the focus is on quality interactions and relationships?
- What new opportunities and challenges might be anticipated in relation to this principle?

Interactions between teachers and students

Understandings of the power of formative assessment were dramatically changed following Black and Wiliam's seminal work that demonstrated strong learning gains when students are actively involved in assessment (Black & Wiliam, 1998). On these foundations Afl practices were built, predicated on strong learning relationships between students and their teachers. As we have already outlined, a consistent theme of the AfL literature is that giving effect to this type of assessment pedagogy is easier said than done. Teachers need to make time for interactions with students, and for peer–peer interactions between students, with a clear focus on what counts as quality work (Booth et al., 2016). If students' active role in assessment is not valued, or teachers do not have the interpersonal skills to build safe relationships with students, this is unlikely to happen. Our impression is that the *relationships* aspect of AfL has received less research attention than building teacher knowledge and skills to respond to students' thinking (i.e., the *cognitive* aspects).[12]

Another less visible aspect of teacher–student assessment interactions concerns the assumptions teachers might make about students' abilities, and the associated expectations they then have. Holding lower expectations for some students has already been highlighted as a problematic aspect of making

12 However, we did read the AfL literature with the idea of assessment capabilities to the fore. We would need to go back and reread the various papers to check this impression.

OTJs (Meissel et al., 2017). Misplaced cultural assumptions can also be problematic when they contribute narratives of underachievement (Cavino, 2013). New ways of framing learning competency as Māori or Pasifika students are needed (Houghton, 2015). This theme will be teased out more fully in the companion paper to this one, which will address assessment from a te ao Māori perspective.

Interactions that support effective professional learning

The international Afl literature emphasises the role of leaders in supporting effective professional learning, and in minimising the impact of perverse incentives from high-stakes assessments (Davies et al., 2014; Hayward, 2015). New Zealand studies also emphasise the necessity for leaders to be engaged learners in their schools and in the recently established Communities of Learning | Kāhui Ako. The Education Council recently commissioned a series of "thought pieces" about the challenges of leading a Kāhui Ako. One paper sets out a demanding profile of the qualities needed of these leaders (Robertson, 2015). Collaboration with other leaders, within and across contexts, to think, and to transform the system of education is one of the wide range of qualities specified. She also notes that effective leaders are digitally confident and competent in e-learning communities and understand the potential of technology, networks, and connectedness for enhancing learning. They are comfortable with ambiguity, complexity, and not-knowing as they learn and adapt within their leadership practice (Robertson, 2015). There are interesting resonances here with the curriculum shifts described in the 21st century literature (see Section 5 of this report).

Teachers also need to be comfortable with ambiguity and not-knowing as they explore new ideas about learning with their students. Support and feedback from their peers is important as they try new pedagogies such as AfL, or apply their expertise to the development of less familiar curriculum outcomes. However, the most recent NZCER National Survey of Primary and Intermediate Schools reported a lack of progress in the development of professional learning cultures, with little change in: teacher sharing; an improvement focus in work together; timely support; coherence in school professional culture; and getting useful feedback (Bonne & Wylie, 2017). The researchers say that these findings suggest "further support is needed for schools to get more out of collective teacher inquiry" (p. 23).

The picture looked a little different when secondary schools were last surveyed in 2015 (Wylie & Bonne, 2016). Compared to earlier surveys, there had been modest gains in the proportions of teachers reporting that they had opportunities such as: support to take risks in their teaching; departmental conversations with a focus on lifting achievement; and access to specialist subject advice. However, there had been no gains in other aspects of professional learning and support, such as getting practical help with engaging Māori and Pasifika students. While not directly about assessment per se, all these opportunities have important implications for making sustainable shifts in assessment pedagogies.

Initial Teacher Education (ITE) teacher educators and schools need to build effective relationships, so that early career teachers are effectively supported to build strong and effective assessment pedagogies. One recent TLRI study, set in four different New Zealand ITE contexts, reported that graduates of ITE programmes leave with a beginning understanding of the roles that students might play in their own assessment. However, they need ongoing support during their 2 years of provisional registration and, for that to happen, closer links are needed between universities and school. It is important to work together to build shared understandings of assessment that will be consistently modelled in practice (Smith et al., 2014). Similar findings are reported in another study set in one ITE context (Hill et al., 2017). Another study that reached similar conclusions is a thesis study that tracked the growth in assessment literacy of three beginning secondary science teachers (Edwards, 2017). These teachers graduated from their ITE programme with sufficient knowledge to carry out summative assessments but they still required support with some aspects of NCEA and assessment task design. Practicum experiences during their ITE had a

strong impact on their growth in knowledge, indicating that mentors during practicum needed to be carefully chosen.

The need for ongoing support for leaders of professional learning and for peer mentors is a clear implication from all these studies. Significant shifts in assessment practice will not happen without sustained, skilful support for ongoing professional learning, at every career stage.

Relationships with families

Relationships between home and school are important. They help ensure that curriculum and its associated assessments are grounded in students' lives and experiences, allowing them equitable opportunities to demonstrate their learning progress. Genuine collaboration between home and school can help build insights about each student's learning progress and challenges (assuming robust progressions are available, as discussed in earlier sections).

Many schools are using digital technologies to communicate point-in-time achievement with parents but two-way digital collaboration does not yet appear to be common. We did find one example of a local digital innovation designed to provide real-time feedback to parents about their child's learning progress. *SchoolTalk* is a cloud-based platform to which teachers at Stonefields School upload learning designs, using a dashboard that shares their learning calendars and planning with other teachers and learners. Clear learning intentions describe the skills, knowledge, attitudes, and values that the student needs to learn. Goals and success criteria (in the form of school-designed progress rubrics) are used to record how well students are progressing, and where their next learning steps are. Students and parents can access these plans, along with supporting resources. A recent evaluation, carried out by the school, showed majority support for the idea that *SchoolTalk* adds value to teaching, learning, and communication between home and school (Stonefields School, n.d.).

There is some evidence of the effectiveness of other types of family–school relationships. Formal interactions (e.g., parent evenings, student-led conferencing) and informal encounters at school provide episodic opportunities for families to interact face to face with their child's teacher(s). The NZCER national surveys include a version for parents. Responses to the most recent primary and intermediate survey suggest that many are generally happy with these interactions. In 2016, a majority of parents and whānau thought their child's teachers were responsive to any concerns they had, and more than half strongly agreed they: felt comfortable talking with their child's teachers and asking about their child's progress; felt welcome in the school; and would recommend the school to others. Compared to 2013 responses, a slightly increased proportion of parents and whānau viewed schools as being respectful and inclusive of their child's cultural identity (Bonne & Stevens, 2017). Similar trends were reported in the 2015 secondary survey. Compared with previous surveys, more parents thought they were getting good information from the school about their child's progress (74%, compared with 63% in 2012 and 53% in 2009), and being genuinely consulted about new directions or issues (47% in 2015, 41% in 2012, and 34% of parents in 2009) (Wylie & Bonne, 2016).

The Teaching and School Practices Survey Tool provides a source of data that tells the story from the perspective of school professionals. In 2017, three of the 12 school-wide practices rated as particularly strong were related to school/home interactions: we welcome questions from parents and whānau about their child's learning in the school; we provide parents and whānau with opportunities to learn how to effectively support their child's learning at the school; and we seek and are responsive to parents' and whānau views about their child's learning (Wylie, McDowall, Ferral, Felgate, & Visser, 2018). The picture at the level of individual teacher practice is more cautionary. The following were among the practices that stood out because they were rated as strong by fewer than 25% of teachers: collaborate with the local

community so that their expertise can be used to support learning in class or other school activities; support the local community by ensuring that students have opportunities to actively contribute to it in ways valued by the community; collaborate with parents and whānau so that their expertise can be used to support collective learning in class or other school activities; and use the knowledge that parents and whānau have about their child to support the child's learning. Collectively, these data imply a level of comfort with high-level communications between home and school, and also that genuinely collaborative relationships that impact on practice within the classroom are much less likely to happen. A recently published re-analysis of ERO school visit data tells much the same story (Education Review Office, 2018a).

