The Impact of High Stakes Testing: The Australian Story

Abstract

High Stakes testing in Australia was introduced in 2008 by way of the National Assessment Program - Literacy and Numeracy (NAPLAN). Currently, every year all students in Years 3, 5, 7 and 9 are assessed on the same days using national tests in Reading, Writing, Language Conventions (Spelling, Grammar and Punctuation) and Numeracy. In 2010 the NAPLAN results were published on the Federal Government MySchool website. The impact of these high stakes tests on jurisdictions, school principals, parents and students is considered in this article. We draw on reported observations from the Australian Primary Principals Association during 2009-10 testing periods across the country and published Australian research on the impact of high stakes literacy and numeracy testing. We also examine alternative approaches that include the use of assessment evidence for learning improvement purposes and for accountability purposes. In considering alternatives to the current large-scale testing approach we draw on key insights from research on teacher judgement, achievement standards and social moderation in the context of national curriculum and assessment reform in support of the suggested directions forward.

Introduction

In this article the authors aim to reflect past, present and possible future directions of the high stakes national testing regimes in Australia with particular reference to literacy and numeracy. The present situation in Australia shows some signs that the approach to accountability through testing runs the risk of repeating the unintended consequences experienced in other countries including the United States and England. These have been well documented by a range of writers who demonstrate the "uses and abuses of testing" (Stobart, 2008) and how "high stakes testing corrupts schools" (Nichols & Berliner, 2007).

The authors recognise that the Australian government's push for high stakes testing is driven by a desire to meet public accountability, demonstrate transparency and maintain public confidence in the standards of schooling. The issue for educators across the country is whether this push will meet accountability demands without negatively impacting on high quality, high equity teaching and learning.

Though in its relative infancy, Australia now has a burgeoning multi-million dollar testing industry, financial incentives for principals to lift tests scores, and "flying squads" to support schools perceived to be underperforming. The history of how the testing regime in Australia has developed to date and the directions forward are discussed. The tensions between the use of test results for diagnostic and improvement purposes are considered in the context of increased pressure on

schools and teachers to account for teacher and school improvement. In the discussion we draw on recently published data in the form of principles governing the use and reporting of results from the country's National Assessment Program - Literacy and Numeracy[i] (NAPLAN) and the accompanying examples of perverse effects from the 2010 tests, reported by the Australian Primary Principals Association (APPA, 2010a).

This is a timely opportunity to reflect on the direction Australia is taking in terms of high stakes assessment programs such as NAPLAN. The authors support the position taken by APPA that such programs can have unintended and negative impact on teaching and learning quality and that schools must be protected from such consequences. Perverse effects are discussed in this article and the authors conclude with directions forward.

Background

Internationally there are at least two major levers for educational reform. First, large-scale high stakes standardised testing in the pursuit of accountability, the focus of this issue of the journal, and second, the understanding of the central role of the teacher in quality assessment practice, understood to be at the heart of learning and learning improvement. Also influential in reform efforts in several countries is the system push for evidence of achievement tied to a commitment of transparency for accountability.

Systems are hungry for such data, which is not a new phenomenon in several countries. In the United States, for example, several writers including Baker and Stites (1990) and Linn (2003) have identified the intensification of government policy interest in externally mandated testing, and the concurrent increased sanctions that have been attached to test scores. Other writers have further identified the impact of accountability-driven testing on racial and socioeconomic equity when the salience of testing for teachers and students has clearly increased (Lee & Wong, 2004). This theme is strikingly clear in the recently published work of Darling-Hammond (2010) where the cautionary voice about the dangers of testing to undermine, or at least work against, equity initiatives is clear. Similarly in Scotland, Hutchinson and Hayward (2005) reported that, "measurement of attainment levels rather than the quality of assessment practices in classrooms became the main focus of schools' planning and action" (p. 1). Commenting on the growth of large-scale testing, Delandshere (2002) highlighted how the policy priorities of testing have occurred even though there has been "almost unanimous recognition of the limitations of current measurement theory and practice" (p. 1461).

The move to foreground the accountability purpose of testing that occurred in England almost twenty years ago is here in Australia today, with schools and teachers being judged on published results and schools being placed in league tables. This has been driven largely by the media and political decision-making at

both state and federal levels. Over the past two years in particular, accountability testing has assumed increasing prominence in public education policy: 2008 marked the extension of Australia's NAPLAN from Years 3, 5 and 7 to include students in Year 9, and a concurrent move driven by Commonwealth funding legislation, from state-based to a national testing system. Beginning 2006, schools were required to make available their individual school performance test data (literacy and numeracy) in at least two of the following forms: hard copy to parents; posting on school websites, or posting on a public visible billboard.

In 2010, the high stakes nature of NAPLAN testing was confirmed with the publication of results on the MySchool website[ii]. The federal government reported high levels of parental support for this initiative indicating that it believed that it serves the best interests of transparency and accountability. Specifically it was claimed that this site enabled parents to have access to evidence of schools' performance in the national literacy and numeracy testing. In effect, the publication of these results was a claim to support parental access to information about the quality of schooling and the results themselves became codes or indexes for the quality status of individual schools and education systems more generally. The suggestion that parents would have the necessary data for choice of school for their children was made, yet silence remains around the issue of equity. It is noteworthy that as the new Australian Prime Minister, Julia Gillard launched the national election in August 2010 with a speech that highlighted initiatives such as national curriculum and the MySchool website as evidencing her track record in accountability and the provision of quality education for all Australian children.

Looking back to the country's recent past position however, and in consideration of equity and testing in particular, it is useful to recall Australia's first National Plan, titled *Literacy for All: The Challenge for Australian Schools* (Department of Employment, Education, Training and Youth Affairs[iii], 1998). This policy document promised a commitment to testing for diagnosis and learning improvement. It successfully aligned the discourse of testing with the discourse of equity, the Plan presenting the notion that testing was in the best interests of all students, including those at risk of not making satisfactory progress. In the late nineties, the stated position was for the data to be used for student identification, targeted intervention and tracking learning growth over time. Professional development was also an identified need in this first national plan. More than a decade later, however, the nexus of testing to learning improvement, especially for students at educational risk, has not been realised. Dawkins (2007) reported that:

> Domestic evidence shows that Australia has not been making any progress on this [improving the balance between equity and quality] front. Data from the 1975 survey of literacy and numeracy levels of Australian students, and subsequent Longitudinal Surveys of Australian Youth (LSAY), show that difference in social background had as much impact on differences in educational achievement in 1998 as they did in 1975. This should be of

> concern to all Australian governments as well as to the Catholic and independent school sectors. (p. 11)

In light of Dawkin's comparison mentioned in this extract, and more recently published Programme for International Student Assessment (PISA) results, it is fair to say that Australia has not achieved high quality and high equity in the national testing initiatives.

The Emergence of Australia's Testing Industry

The emergence of Australia's testing industry can be understood against a background of three main historical phases in assessment: industrialisation and universal schooling at the turn of the 20th century; the rise of the middle class and capitalism in the middle of the century; and the emergence of calls from the field to centre on views of education and purposes for schooling and assessment (Earl, 2005). In the first phase, industrialisation and universal schooling at the turn of the last century led to schools becoming significant social institutions and places where evaluation of student achievement was used to serve economic imperatives. The main purpose of assessment was to sort students (in and out of schools), with assessment serving economic and therefore, political purposes.