Thrupp and White (2013) interviewed parents in a number of schools as part of their study of the impact of National Standards. They said that parents wanted well-rounded information about their child's learning. Progress, attitudes, and socialisation were all important to them. Most parents did not understand National Standards, even in schools that had made the most effort to inform them about these. Many thought the judgement was made on the basis of a test and most appeared not to know anything about the challenges of making OTJs. Thrupp and White warn against using "simplistic" measures when aiming for clear communication. If reporting progress in the most familiar types of learning outcomes (reading, writing, mathematics) is challenging, we wonder about the implications for reporting on new and different learning outcomes such as those discussed in Section 5 of this report. The literature discussed in the next paragraph implies that one new strategy might entail collaborative involvement of at least some parents in building new reporting and data management systems.

Section 5 noted the potential of SMS/LMS to present outputs in readily readable forms (e.g., dashboards) for communication purposes. How well they can do so depends on the nature and quality of the data captured in the system being used (Gulson et al., 2018). Some of the literature we found looks beyond individual examples of local innovation to consider implications of these types of developments for the system as a whole. It is already possible to build complex "knowledge infrastructures" that raise important questions about ownership and power (Buckingham Shum, 2018). Buckingham Shum lists the following as questions that need to be addressed:

- How do we engage with the teams designing the platforms our schools and universities may be using next year?
- Who owns the data and algorithms, and in what senses can an analytics/AI-powered learning system be "accountable"?
- How do we empower all stakeholders to engage in the design process?
- Since digital infrastructure fades quickly into the background, how can researchers, educators, and learners engage with it mindfully?

Buckingham Shum says that the learning analytics community is wrestling with these questions, mindful of the challenge that "human factors" can make or break the use of the innovations they build. However, the implications arguably spread well beyond this academic community, and need to be addressed at the whole-system level.

An increasing emphasis on collaboration

Collaboration is consistently included in lists of 21st century competencies (Voogt & Pareja Roblin, 2012). It is seen as vital for both work and life in a globalised world (OECD, 2018). Building relationships for digital collaboration is now considered to be so important that it is becoming a specific assessment target. International research efforts are being made to investigate ways of assessing collaboration in digital learning environments (see, for example, Horwitz, 2018; Siddiq, 2016). Some researchers say that the skills

needed for digital collaboration are so complex that bespoke assessments should be designed (Care & Vista, 2017). As we noted in earlier sections, evidence-centred design processes are seen as important for meaningful alignment of the activity to be completed, and the underlying data capture and analytics. These digital assessment resources are expensive to design and trial, and not easy to get right. Validity and reliability issues are complex and need careful consideration (Shute & Rahimi, 2017). All the system-level dilemmas related to purchasing expensive goods and services apply.

One research team associated with the ACTS21[13] project has recently reported widely varying levels of capability among 11–15-year-olds in four different nations (Wilson, Gochyyev, & Scalise, 2016). In their trial of an innovative digital assessment approach, some students collaborated easily and produced professional-looking reports while others struggled. The researchers expressed concern about a growing divide between haves and have-nots when some students have clearly had opportunities to build collaborative skills while others do not appear to have been taught them. There are clear equity implications when students do not have similar access to important learning opportunities.

Building a connected, collaborative system

One of the underpinnings of the Tomorrow's Schools reforms was the idea that competition is an important driver of quality in the school system. A cost of this system change was that professional learning networks broke down and there was no systematic way to circulate important new professional knowledge within the system. One critical commentary estimates that more than a decade of progress and professional growth was lost as a consequence (Wylie, 2012).

A recent ERO report emphasises the importance of building collaborative cultures where responsibility for the success of all students is shared among all members of the community (Education Review Office, 2017a). This requires shifts in thinking and practice because their own school is the teaching and learning community most participants know best. Teachers who work across a Kāhui Ako also have specific development needs for these roles. They need to build close relationships with individual school principals and managers. ERO reported some instances of teachers meeting resistance from individual principals, specifically about their role in critiquing data and practice (Education Review Office, 2017a). It is likely that the challenge of building complex networks of relationships was underestimated, and it will take time for clarity about roles to be established.

Robertson (2015) also notes that effective Kāhui Ako leaders care as much for the students in the institution down the road as they do for those in their own—collaboration trumps competition. Similar ideals are espoused by Thrupp and White (2013) who invoked Dewey to make the same point: "what the best and wisest parent wants for his own child, that must the community want for all its children" (Dewey 1902, cited in Thrupp & White, 2013). This is a different ethos from the competitive relationships between schools that arose following the implementation of the Tomorrow's Schools reforms and adds new layers of nuances to the relationships discussed in this section.

In summary

Effective relationships are important in every aspect of assessment and reporting practice. For example, they are a critical component of:

- effective use of AfL pedagogies in the classroom
- mentoring of early career teachers in assessment pedagogies and practices (which are known to be in the very early stages of development when they graduate from their ITE programmes)

13 Assessment of 21st Century Skills (http://www.atc21s.org/).

- peer learning for teachers and school leaders as they seek to shift established teaching and assessment practices
- collaboration across other types of communities of learning (e.g., Kāhui Ako)
- working with families to support achievement of every student
- working and communicating in digital environments.

The need for ongoing support for leaders of professional learning and for peer mentors is a clear implication from all these studies. Significant shifts in assessment practice will not happen without sustained, skilful support for ongoing professional learning, at every career stage.

9. Themes related to system-level accountabilities

Principle: An assessment capable system is an accountable system

This section addresses the sixth principle in the set. It is one of two principles that refer to the idea of assessment capability. An earlier section of the report has established that assessment capability is an important concept for focusing on what needs to happen to achieve system-wide shifts in classroom assessment practices, particularly in relation to AfL. What does this principle add that is different?

We do think there is something new to be added—specifically the concept of system-level capabilities and accountabilities. However, the *relationship* between these ideas is not especially clear in the wording of the principle itself. The intention of this principle could be read as ensuring that system accountabilities are designed to enable a valid and fair account of learning to be given, for every student. This is the reading we will take in this section, although at the moment the principle is written the other way around. It seems to assume that, once the system is assessment capable, accountability will no longer be an issue. This is problematic.

One clear message in previous sections of this report is that building assessment capability is easier said than done, and needs some fresh approaches if it is to be achieved. Many challenges that impede assessment capability have been outlined. Fresh approaches might include: sustained provision of the support to build teachers' assessment capabilities; attention to the design and processes of high-stakes assessments so that they do not send mixed signals about what is important; resources and support for shifting curriculum thinking, with its associated assessment challenges; and the strategic introduction of digital assessment opportunities that support the other goals just named. A similar mix has been announced recently for the review of Ontario's assessment and reporting systems (Campbell et al., 2018).

On this basis, the principle might be reworded along the following lines: *The system supports the ongoing development of assessment capability fit for 21st century learning*. The following discussion works through some of the system-level challenges that might be anticipated from the fresh approaches just suggested. Several new opportunities afforded by rapid developments in digital technologies provide the context for working through the systems-level dynamics.

Microcredentials and "ecologies" of assessment systems

Microcredentials allow achievements of a finer grain-size to be acknowledged and rewarded. A contained aspect of learning can be assessed, with a specific token or badge awarded if the specified learning is successfully demonstrated. Because performance assessment methods can be used, microcredentials

have the potential to capture aspects of learning that are problematic to assess by traditional high-stakes assessment methods. Performance assessments allow assessment of competencies/capabilities, specific context-dependent professional skills, and metacognition (Milligan, Kennedy, & Israel, 2018).