The rise of the middle class and capitalism reshaped the role of schools and schooling. The purposes of assessment to measure, sort and segregate were consolidated. It was at this time that assessment was aligned with what Earl (2005) and others identify as seemingly scientific and objective mechanisms for measuring student achievement. By extension, such mechanisms are often associated with tests, and more specifically multiple-choice tests. The pervasive influence of this development, even to the present, is that test scores decide what each person's role in society should be (Earl, 2005). Several writers have made the point previously that the perceived merit (and by implication, fiscal viability) of large scale testing initiatives came to be invested in the claim, usually uncontested, that tests, and more specifically, multiple choice tests, could deliver objective measurements in which society could have confidence. In short, tests as developed outside schools and classrooms – uncontaminated in their development or scoring by teachers – came to be construed as scientifically developed instruments, capable of yielding objective measures of a student's real achievement. The traditional divide between objective and subjective judgement became established, the former routinely associated with standardised testing, and the latter, teacher judgement. Underpinning the divide was the ill-conceived notion that standardised testing led to more reliable judgement, especially where marking was regulated (e.g., by machine marking), and relied less on the human brain for decision-making.

The risks associated with the assumption about the intrinsic use of testing best placed at a distance from the work of teachers are heightened when there is a clear need for politicians to be seen to deliver improved outcomes in education. At issue then is the relationship between the political appeal of particular education policy

directions and the more fundamental matter of fitness for purpose. Broadfoot and Black (2004) note, for example, that "decisions about assessment procedures – particularly those concerning high stakes testing of various kinds – are as often based on perceived *political* appeal as they are on a systematic knowledge on the scientific evidence concerning fitness for purpose" (p. 3).

In Australia, the Ministerial Council for Education, Employment, Training and Youth Affairs (MCEETYA) in April 2009 announced the decision to develop a system for comparing the performance of schools using NAPLAN results and other sources of data. This was considered by state and territory Education Ministers to be a step towards greater transparency. State governments are now keen to raise standards as represented by the results of NAPLAN tests. For example, in Queensland[iv] in 2009, the Premier advised schools to sit practice NAPLAN tests in Years 3, 5, 7 and 9 as she was dissatisfied by the overall results of the previous year's tests which she stated were designed to assess if students were meeting "national standards in numeracy, reading, writing, spelling, punctuation and grammar" (Bligh, 2009). At the national level, there are no officially endorsed statements about the expected learning of literacy and numeracy as cross-curriculum priorities and no official descriptors of standards to inform teachers about the expectations of quality, except for those produced after the testing is complete. Summary statements of skills assessed to inform parents about their child's report are provided. These are benchmark standards representing the level below which a student is considered to be at educational risk. Teachers are now using practice tests to familiarise students with test conditions and the types of anticipated test questions to measure students' improvement efforts because of the lack of information about expected qualities of performance, or how these tests relate to learning in the curriculum, or to specific curriculum domain standards (Klenowski & Wyatt-Smith, 2010).

Media reports such as "Literacy and numeracy problems unchecked for a decade", "Publish schools results: Rees", "Crunch Time at School: National test results must be publicly available to all states", "Gillard praises state for lead on ratings", " 'Flying squads' to lift state teaching" are common in Australia. The use of teams of teachers headed by a school principal evaluating teaching practices and making recommendations to improve them is a clear sign that teachers are now being required to account for test scores and the assessment stakes have been raised considerably. Such quick fix approaches to improvement are not effective as the overriding goal is now focused on higher grades or results per se rather than the issues related to assessment *and* learning.

The testing regime in Australia is expanding with the Director of the Australian Council for Educational Research, Geoff Masters, commissioned by the Queensland Premier, to conduct a review of Queensland's education system. The resultant Masters' report (Masters, 2009) described the system as lagging behind other states in national exams (Dunleavy, 2009). The recommendations included the use of practice tests and "that standard science tests be introduced at Years 4, 6, 8 and 10 for school use in identifying students who are not meeting year-level expectations

and for monitoring student progress over time." It was subsequent to this report that the $9 million dollar initiative to set up "flying squads" was proposed.