Milligan and her colleagues studied the microcredentials offered in various different initiatives. They looked at the structure of microcredential systems and the processes used to make them manageable and credible. They say that it is important to design a structure where the individual microcredentials are "stackable". This means that each individual award contributes to a bigger whole, with the pieces often built sequentially. In the school sector, an example of this can be seen in innovative systems of microcredentials awarded for demonstrations of professional capabilities. Stackable systems specify the pieces that can be fitted together to demonstrate an overall professional standard. (The structure of the professional standards for New Zealand teachers would readily lend itself to this design.) A major US study of microcredentials for teacher professional standards found considerable variation across different American states—some are designed to be stackable but others are not (Kuriacose & Warn, 2018). On the plus side, this research noted that teachers really liked being able to work towards larger qualifications in manageable pieces, given their heavy workloads.

Clearly, structure needs careful attention if the whole is intended to amount to something more than simply the sum of the parts. Milligan et al. say that stacking is an innovative response to the critique that microcredentials lead to fragmentation of learning programmes, and that the overall design of a qualification can be quite complex when stacking is used (Milligan et al., 2018). Could we apply the idea of stacking to the existing structure NCEA, in effect treating each standard achieved as one microcredential? What advantages might be gained? What else would need to happen to make the idea credible and practical?

The questions just posed draw attention to another aspect of microcredential systems that is also critically important. The *credibility* of a microcredential system rests on a complex web of relationships. In effect, a whole "ecology" of people, processes, and documentation has to be carefully built. Consider the web of relationships that needs to be built for the following components to align well and work together:

- Task design: All the design challenges that apply in other assessment contexts are also relevant here. The learning to be credentialled is likely to require a performance of some sort and that performance will be contextually bound. Credibility begins with the clarity of the task specifications and expectations.
- Evaluation of the evidence presented: The literature is clear that judgements should be made by evaluators with deep expertise in the area being assessed. Again, all the challenges that apply in other assessment contexts also apply here, including focus and clarity in the schedules used to judge the evidence.
- Moderating judgements: This is often carried out by different people than those who make the initial judgements. Clarity of expectations now extends to clear understanding between judges and moderators, again with all the challenges inherent in traditional moderation systems.
- Awarding the credential: Who makes the awards, or endorses them, is critical to credibility. The support of trusted agencies, particularly for professional microcredentials, helps to ensure that they have value for the awardee, beyond the simple satisfaction of gaining them.
- End users: On the plus side, some commentators note that employers are increasingly less impressed by formal qualifications and more interested in what a given individual can do. Some badging platforms integrate directly with LinkedIn, which allows employers to explore the skills and abilities of prospective employees (Murgatroyd, 2018a). Of concern, however, is the rapid proliferation of different microcredentials offered by diverse groups. How do potential end users know which can be trusted to have credibility? Milligan et al. describe considerable efforts made by some providers to bolster credibility (for example, working with professional agencies if the credentials apply to professional learning) (Milligan et al., 2018).

If system-level collaboration within the overall ecology of the microcredentialling system is key to credibility and quality, then all these steps must be robust and well-aligned. This is an area in need of urgent consideration, given the rapid proliferation of microcredentials already offered by different providers. Tertiary education is ahead of the schooling sector in innovating the reporting and awarding of microcredentials (Milligan et al., 2018). Murgatroyd (2018a) says that use of digital badges already happens in one in five US colleges and universities. Just as we were completing this report, NZQA announced a new system for awarding microcredentials in the tertiary sector.[14] We note that the overview material makes no mention of a need to design for stacking. The schooling sector now has an opportunity to review this innovation and plan for well-aligned use of microcredentials in schools.

In many ways, all the issues that apply to the design and use of fit-for-purpose data management systems (see Section 8) could also be seen as applying to microcredentialling systems. The two are interconnected because data have to be stored somewhere—which also brings innovations such as blockchain into the picture, with its potential for personalising assessment and credentials (Vander Ark, 2017). Do these fluid possibilities and complexities point to the need for input from students, parents, employers, etc. into planning and design of microcredential systems for the schooling sector (as has been suggested for database design)?

Ideally, collaboration and coherence will create a positive experience for the learner who sits at the heart of the process. Some futures thinkers say that a positive learner experience should be one of the indicators of quality in national assessment systems that are rapidly evolving to meet the demands of a digital era (Murgatroyd, 2018c).

AI at the interface between biological systems and learning

Artificial intelligence (AI) can be defined in a number of ways. The definition preferred by one group of researchers is simple and clear: "Artificial intelligence (AI) describes the use of computers to do the kinds of things that minds can do" (Gulson et al., 2018, p. 4). Gulson et al. describe a dynamic, evolving context in which the ways that education systems are currently shaped will influence, but also be influenced by, innovations in AI technologies and the corporate interests behind these. The "major shift" they identify is toward personalised learning and assessment, as outlined in earlier sections of the report. They are clear that teachers will still be needed, but their roles will change as more routine aspects of teachers' work are taken over by AI (e.g., many forms of marking and administration).

Milligan has also recently written about the challenges AI poses for education. Her focus is on how we think about learning per se: our assumptions about its nature and about learners; how we understand "measurement" of learning; the criteria we apply when assessing quality of data; and the standards of proof we seek when establishing validity and reliability, and so on. All these are open questions in traditional scholarship, and must also be taken seriously by AI scholars (Milligan, 2018). However, it is not only assessment scholars who need to be clear about how these challenges apply in AI contexts. Gulson et al. note that it is important to educate young people to understand how AI processes work, to inform their personal decision making, and to build insights about ways AI processes and outcomes shape realities (Gulson et al., 2018).

One specific example of the way in which realities might be shaped is the use of AI to identify and track "at-risk" students. Data sources integrated in this process will extend beyond traditional measures of achievement (which will in any case change with use of adaptive assessment) to include affective measures of learning and interaction (as a way to track engagement and collaboration). New data sources might also include movement around a campus (the example given is related to tertiary students) (Gulson et al., 2018).

14 https://www.nzqa.govt.nz/about-us/news/micro-credentials-system-launched/

They also say that the combination of robotics and AI, when built into digital learning and assessment resources, opens up new opportunities and challenges for students with a range of specific disabilities (Gulson et al., 2018). At the edge of the range of possibilities, some researchers claim that neuro-cognitive therapies can treat learning disorders. For example, they say that differences in the way nerves are sheathed (myelination) in the brain's corpus callosum are associated with reading ability. They cite evidence that, given the brain's neuroplasticity, specific therapies can make a difference for children with this biological problem (Boyd, 2016). We have been told that there is critique of this work, partly on the grounds that it has been conducted by those with a vested interest in ensuring the success of the school that offers the therapies. However, we have not been able to source this critique in the time available. Here, too, there are concerns that differential access to such therapies will exacerbate existing social and economic inequalities (Gulson et al., 2018). They say this is an example of an area that will need to be regulated in new ways.

As in the case of microcredentials, AI opens up many potentially highly disruptive possibilities, but also raises important questions about regulation and alignment within the overall system. Issues to be considered include: equity of access, with associated possibilities for exacerbating existing inequalities; actual benefits vs. those claimed; costs of AI services to individuals and to the system; planning for changing patterns of work; educating young people to understand AI and its impacts and so on. The sheer range and reach of these issues suggests that decision making should not be left to chance. On the basis of their literature review for the Gonski Institute, Gulson et al. (2018) make the following six recommendations:

- Establish a cross-sector representative body to provide advice on the development of AI in Australian education systems.
- Establish guidelines for the introduction of adaptive and personalized learning to ensure a focus on educational and equity principles.
- Consult with the teaching profession, teacher educators, and industry to formulate professional development strategies that help educators manage the introduction of AI into their workplaces.
- Develop a set of procurement guidelines that encourage the ethical design and transparency of AI systems purchased by Australian education systems.
- Review data protection legislation internationally to help develop an approach for Australian education systems.
- Increase access to resources that help educators and learners to develop AI-complementary skills, including 21st century skills. (Gulson et al., 2018, pp.2-3)

All of these recommendations resonate with issues raised throughout this report.

Who should do the regulating?

The Gordon Commission on the Future of Assessment noted that the cost-effectiveness of traditional assessment systems now stands in the way of the call for authentic, performance-based assessments such as those discussed in this report. In the US context, they recommended the creation of a permanent Council on Educational Assessments with evaluative, research, and communication functions.

They noted that this group would need evaluation functions to assess [commercial] assessment systems and resources for their impacts on teaching and learning and to help [states] make good purchasing decisions.

Research functions for the proposed Council include:

- determining how best to broaden the targets of assessment so that a "wider range of human abilities" is assessed

- setting performance-level targets [by which they might mean progressions?]
- determining how best to secure data privacy
- scoping challenges of equitable assessment, including ways that agency, dispositions and cultural identities influence the nature and quality of assessment performances
- exploring ways to differentiate assessments to "capture intellective competence as a property of individuals and as a function of collaboration between persons" (The Gordon Commission on the Future of Assessment in Education, 2013, p. 24).

It is clear from this list that problems inherent in AI, as outlined above, also apply more widely across the system. The comprehensive approach outlined here provides a useful guide to what we would need in New Zealand for the system changes needed if we are to realise the assessment principles discussed in this report.

In summary

Complex new accountability issues arise at the intersection of digital tools, AI, new methods of credentialling learning, and rapid changes in the worlds of work and life. The principle that addresses system-level accountabilities probably needs to be updated or replaced to reflect the dramatic changes that have taken place since 2011.

New "ecologies" of learning and assessment have been opened up by digital innovations such as microcredentialling. Purposeful design, alignment between different key components, collaboration between those involved in the different processes, transparency of decision making, attention to database design, and new possibilities for personalising the storage of data (e.g., blockchain) all point to the need for oversight of the ecology of the system as a whole.

We cannot take the act of "learning" as a given. Yet more new possibilities arise at the intersection of neuro-biology, education, and new digital technologies. These innovations tend to be driven by vested commercial interests and caution needs to be exercised when making investment decisions. Even within the traditional school curriculum, new types of outcomes are now seen as desirable, and hence new forms of assessment are needed. Again, many innovative solutions will come at a cost. Oversight of quality and cost/benefits is advisable.

While they are worded differently, recent research-based recommendations for system-level oversight of education in Australia and in the US both point towards greater central oversight and more deliberate planning to ensure the education system can be responsive to the rapid changes taking place in life and work in the 21st century.

References

Absolum, M., Flockton, L., Hattie, J., Hipkins, R., & Reid, I. (2009). *Directions for Assessment in New Zealand (DANZ)*: *Developing students' assessment capabilities.* Wellington: Ministry of Education.

Aitken, G., Sinnema, C., & Meyer, F. (2013). Initial *teacher education: Standards for graduating teachers. A paper for discussion*. Wellington: Ministry of Education. Retrieved from https://www.educationcounts.govt.nz/publications/ECE/2511/initial-teacher-education-outcomes

Aitken, J., Villers, H., & Gaffney, J. (2018). Guided reading: Being mindful of the reading processing of new entrants in Aotearoa New Zealand primary schools. *set: Research Information for Teachers*, (1), 25–33. https://doi.org/10.18296/set.0099

Ayca, C., & Hasan, K. (2017). An application of fuzzy analytic hierarchy process (FAHP) for evaluating students' project. *Educational Research and Reviews*, *12*(3), 120–132. https://doi.org/10.5897/ERR2016.3065

Baird, J.-A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy & Practice*, *24*(3), 317–350. https://doi.org/10.1080/0969594X.2017.1319337

Baker, E. (2018). Design for assessment change. *European Journal of Education*, *53*(2), 138–140. https://doi.org/10.1111/ejed.12275

Bendikson, L. (2015, 17 November). *Community of Schools' (CoSs) leadership—Throwing money and hoping for success.* (Leadership for Communities of Learning—Five Think Pieces. Discussion papers.) Wellington: New Zealand Education Council. Retrieved from https://educationcouncil.org.nz/sites/default/files/Education%20Council%20Five%20Think%20Pieces%200612.pdf

Bereiter, C., & Scardamalia, M. (n.d.). *What will it mean to be an educated person in the mid-21st century?* Gordon Commission for the Future of Assessment. Retrieved from https://www.ets.org/Media/Research/pdf/bereiter_scardamalia_what_will_mean_educated_person_century.pdf

Birenbaum, M., DeLuca, C., Earl, L., Heritage, M., Klenowski, V., Looney, A., ... Wyatt-Smith, C. (2015). International trends in the implementation of assessment for learning: Implications for policy and practice. *Policy Futures in Education*, *13*(1), 117–140. https://doi.org/10.1177/1478210314566733

Bisson, M.-J., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education*, *2*(2), 141–164. https://doi.org/10.1007/s40753-016-0024-3

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Bolden, B., & DeLuca, C. (2016). Measuring the magical: Leveraging assessment for emergent learning. *Assessment Matters*, *10*, 52–73. https://doi.org/10.18296/am.0017

Bonne, L. (2016). *National standards in their seventh year*. Wellington: New Zealand Council for Educational Research. Retrieved from http://www.nzcer.org.nz/system/files/NZCER%20National%20Standards%20Report.pdf

Bonne, L., & Stevens, E. (2017). *Parent and whānau perspectives on their child's schooling: Findings from the NZCER national survey of primary and intermediate schools 2016.* Wellington: New Zealand Council for Educational Research. Retrieved from http://www.nzcer.org.nz/research/publications/parent-and-wh-nau-perspectives-their-child-s-schooling-findings-nzcer-national

Bonne, L., & Wylie, C. (2017). *Teachers' work and professional learning: Findings from the NZCER national survey of primary and intermediate schools 2016*. Wellington: New Zealand Council for Educational Research. Retrieved from http://www.nzcer.org.nz/system/files/National%20Survey_Teacher%20Work_Nov17.pdf

Booth, B., Dixon, H., & Hill, M. (2016). Assessment capability for New Zealand teachers and students: Challenging but possible. *set: Research Information for Teachers*, (2), 28–35. https://doi.org/10.18296/set.0043

Bourke, R., & Mentis, M. (2014). An assessment framework for inclusive education: Integrating assessment approaches. *Assessment in Education: Principles, Policy & Practice*, *21*(4), 384–397. https://doi.org/10.1080/0969594X.2014.888332

Boyd, L. (2016). *Arrowsmith brain imaging study: End of year update and future plans!* Vancouver: University of British Columbia (UBC), Faculty of Medicine. Retrieved from http://www.eatonarrowsmith.com/wp-content/uploads/ubcresearchupdateapril2016-1.pdf

Breakspear, S. (2013). *New metrics briefing 2: Feedback to the GELP Metrics co-design group.* Retrieved from: https://www.gelponline.org/sites/default/files/new_metrics_briefing_2_-_simon_breakspear.pdf

Briggs, S. (2018). *Blockchain technology: Can it change education?* Retrieved from https://www.opencolleges.edu.au/informed/edtech-integration/blockchain-technology-education/

Buckingham, J., & Joseph, B. (2018). *What the Gonski 2 review got wrong*. Sydney, NSW: The Centre for Independent Studies.

Buckingham Shum, S. (2018, June). Simon.BuckinghamShum.net. Retrieved from http://simon.buckinghamshum.net/2018/06/icls2018-keynote/

Burgon, J. (2018). *Special assessment conditions—barriers to use*. Wellington: New Zealand Council for Educational Research.

Burgon, J., Eyre, J., & Stevens, E. (2017). *Describing progress for children and young people learning long term within level 1 of the New Zealand curriculum: Communication.* Report prepared for the Ministry of Education. Wellington: New Zealand Council for Educational Research.

Burgon, J., Roberts, J., Darr, C., & Hipkins, R. (2017). *Reporting progress and achievement for students learning long-term within level 1 of the New Zealand curriculum: An exploratory study.* Report to the Ministry of Education. Wellington: New Zealand Council for Educational Research and Ministry of Education.