It was also a consequence of this report that testing became enlisted in the service of improving teacher quality. Even though there have been high levels of accountability and quality assurance for several years in pre-service teacher preparation in Queensland, the government approved that pre-service teachers' content and pedagogic knowledge and skills in literacy, numeracy and science be tested in the Bachelor of Education Primary Program (Masters, 2009). Interestingly, the pre-service teacher's test results in these areas (in addition to their academic results) and registration requirements have become tied. This is another addition to Australia's testing regime. At the time of writing the implications of failing the test are not known including the legal implications.

Many teachers, principals, parents and the teachers' union in Australia are critical of the expanding testing regime and the 'quick fix' approaches being taken in some states. Some principals in particular are taking action to resist the lure of simplistic measures that appear to communicate indexes of quality in which it is claimed that the community, parents and students can have faith.

Australian Primary Principals Association (APPA) Position and Principles

The Australian Primary Principals Association representing 7200 Government, Catholic and Independent principals recently released a position paper to make its stance clear on the publication of nationally comparable school performance information. While APPA is supportive of high academic standards and the use of achievement data in the key areas of the primary curriculum in its position paper (APPA, 2009) there is a call for the responsible release of information about the resources available to schools and the performance of their students.

APPA acknowledges the negative impact of high stakes assessment on the quality of teaching and learning when there is a shift to focus only on the results in the evaluation of school performance and when sanctions are imposed. These unintended consequences have been identified in terms of: the narrowing of the curriculum as teachers teach only that which is to be tested; curriculum areas that are not tested are neglected; higher order thinking skills that are difficult to assess in such paper and pencil formats are also neglected; time is spent on coaching and practice tests; schools participate in perverse practices designed to improve achievement data, and finally, as stated in the APPA position paper, "a testing industry grows which is driven by its own commercial interests" (p. 5).

In this article the authors refer directly to the examples of such unintended or perverse practices identified by APPA from the 2010 NAPLAN tests. While only a few instances have been identified, it is timely to heed the signs from the United States where Amrein-Beardsley, Berliner, and Rideau (2010) have researched how increased pressures on teachers to ensure improved tests results leads to cheating. These researchers studied the types and degrees to which a sample of teachers were

aware of, or had participated in, these practices because of the *No Child Left Behind* (NCLB) high stakes testing policies. A taxonomy of cheating based on definitions of first, second and third degree offences in the field of law was identified. They concluded that:

> Policies that clearly undermine the moral and professional behavior of America's teachers need to be debated more thoroughly, and such policies must be challenged if their negative effects outweigh their positive effects on the educational system of our nation. It serves no one's interest to have policies that inherently promote cheating, and even justifications for cheating by educators, because the policy environment in which they work has become so onerous. There are better ways to design accountability systems. (p. 27)

APPA has identified practices and related consequences that send a clear message about the unintended consequences that are emerging in Australia. These include pressures on leaders to lift performance, threats to their jobs if results do not improve, more attention given to those students who are more likely to achieve better grades, neglect of those students who have the greatest need for support, the emergence of commercial tests that have not been quality assured, increased absenteeism for low performing students on the day of the test and increased instances of cheating. Several researchers have identified such consequences as they have occurred in other countries including the United States (Darling-Hammond, 2010), England (Stobart, 2008), and Singapore (Kramer-Dahl, 2008). Common across these contexts is the high-stakes and high accountability nature of testing that gives prominence to a narrow set of outcomes which tend to distort learning and teaching. It should be noted that not all countries have made public high stakes testing results in league tables: New Zealand, Scotland, Wales and the Republic of Ireland being notable examples.

In the Australian context, however, there has been a strong accountability push to install national testing and the concurrent publication of results in literacy and numeracy. With the introduction of the Australian Curriculum planned for first time offer in 2011, it is significant to note the ongoing debates around standards, even at the time of such offer. While the matter of testing and national curriculum has not surfaced to date, drawing on APPA (2010), the identification of test related developments below is timely. These include:

- Some line managers exerted **pressure on principals to improve** their test results at all costs without taking into account what the school has been doing to improve the students' performance and the particular factors that have made progress so challenging. As a consequence, principals reported feeling unfairly 'threatened' if they failed to treat raising the average test performance as their absolute goal. It was implied that their job would be on the line if the school's results did not improve.