Caldwell, A., & Hawe, E. (2016). How teachers of years 4–8 students analyse, interpret and use information from the Progressive Achievement Test: Mathematics. *Assessment Matters*, *10*, 100–125. https://doi.org/10.18296/am.0019

Call, K. (2018). Professional teaching standards: A comparative analysis of their history, implementation and efficacy. *Australian Journal of Teacher Education*, *43*(3).

Cameron, M., & Baker, R. (2004). *Research on initial teacher education in New Zealand: 1993–2004 Literature review and annotated bibliography*. Wellington: New Zealand Council for Educational Research. Retrieved from https://uatmain.educationcouncil.org.nz/sites/default/files/itelitreview.pdf

Campbell, C., Clinton, J., Fullan, M., Hargreaves, A., James, C., & Longboat, K. D. (2018). *Ontario: A learning province: Findings and recommendations from the Independent Review of Assessment and Reporting*. Ottawa, Canada. Retrieved from http://www.edu.gov.on.ca/CurriculumRefresh/learning-province-en.pdf

Care, E., & Vista, A. (2017, March). *Education assessment in the 21st century: New skillsets for a new millennium.* Retrieved from https://www.brookings.edu/blog/education-plus-development/2017/03/01/education-assessment-in-the-21st-century-new-skillsets-for-a-new-millennium/

Casey, G. (2013). Building a student-centred learning framework using social software in the middle years classroom: An action research study. *Journal of Information Technology Education: Research*, *12*, 159–198.

Cavino, H. M. (2013). Across the colonial divide: Conversations about evaluation in indigenous contexts. *American Journal of Evaluation*, *34*(3), 339–355. https://doi.org/10.1177/1098214013489338

Christodoulou, D. (2018, April). *Assessing standards with comparative judgement.* Retrieved from https://blog.nomoremarking.com/assessing-standards-with-comparative-judgement-da20d64c5de1

Cowie, B., & Cooper, B. (2017). Exploring the challenge of developing student teacher data literacy. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 147–163. https://doi.org/10.1080/0969594X.2016.1225668

Cowie, B., & Moreland, J. (2015). Leveraging disciplinary practices to support students' active participation in formative assessment. *Assessment in Education: Principles, Policy & Practice*, *22*(2), 247–264.

Crisp, G. T. (2014). Assessment in next generation learning spaces. In K. Fraser (Ed.), *International perspectives on higher education research* (Vol. 12, pp. 85–100). Emerald Group Publishing. https://doi.org/10.1108/S1479-362820140000012009

Darling-Hammond, L., & Falk, B. (2013). *Teacher learning through assessment: How student-performance assessments can support teacher learning*. Washington DC.: Center for American Progress. Retrieved from https://cdn.americanprogress.org/wp-content/uploads/2013/09/TeacherLearning.pdf

Darr, C. (2018). A return to assessment for learning: Back to the future. *Set: Research Information for Teachers*, (1), 46–48. https://doi.org/10.18296/set.0102

Davies, A., Busick, K., Herbst, S., & Sherman, A. (2014). System leaders using assessment for learning as both the change and the change process: Developing theory from practice. *The Curriculum Journal*, *25*(4), 567–592. https://doi.org/10.1080/09585176.2014.964276

DiCerbo, K., Shute, V., & Kim, Y. J. (2017). The future of assessment in technology-rich environments: Psychometric considerations. In M. Spector, B. Lockee, & M. Childress (Eds.), *Learning, design, and technology*. (pp.1–21). Switzerland: Springer International Publishing.

Drummond, R. (2017). *Extending into the future: How extended project work can help prepare students for success at school, at university and in the careers of tomorrow*. Oxford, UK: Oxford International AQA Examinations. Retrieved from https://cf.oxfordaqaexams.org.uk/oaqaresources/projects/extending-into-the-future1.pdf

Edmunds, B., & Hartnett, M. (2014). Using a learning management system to personalise learning for primary school students. *Journal of Open, Flexible and Distance Learning*, *18*(1), 11–29.

Education Council. (2017). *Our code, our standards: Code of professional responsibility and standards for the teaching profession; Ngā Tikanga Matatika, Ngā Paerewa: Ngā Tikanga Matatika mō te Haepapa Ngaiotanga me ngā Paerewa mō te Umanga Whakaakoranga*. Wellington: Author.

Education Review Office. (2007). *The collection and use of assessment information.* Wellington: Author.

Education Review Office. (2015). *Inclusive practices for students with special education needs in schools*. Wellington: Author. Retrieved from http://www.ero.govt.nz/publications/inclusive-practices-for-students-with-special-education-needs-in-schools/

Education Review Office. (2017a). *Communities of learning 1. Kahui Ako in action. What we know so far.* Wellington: Author. Retrieved from http://www.ero.govt.nz/assets/Uploads/Communities-of-Learning-Kahui-Ako-Action.pdf

Education Review Office. (2017b). *Newly graduated teachers—preparation and confidence to teach.* Wellington: Author. Retrieved from http://www.ero.govt.nz/publications/newly-graduated-teachers-preparation-and-confidence-to-teach/

Education Review Office. (2018a). *Building genuine learning partnerships with parents*. Wellington: Author. Retrieved from http://www.ero.govt.nz/publications/building-genuine-learning-partnerships-with-parents/

Education Review Office. (2018b). *Evaluation at a glance. A decade of assessment in New Zealand primary schools—practice and trends*. Wellington: Author. Retrieved from http://www.ero.govt.nz/publications/evaulation-at-a-glance-a-decade-of-assessment-in-new-zealand-primary-schools-practice-and-trends/

Educational Assessment Research Unit. (n.d.). *Wānangatia te Putanga Tauira: National monitoring study of student achievement*. Dunedin, Wellington: University of Otago, New Zealand Council for Educational Research. Retrieved from http://nmssa.otago.ac.nz/files/EG_Issue_13.pdf

Edwards, F. (2017). *The development of summative assessment literacy. An exploration of the experiences of beginner secondary science teachers.* A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Education, University of Waikato. Retrieved from https://waikato.researchgateway.ac.nz/bitstream/handle/10289/11567/thesis.pdf?sequence=4&isAllowed=y

Edwards, S. (2017). Assessment of English language learners in New Zealand primary schools: Purposes, principles, and practices. *Assessment Matters*, *11*, 53–74 https://doi.org/10.18296/am.0024

Flockton, L. (2012). Commentary: Directions for assessment in New Zealand. *Assessment Matters*, *4,* 129–149.

Gadd, M. (2014). *What is critical in the effective teaching of writing? A study of the classroom practice of some year 5–8 teachers in the New Zealand context.* Unpublished doctoral dissertation, The University of Auckland, Auckland.

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. Retrieved from https://docs.education.gov.au/node/50516

Groff, J. (2018). The potentials of game-based environments for integrated, immersive learning data. *European Journal of Education*, *53*(2), 188–201. https://doi.org/DOIL 10.1111/ejed.12270

Grudnoff, L., Haigh, M., Hill, M., Cochran-Smith, M., Ell, F., & Ludlow, L. (2017). Teaching for equity: Insights from international evidence with implications for a teacher education curriculum. *The Curriculum Journal*, *28*(3), 305–326. https://doi.org/10.1080/09585176.2017.1292934

Guerin, A. (2015). *'The inside view' Investigating the use of narrative assessment to support student identity, wellbeing, and participation in learning in a New Zealand secondary school.* Unpublished doctoral dissertation, University of Canterbury, Christchurch. Retrieved from https://ir.canterbury.ac.nz/handle/10092/10486

Gulson, K., Murphie, A., Sellar, S., & Taylor, S. (2018). *Education, work and Australian society in an AI world: A review of research literature*. Sydney: University of New South Wales, Gonski Institute.

Harper, A., & Brown, G. (2017). Students' use of online feedback in a first-year tertiary biology course. *Assessment Matters*, *11*, 99–121. https://doi.org/10.18296/am.0026

Harris, L., Brown, G., & Harnett, J. (2014). Understanding classroom feedback practices: A study of New Zealand student experiences, perceptions, and emotional responses. *Educational Assessment, Evaluation and Accountability*, *26*(2), 107–133.

Hawe, E., & Parr, J. (2014). Assessment for learning in the writing classroom: An incomplete realisation. *Curriculum Journal*, *25*(2), 210–237.

Hayward, L. (2015). Assessment is learning: The preposition vanishes. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 27–43. https://doi.org/10.1080/0969594X.2014.984656

Heard, J., & Hollingsworth, H. (2018). *Continuous student reporting—the next step?* Melbourne, VIC: Australian Council for Educational Research—ACER. Retrieved from https://www.teachermagazine.com.au/articles/continuous-student-reporting-the-next-step

Hill, M. (2011). 'Getting traction': Enablers and barriers to implementing assessment for learning in secondary schools. *Assessment in Education: Principles, Policy & Practice*, *18*(4), 347–364. https://doi.org/10.1080/0969594X.2011.600247

Hill, M., Ell, F., Grudnoff, L., Haigh, M., Cochran-Smith, M., Chang, W.-C., & Ludlow, L. (2017). Assessment for equity: Learning how to use evidence to scaffold learning and improve teaching. *Assessment in Education: Principles, Policy & Practice*, *24*(2), 185–204. https://doi.org/10.1080/0969594X.2016.1253541

Hill, M., & Eyers, G. (2016). Moving from student to teacher: Changing perspectives about assessment through teacher education. In *The handbook of human and social conditions in assessment* (pp. 57–76). New York: Routledge.

Hill, M. F., Ell, F. R., & Eyers, G. (2017). Assessment capability and student self-regulation: The challenge of preparing teachers. *Frontiers in Education*, *2*. https://doi.org/10.3389/feduc.2017.00021

Hipkins, R. (2013). *NCEA one decade on*. Wellington: New Zealand Council for Educational Research. Retrieved from http://www.nzcer.org.nz/research/publications/ncea-one-decade

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

Hipkins, R. (2017). *Weaving a coherent curriculum: How the idea of capabilities can help*. Wellington: New Zealand Council for Educational Research. Retrieved from http://www.nzcer.org.nz/research/publications/weaving-coherent-curriculum-how-idea-capabilities-can-help

Hipkins, R., & Dingle, R. (2013). Student management systems in secondary schools. *set: Research Information for Teachers*, (2), 29–34.

Hipkins, R., Johnston, M., & Sheehan, M. (2016). *NCEA in context*. Wellington: NZCER Press.

Hipkins, R., & Robertson, S. (2012). The complexities of moderating student writing in a community of practice. *Assessment Matters*, *4*, 30–52.

Horwitz, P. (2018). *What happens when students try to work collaboratively?* Retrieved from https://concord.org/newsletter/2018-spring/what-happens-when-students-try-to-work-collaboratively/?utm_source=The+Concord+Consortium+List&utm_campaign=dc1a7e8f9b-Spring_2018_%40Concord_Announcement&utm_medium=email&utm_term=0_4ca9f8d47e-dc1a7e8f9b-313227193

Houghton, C. (2015). *Underachievement of Māori and Pasifika learners and culturally responsive assessment.* Retrieved from https://ir.canterbury.ac.nz/handle/10092/11437

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

Jones, I., & Henderson, B. (2016, November). *Snapshots of deep learning over time: A novel approach to measuring student progress*. Presentation to AEA conference, Cyprus.

Jonsson, A., Lundahl, C., & Holmgren, A. (2015). Evaluating a large-scale implementation of assessment for learning in Sweden. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 104–121. https://doi.org/10.1080/0969594X.2014.970612

Kucirkova, N. (2018). Is Silicon Valley standardizing learning? *Education Week*, pp. 22–23.

Kuriacose, C., & Warn, A. (2018). *An exploration of six educator micro-credential programs | CCE—Center for Collaborative Education.* Retrieved 6 June 2018, from http://cce.org/paper/personalized-professional-learning-micro-credential-programs

Masters, G. (2017, February 7). *Rethinking how we assess learning in schools.* Retrieved 5 June 2018, from http://theconversation.com/rethinking-how-we-assess-learning-in-schools-71219

McDowall, S., & Hipkins, R. (2018). *How the key competencies evolved over time: Insights from the research*. Wellington: New Zealand Council for Educational Research. Retrieved from http://www.nzcer.org.nz/system/files/Paper%202%20KCs%20research%20_final.pdf

McIlroy, A.-M. (2017). *'The myth of inability': Exploring children's capability and belonging at primary school through narrative assessment.* Unpublished doctoral dissertation, University of Canterbury, Christchurch. Retrieved from https://ir.canterbury.ac.nz/handle/10092/14939

McPhail, G. (2018). Curriculum integration in the senior secondary school: A case study in a national assessment context. *Journal of Curriculum Studies*, *50*(1), 56–76. https://doi.org/10.1080/00220272.2017.1386234

May, S. (n.d.). *Assessment: What are the cultural issues in relation to Pasifika, Asian, ESOL, immigrant and refugee learners?* Report prepared for the Ministry of Education, for DANZ working group. Retrieved from https://assessment.tki.org.nz/Research-and-readings/Research-behind-DANZ

Meissel, K., Meyer, F., Yao, E. S., & Rubie-Davies, C. M. (2017). Subjectivity of teacher judgments: Exploring student characteristics that influence teacher judgments of student ability. *Teaching and Teacher Education*, *65*, 48–60. https://doi.org/10.1016/j.tate.2017.02.021

Milligan, S. K. (2018). *Methodological foundations for the measurement of learning in learning analytics* (pp. 466–470). ACM Press. https://doi.org/10.1145/3170358.3170391

Milligan, S. K., Kennedy, G. E., & Israel, D. (2018). *Assessment, credentialling and recognition in the digital era: Recent developments in a fertile field*. Centre for Strategic Education: Melbourne. Retrieved from http://www.cse.edu.au/content/assessment-credentialling-and-recognition-digital-era-recent-developments-fertile-field

Ministry of Education. (2007). *The New Zealand curriculum*. Wellington: Learning Media.

Ministry of Education. (2011). *Ministry of Education position paper: Assessment: Schooling sector.* Wellington: Author. Retrieved from http://assessment.tki.org.nz/Media/Files/Ministry-of-Education-Position-Paper-Assessment-Schooling-Sector-2011

Ministry of Education. (2014). *Review of special assessment conditions for National Certificate of Educational Achievement (NCEA)*. Wellington: Author.

Ministry of Education. (2017). *Development paper: Revising the technology learning area to strengthen digital technologies in the New Zealand curriculum: Proof of concept, development, and testing*. Wellington: Author. Retrieved from http://technology.tki.org.nz/Technology-in-the-NZC/DDDO-Progress-outcomes-exemplars-and-snapshots

Ministry of Education. (2018). *Achievement and progress in mathematics, reading and writing in primary schooling. Analysis of e-asTTle assessment data 2011–2016*. Wellington: Author. Retrieved from https://www.educationcounts.govt.nz/__data/assets/pdf_file/0019/185023/20171213-Achievement-and-Progress-in-mathematics-reading-and-writing.pdf

Morton, M. (2014). Using DSE to "notice, recognise and respond" to tools of exclusion and opportunities for inclusion in New Zealand. *Review of Disability Studies*.

Mosher, F. (2011). *The role of learning progressions in standards-based education reform*. Consortium for Policy Research in Education. https://doi.org/10.12698/cpre.2011.rb52

Munro, I. (2016). *When special assessment conditions for some become standard assessment conditions for all: The implications for SAC as assessment moves on-line*. Unpublished paper, New Zealand Qualifications Authority, Wellington.