- Some schools were required by their line manager to **lift their results** by a certain percentage. These schools then identified the students most likely to show improvement if given extra assistance. They then allocated their resources to this select group of students. Other students with greater needs did not receive as much attention for the first five months of the year until the completion of the NAPLAN tests.
- A plethora of **commercial products** have been produced and are now available from retail outlets. Companies are contacting schools offering to test their children and provide the results prior to NAPLAN testing. Assessment of this kind is inappropriate and can undermine good teaching.
- The media has reported that some schools have encouraged parents to **keep their children at home** on test day if the school judged that the student would not perform well in the tests.
- A small number of teachers have provided **assistance to students while sitting the tests to improve their test results**, in some cases arguing that the students knew the answers but were confused or overly anxious on the day. (APPA, 2010, p. 6)

A set of guiding principles governing the reporting and use of NAPLAN has been developed by APPA (APPA, 2010, p. 2) in an effort to protect primary schools from the unintended consequences and to ensure that the national transparency agenda has a positive impact on the primary curriculum. These principles include first, "making informed and balanced judgements" that involve evaluations of schools' and systems' performance based on multiple sources of reliable evidence that relate to not just the academic goals but include the key socio-emotional goals of schooling. This is recommended to be a responsibility of the school and to take place through the development of appropriate appraisal systems rather than have the Australian government develop more quantitative indicators.

There is increasing recognition in the published research on inquiry approaches to the relationship between curriculum, learning and assessment and conceptual understandings of such relationships (Assessment Reform Group 2002a, 2002b; Wyatt-Smith, Klenowski, & Gunn, 2010; Looney & Klenowski, 2008). In this growing body of work there is clear focus on how curriculum, assessment (including testing) and reporting standards align (Biggs, 1996; Harlen & James, 1996; Meisels, Atkins-Burnett, Xue, & Bickel, 2003), bringing with it a heightened interest in teacher judgement, especially in the context of standards-reform (Klenowski & Wyatt-Smith, 2010; Wyatt-Smith & Klenowski, 2010), and social moderation, especially as it involves teacher use of defined standards.

The second principle recognises the complexity of the factors that impact on school performance. To illustrate: "… there are many systemic and local factors that mediate the performance of students on NAPLAN and which invalidate simple comparisons of school performance" (APPA, 2010, p. 2). Currently the MySchool website publishes 'like school' comparisons of school performance based on the

Index of Community Socio-Educational Advantage (ICSEA) scale. This scale was developed for the MySchool website to identify schools with similar student populations and to measure key factors that correlate with achievement indicators as suggested by NAPLAN results rather than general measures of socio-economic status.

The MySchool website claims that "it uses a new index of student and school characteristics, developed specifically for the purpose of identifying schools serving similar student populations. This enables schools' results on national tests to be understood in a fair and meaningful way, and enables schools seeking to improve their performance to learn from other schools with statistically similar populations" (www.myschool.edu.au). Comparisons are made between average NAPLAN scores achieved by students at a particular school that is being viewed on the MySchool website and the average for the group of schools to which it is statistically similar. APPA (2010) reports that the ICSEA scale "does not produce results that are fine-tuned enough to yield an accurate score for all schools" and therefore in 2010 "misrepresented the differences in the intake for schools that were supposedly alike" (p. 2). APPA recommends that MySchool be developed for purposes of inquiry but not to publish school results and that the 'like schools' concept be abandoned.

In a recent public address Professor Harvey Goldstein discussed the MySchool website and stated that "in comparing the performance of schools, it is important to take into account differences in their student intakes" (Goldstein, 2010). He made the telling point that "comparisons of schools that are not statistically similar can lead to misleading conclusions about their performance". He went on to indicate the approach taken in Australia to identify "a set of variables that best predicted student performance on the combined NAPLAN tests on reading and numeracy, and then use these to create an index for grouping 'similar' schools." In practice this approach means that "if it is a good predictor then 'similar' schools are those with similar mean test scores – so schools are compared just with those having similar performance!" However, there are recognised concerns that emerge from a close examination of the current 'prediction' formula that combines: parental background information, occupation and education, post code. That is, a derived socioeconomic variable. Specifically the concerns include the reliability of the measures for cross-school comparisons. Public confidence in the certainty of the published results and what they represent is ill-founded (Goldstein, 2010; Wu, 2010). For example, parents and the wider community have asked what is the relevance of identifying 'like schools' that are located in regions or even states apart, if the intention is to inform choice of school and convey clear messages of quality of performance. The statistical challenge remains: how do league tables factor in contextual variables that directly relate to quality education in local sites? Further, the authors question the relevance of investing millions of dollars in the development and use of prediction formulae and national testing when a more appropriate expenditure of such funds would be at the local level. Given that the main predictor of quality student outcome is quality classroom teaching and assessment and further, that we

know that teachers' professional knowledge in quality assessment is limited, why is the money not directed to professional development with a focus on literacy, numeracy and assessment?

The ethical use of rewards and sanctions is the third overarching principle. This principle aims to prevent the unintended consequences of pressure on schools to improve results "at all costs". Presently governments are allocating funding of "hundreds of millions of dollars to states and territories that achieve performance targets" (APPA, 2010, p. 3). This reductionist approach provides ambiguous and narrow meaning. "When funding decisions are treated as unambiguous, and when single scores are generalised beyond justification as true characterisations of individuals and systems, the potential for mischief is enormous" (Shavelson, Black, Wiliam, & Coffey, 2004, p. 35). Such unintended consequences were recently reported in Australia at the time of the publication of the NAPLAN results. In Queensland claims of cheating including the provision of extra time for students to complete the test were investigated. Such investigations about accountability were given prominence at the school and departmental levels. In recognising the challenges facing schools and education systems APPA recommends that Ministerial Council for Education, Early Childhood Development and Youth Affairs (MCEECDYA) should provide guidelines that specify what is possible for states, education systems and schools to do to enhance NAPLAN results.

The fourth principle is that schools should have the capacity to challenge inferences drawn from NAPLAN results about their performance that are misleading and damaging to their reputation. In 2009 the publication of NAPLAN results on the MySchool website led to the media publishing grossly simplistic results or even misinterpreted the results pertaining to individual schools yet governments did not intervene. An independent ombudsman with the capacity to quickly follow up complaints from schools is a further recommendation from APPA (APPA, 2010, p. 3). Similarly it has been recommended that an independent body should be established to monitor the impact of NAPLAN to alleviate any unintended consequences (APPA, 2010, p. 4). APPA suggests that MCEECDYA needs to appoint an independent group to monitor the implementation of MySchool and report on an annual basis. Even if the reliability of the measures used for reporting large scale test results could be addressed as discussed earlier such a group could consider how schools could account locally using classroom summative assessment data to supplement that provided in the form of tests results. This would address the currently unresolved issues between system and site validity issues (Freebody & Wyatt-Smith, 2004).

The final principle is to make "NAPLAN fully transparent" (APPA, 2010, p. 4) as presently researchers and policy analysts cannot access information regarding the development of the tests, their properties and other aspects of the information contained on the website. Opening up access to researchers "to de-identified data to replicate findings reported by ACARA" and to research and carry out different analyses would allow the NAPLAN data base to be used as a powerful research tool (APPA, 2010, p. 4).