Murgatroyd, S. (2018a). New approaches to the assessment of learning: New possibilities for business education. In A. Khare & D. Hurst (Eds.), *On the line—business education in the digital age* (pp. 141–155). Switzerland: Springer.

Murgatroyd, S. (2018b, in press). Recent developments in assessment. In M. Makhanya (Ed.), *Global best practices in online teaching and learning—around the world with 30 stops*. Pretoria: University of South Africa Press.

Murgatroyd, S. (2018c, June). Teaching and learning in the digital age: A new understanding of quality. Retrieved from https://teachonline.ca/tools-trends/insights-online-learning/2018-02-07/teaching-learning-digital-age-new-understanding-quality

OECD. (2005). *The definition and selection of key competencies: Executive summary*. Paris: Author. Retrieved from http://www.oecd.org/pisa/35070367.pdf

OECD. (2018). *The future of education and skills: Education 2030*. Paris: Author. Retrieved from http://www.oecd.org/education/2030/E2030%20Position%20Paper%20(05.04.2018).pdf

Pane, J., Steiner, E., Baird, M., & Hamilton, L. (2015). *Continued progress: Promising evidence on personalized learning*. RAND Corporation. https://doi.org/10.7249/RR1365

Parr, J. M. (2016). Accelerating student progress in writing: Examining practices effective in New Zealand primary school classrooms. In E. Ortlieb, E. H. Cheek, & W. Verlaan (Eds.), *Literacy Research, Practice and Evaluation* (Vol. 7, pp. 41–64). Emerald Group Publishing. https://doi.org/10.1108/S2048-045820160000007002

Perry, K. (2015). *Moving the secondary economics classroom towards formative assessment using instant feedback techniques.* A dissertation submitted in partial fulfilment of the requirements for the degree of Masters of Professional Studies in Education, The University of Auckland, Auckland.

Poskitt, J., & Mitchell, K. (2012). New Zealand teachers' overall teacher judgements (OTJs): Equivocal or unequivocal? *Assessment Matters*, *4*, 53–75.

Price, D., Smith, J. K., & Berg, D. A. G. (2017). Personalised feedback and annotated exemplars in the writing classroom: An experimental study in situ. *Assessment Matters*, *11*, 122–144. https://doi.org/10.18296/am.0027

Ramsay, J., Vetelino, C., Dewar, S., & Barker, P. (2018). *Teacher learning communities: A way to effect change in formative assessment and teaching practices*. Wellington: Ministry of Education. Retrieved from http://thehub.superu.govt.nz/resources/teacher-led-innovation-fund-tlif-summaries-of-completed-projects/

Rao, K., Ok, M. W., & Bryant, B. R. (2014). A review of research on universal design educational models. *Remedial and Special Education*, *35*(3), 153–166. https://doi.org/10.1177/0741932513518980

Robertson, J. (2015, 17 November). *Think-piece on leadership education in New Zealand.* (Leadership for Communities of Learning—Five Think Pieces. Discussion papers.) Wellington: New Zealand Education Council. Retrieved from https://educationcouncil.org.nz/sites/default/files/Education%20Council%20Five%20Think%20Pieces%200612.pdf

Rose, D., Meyer, A., Strangman, N., & Rappolt, G. (2002). *Teaching every student in the digital age*. Alexandria, VA: ASCD. Retrieved from http://www.ascd.org/publications/books/101042/chapters/Using-UDL-to-Accurately-Assess-Student-Progress.aspx

SACE Board of South Australia. (2018). *Research project subject assessment advice*. Adelaide, SA: Government of South Australia.

Scoular, C. (2018). *Equipping teachers with tools to assess and teach general capabilities*. Presented at the Research Conference 2018, Australian Council for Educational Research, Sydney.

Scoular, C., & Heard, J. (2018). Teaching and assessing general capabilities. *Teacher Magazine*, (June 5). Retrieved from https://www.teachermagazine.com.au/articles/teaching-and-assessing-general-capabilities?utm_source=CM&utm_medium=bulletin&utm_content=June5

Searle, M., Elrofaie, A., Kirkpatrick, L. C., Sauder, A., & Brown, H. M. (2017). Investigating the use of a one-to-one technology programme on formative assessment practices in grades 7 to 9 classroom learning environments. *Assessment Matters*, *11*, 145–170. https://doi.org/10.18296/am.0028

Sellar, S., Rutkowski, D., & Thompson, G. (2017). *The global education race: Taking the measure of PISA and international testing*. Edmonton, Alberta: Brush Education.

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

Siarova, H., Sternadel, D., & Mašidlauskaitė, R. (2017). *Assessment practices for 21st century learning review of evidence: Analytical report.* Luxembourg: European Commission. Retrieved from http://nesetweb.eu/wp-content/uploads/NESET-II_AR1_2017s.pdf

Siddiq, F. (2016). *Learning in digital networks—a novel assessment of students' ICT literacy*. Presented at the Annual Conference of the Association for Educational Assessment – Europe, Cyprus.

Sinnema, C., Alansari, M., & Turner, H. (2018). *The promise of improvement through and of the Teacher-Led Innovation Fund. Evaluation of the Teacher-Led Innovation Fund: Final report*. Auckland: Auckland UniServices Limited. Retrieved from https://www.educationcounts.govt.nz/publications/schooling/evaluation-of-the-teacher-led-innovation-fund-final-report

Smaill, E. (2018). *Social moderation: Assessment for teacher professional learning.* Unpublished Doctor of Philosophy thesis, University of Otago, Dunedin. Retrieved from http://hdl.handle.net/10523/7850

Smith, J., Gilmore, A., Berg, D., Smith, L., & Jameson-Charles, M. (2012). What makes performace tasks motivating? *Assessment Matters*, *4*, 76–94.

Smith, L., Hill, M., Cowie, B., & Gilmore, A. (2014). Preparing teachers to use the enabling power of assessment. In C. Wyatt-Smith, V. Klenowski, & P. Colbert (Eds.), *Designing Assessment for Quality Learning* (Vol. 1, pp. 303–323). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-5902-2_19

Stevens, E., & Hipkins, R. (2016). *Review and Maintenance Programme (RAMP): Learning languages*. Wellington: New Zealand Council for Educational Research. Retrieved from Learning languages—NCEA on TKI https://ncea.tki.org.nz/…/RAMP%20Learning%20Languages%20Literature%20Overview

Stonefields School. (n.d.). *SchoolTalk*. Retrieved 30 May 2018, from http://schooltalk.co.nz

Sturgis, C. (2015, July). Learning progressions: Are student-centered state standards possible? Retrieved from https://www.competencyworks.org/analysis/are-student-centered-state-standards-possible/

Sturgis, C., & Casey, K. (2018). *Designing for equity: Leveraging competency-based education to ensure all students succeed*. CompetencyWorks, iNACOL. Retrieved from https://www.inacol.org/resource/designing-equity-leveraging-competency-based-education-ensure-students-succeed/

Sunnybrae School. (2018). *TLIF2-026: Supporting success on school entry and the first year of instruction. Final project report to the Teacher-Led Innovation Fund (unpublished).*

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. Retrieved from http://www.gordoncommission.org/rsc/pdfs/gordon_commission_technical_report.pdf

Thorpe, V., Gilmour, H., & Walton-Roy, K. (2017). Shared understanding: Using a conceptual model to support the assessment of NCEA group composing. *Assessment Matters*, *11*, 75–98. https://doi.org/10.18296/am.0025

Thrupp, M., & White, M. (2013). *Research, analysis and insight into national standards (RAINS) project final report: National standards and the damage done*. Commissioned report for external body. Wilf Malcolm Institute of Educational Research, The University of Waikato. Retrieved from https://hdl.handle.net/10289/8394

Tolmie, E. (2016). *Implementing personalised learning in New Zealand primary schools' innovative learning environments.* A thesis submitted in partial fulfilment of the requirements for the degree of Master of Educational Management and Leadership, UNITEC Institute of Technology, Auckland. Retrieved from http://unitec.researchbank.ac.nz/bitstream/handle/10652/3655/MEdLM_2016_Emma%20Tolmie_1387132_Final%20Research.pdf?sequence=1

Tondeur, J., van Braak, J., Siddiq, F., & Scherer, R. (2016). Time for a new approach to prepare future teachers for educational technology use: Its meaning and measurement. *Computers & Education*, *94*, 134–150. https://doi.org/10.1016/j.compedu.2015.11.009

Vander Ark, T. (2017). *How blockchain will transform credentialing (and education)*. Retrieved from http://www.gettingsmart.com/2017/12/blockchain-will-transform-credentialing-education/

Voogt, J., & Pareja Roblin, N. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, *44*(3), 299–321.