It is vitally important that the reliability and validity of the tests are researched. For doubts exist about what some NAPLAN tests are actually testing and how they relate to the support of learning. Willett and Gardiner (2009) in their critique of the NAPLAN spelling test draw on their longitudinal equating study in raising significant questions about the validity and reliability of NAPLAN achievement data of this test. These researchers indicate how this type of test does not assist teachers' practice or student learning rather the information provided to teachers is unhelpful and incorrect and likely to have a negative impact by encouraging discredited spelling constructs and teaching methods. They illustrate how the construction of spelling items is formulaic which they suggest is due to the lack of an articulated research-based framework and the desire to keep the item 'pure' by trying to ensure that the items have a single item demand (Willett & Gardiner, 2009, p. 5).

Table 1: Formula for creation of spelling items (Willett & Gardiner, 2009)

Leave out a letter	craked (cracked), weel (wheel), frends (friends) Used at the syllable junction eg swiming (swimming), disapointed (disappointed)
Add a letter	Used at the syllable junction consummed (consumed), fittness (fitness)
Use a different vowel combination	broun (brown), arownd (around), lowdly (loudly), seet (seat),
Substitute a letter	cumplained (complained), sinse (since)
Reverse a letter sequence	muscel (muscle), marothan (marathon)

Constructing items in this formulaic manner contrasts with the authentic student spelling errors. This approach is likely to encourage the teaching of test preparation rather than productive spelling knowledge and skills. These researchers were also able to demonstrate the negative impact of such testing by providing evidence of how the misspelling of the first syllable (com) in complain as cumplain, is not an error that Year 3 students make, yet after exposure to the NAPLAN error, students in their study used this misspelling when attempting to spell the word! This exemplifies the point made that the way students are assessed impacts on the way they learn and what they learn.

There is an already strong and growing body of published research attesting to the limitations of large-scale high stakes testing. There is a resonance in several countries including England and the United States between this work and the concerns raised by APPA. Recurring themes in the discussion of the consequences of high stakes assessment include: narrowing of the curriculum (Wearmouth, 2008); children experiencing constant stress throughout their school lives (SFS Group, 2008), and the use of published test results as a tool for control and to encourage parents to use the information to select schools for their children (Hall & Ozerk, 2008). Additionally there are well recognised concerns about under-utilising the

professional abilities of teacher and focusing disproportionate resources on borderline students to raise their achievement outcomes (House of Commons, Children, Schools and Families Committee, 2008). Importantly while there is a reported over-emphasis on basic skills and a concurrent neglect of higher order and critical thinking in both testing and classroom practice, more recently there are clear signs about the limitations of current print-dependent testing. These draw attention to the exclusion in testing practice of 21st century skills including working in teams and online, to use and create knowledge.

Directions Forward

In presenting the directions forward for quality assessment in the context of national curriculum in Australia we draw on some recent research studies of direct relevance to the authors' main concerns in this article with classroom assessment and its relation to testing and accountability. These include a major federally funded investigation of teacher judgement and moderation in the context of standards-driven reform in the middle years of schooling (Wyatt-Smith, Klenowski, & Gunn, 2010), a study of teacher generated assessment tasks, standards and moderation (Wyatt-Smith & Bridges , 2008) and a further study of online assessment of centrally devised curriculum tasks to achieve system accountability purposes (Klenowski, 2007). Also relevant is a study of teacher use of assessment evidence including classroom based assessment and literacy testing results (Cumming, Wyatt-Smith, Elkins, & Neville, 2006).

Building on the principles and recommendations put forward by APPA (APPA, 2010; 2009) we suggest that national testing can provide limited data for diagnostic use to inform teaching and pedagogical interventions for the improvement of learning. Testing of this type realistically can only ever deliver a snapshot, point in time evidence of performance achievement. Currently the Australian NAPLAN testing regime has limited utility in informing the Australian people how children are learning in the curriculum.

The direction therefore is for a modest recognition of what these tests can achieve and communicate about student learning. A related message is for a richer and comprehensive set of achievement indicators for student learning.