Ward, J., & Thomas, G. (2016). *National standards: School sample monitoring & evaluation project, 2010–2014*. Wellington: Ministry of Education. Retrieved from https://www.educationcounts.govt.nz/__data/assets/pdf_file/0009/171684/National-Standards-School-Sample-Monitoring-and-Evaluation-2010-2014.pdf

White, Karyn, & Hipkins, R. (2017). An innovative approach to assessing more complex outcomes of learning. *set: Research Information for Teachers*, (3), 15–19. https://doi.org/10.18296/set.0087

White, Katie. (2017, November 21). *Can assessment and open-ended contexts coexist?* Retrieved from http://allthingsassessment.info/2017/11/21/assessment-and-open-ended-contexts/#more-1113

Willis, J., & Klenowski, V. (2018). Classroom assessment practices and teacher learning: An Australian perspective. In J. Heng & M. Hill (Eds.), *Teacher learning with classroom assessment: Perspectives from Asia Pacific*. (pp. 19–37). Singapore: Springer Singapore.

Wilson, M., Gochyyev, P., & Scalise, K. (2016). Assessment of learning in digital interactive social networks: A learning analytics approach. *Online Learning*, *20*(2), 97–119.

Wylie, C. (2012). *Vital connections: Why we need more than self-managing schools*. Wellington: NZCER Press.

Wylie, C., & Bonne, L. (2016). *Secondary schools in 2015: Findings from the NZCER national survey*. Wellington: New Zealand Council for Educational Research. Retrieved from http://www.nzcer.org.nz/system/files/NZCER%20Secondary%20Survey%202015_%20Full%20report_0.pdf

Wylie, C., McDowall, S., Ferral, H., Felgate, R., & Visser, H. (2018). *Teaching practices, school practices, and principal leadership: The first national picture in 2017*. Wellington: New Zealand Council for Educational Research. Retrieved from https://www.tspsurveys.org.nz/images/TSP_National_Report_2017.pdf

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, *58*, 149–162. https://doi.org/10.1016/j.tate.2016.05.010

Appendices

APPENDIX 1
Acknowledgements

The following people were generous in responding to requests for help or in joining with us to think through the challenges a wide-ranging review will inevitably surface. We thank them all.

Denise Arnerich: Ministry of Education

Margaret-Anne Barnett: Ministry of Education

Jacky Burgon: NZCER

Bronwen Cowie: University of Waikato

Kate Curtis: Ministry of Education

Charles Darr: NZCER

Cathy Diggins: Ministry of Education

Carolyn English: CORE Education

Miriam Gibson: Ministry of Education

Kalervo Gulson: University of New South Wales

Mary Hill: University of Auckland

Sally Jackson: Ministry of Education

Ian Jones: Loughborough University, UK

Rebecca Lythe: NZCER

Sheridan McKinley: NZCER

Sarah Martin: Stonefields School

Stephen Murgatroyd: Murgatroyd Communications & Consulting, Alberta, Canada

Mark Osborne: Leading Learning

Claire Scoular: Australian Council for Educational Research

Taylor Webb: The University of British Columbia

Heleen Visser: NZCER

Joanne Walker: University of Auckland

Cathy Wylie: NZCER

APPENDIX 2
Tags used to organise the Zotero data base

Assessment capability principle

Tag name	Scope
Assessment for learning	all papers with an explicit AfL focus (and assessment capability per se)
Assessment/data literacy	commentary about teachers' ability to make effective use of data
Evidence from practice	all the New Zealand studies of actual classroom-based assessment
Exemplars	specifically the need for teacher and student support materials
Formative assessment	includes how teachers use data to meet students' learning needs etc. (unlike AfL, there is no suggestion of intent to involve students in these decisions)
Progression	papers that discuss the use of pathways by which learning unfolds, with the intention to support assessment for learning
Theoretical argument	papers that discuss the relationship between cognitive processes and assessment practices

Curriculum principle

Tag name	Scope
Assessing competencies	"21st century" learning outcomes are the focus (both competencies and dispositions included under this tag)
Assessing emergent outcomes	paying attention to more complex outcomes, and those that have not been anticipated in advance
Assessment design	papers that discuss how assessment tools and tasks are structured to achieve their intended purposes and to address issues of validity, reliability, etc.
Comparative judgement	A technology-enabled method for making judgements about complex learning outcomes
NCEA	papers that discuss aspects of assessment for an NCEA qualification
Technology-enabled pedagogies	this theme addresses use of spaces (MLEs etc.) and ability to assess learning that happens beyond the classroom
Types of knowledge	papers that do not take "knowledge" as a given. Focus might be epistemic, nature of conceptual understanding, practical knowledge, etc.

Student at the centre principle

Tag name	Scope
Equity	papers that discuss impact of assessment opportunities, practices, and impacts for specific groups, including Māori and Pasifika learners
Learning stories	papers that focus on narrative assessment strategies
Personalising assessment	particularly with an emphasis on the affordances of digital technologies for this purpose
Progression	papers that discuss the use of pathways by which learning unfolds, with the intention to support assessment for learning
Self and peer assessment	papers that focus on students as active meaning-makers in assessment activities, and/or where they have the locus of control for assessing their own work
Special needs	papers that discuss assessment for students who have specific learning challenges and needs, including ESOL
Stealth assessment	assessment of learning in online game-like environments where the students' actions and choices provide the data that constitute "evidence" for assessment purposes

Range of evidence principle

Tag name	Scope
Comparative judgement	a method of assessment based on multiple pair-wise comparisons of student work, fed into an algorithm for the overall judgement
Group assessment	any papers that address assessment of work completed by groups rather than individuals
Moderation	teachers' assessment decision making is the focus of professional conversations
Overall teacher judgements	how evidence from different assessment events is brought together to determine overall progress
Performance assessments	Assessment of something students actually do (as opposed to what they write down
Student management systems	Databases for collating student progress and achievement data

Quality interactions principle

Tag name	Scope
Family/whānau engagement	specifically in assessment and/or reporting of students' learning and achievement (includes papers that discuss a role for digital technology in school–home communication)
Initial teacher education	papers that advocate for better/different preparation for teachers' assessment roles (in relationships because some advocate for better research/practice alignment)
Leadership of assessment	Senior and middle leadership roles in supporting assessment practice/ assessment change
Professional learning	of practising teachers, and some commentary on leadership of assessment learning

Accountable system principle

Tag name	**Scope**
Data analytics	issues to do with how large data sets are used to create inferences and inform policy
E-assessment tools	commentary on design and provision of e-enabled tools, with implications for policy
Equity	papers that discuss consequences of assessment and accountability practices for specific groups, including Māori and Pasifika learners
Microcredentials	papers that discuss challenges of establishing and using systems of microcredentials
Neurotherapy	Leveraging links between biology, learning, and assessment
Policy	commentary on role of policy in ensuring an aligned/coherent system, of which assessment is one key part
SMS	papers that focus on how achievement data are recorded for subsequent potential use in data analytics

NZCER
Rangahau Mātauranga o Aotearoa

New Zealand Council for Educational Research
facebook.com/nzcer
@NZCER
www.linkedin.com/NZCER
www.nzcer.org.nz