If national testing programs are to have a genuine purpose of improving outcomes, as distinct from reporting outcomes, then we need to reach agreement that the teacher, not the test, is the primary change agent. If we agree on this, then we must bring teacher judgement to centre stage. The point is that teacher judgement is central to a much-needed review and discussion of all performance evidence, including that generated in standardised testing and in classroom-based programs.

The direction from this observation is that the professional abilities of teachers should not be minimised by high stakes testing. Instead teacher judgement is at the heart of efforts to improve learning outcomes especially for those at educational risk. A related direction is to divert some of the funding for test development and trialling into

professional development opportunities to build teacher assessment capabilities especially in task design and the use of achievement standards.

In the context of Australia's planned move to a national curriculum there is an urgent need to make explicit performance expectations for literacy and numeracy education and the relationship with *curriculum literacies* (Wyatt-Smith & Cumming, 2003).

This is a call for achievement standards to be inclusive of curriculum knowledges and capabilities as well as literacy and numeracy demands of the curriculum to improve learning. A related call is for exemplars with an accompanying commentary to be developed as concrete demonstrations of ways to meet the standards.

To achieve the intended diagnostic purpose of reported test results the capacity of teachers to interrogate and analyse achievement data is needed. The assessment literacy of teachers especially in regard to using and interpreting assessment evidence is particularly important in the context of national curriculum and achievement standards reform.

Teacher professional development programs are needed to assist teachers to distinguish between assessment with teaching-learning significance and assessment with measurement significance.

Quality and equity are the central to efforts to achieve real improvement. Assessment is necessarily contextualised and value-laden. There is no such thing as value-free assessment. Dwyer (1998) made the point about cut scores, writing that "any use of a cut point, no matter how sophisticated or elaborate its technical apparatus, is at heart a values decision. The underlying question in setting any cut score can be phrased quite simply: 'How much is enough?' There is, of course, no technical answer to that question; there is always a value answer to it" (p. 18). There is a need for debate about the utility of large-scale high stakes testing for informing the public about the quality of schools, teachers and systems.

This suggests a need for public scrutiny of test design, principles and practices and government expenditure and resourcing for testing. In addition, there is a need to review the place of web reporting of test results and the media production of simplistic league tables.

It is time to critique the flawed thinking associated with an assumed connection between testing and learning improvement. The divergent priorities and goals of key education stakeholders in Australia are well recognised as is the pressure on educational leaders to follow short-term political imperatives of appearing to be delivering improved results. As many have argued the challenge for the educational community is to ward off this pressure, focussing instead on providing support for the long-term professional development change necessary to effect actual pedagogical change and improved outcomes and a more equitable society. This includes attention to avoiding test irregularities such as providing answers to exam

questions and the reduction in Native language and culture responsive teaching (McCarty, 2009; Patrick, 2008).

Finally, we note that there has been a pervasive silence around the rights of the child/student and the ways in which they have been positioned by testing and accountability priorities. There are examples of alternative systems of accountability that have been described as more intelligent and that recognise the complexities of assessment purposes, modes, conditions and contexts. It is timely to investigate these alternatives with several countries including New Zealand, Finland, Scotland, Wales and England committed to introducing more inclusive, equitable and balanced assessment systems. These include national tests complemented by teacher assessment and moderation practice and sampling rather than census testing.

In reflecting on the diversity of practices across the alternative systems being adopted for accountability and learning improvement the authors reassert the central relationship between curriculum and assessment. In any reform it is important to keep the learner, learning and curriculum to the fore ensuring that assessment practices have integrity with the intended curriculum, validity and reliability of reported results. Where such integrity is compromised the demands placed on testing and public reporting of results can give undue dominance to efforts to lift test results as distinct from learning improvement.

Readers may be interested to learn that, in the period of the review of this article, there has been a change of Federal Minister of School Education and changes and additions to the MySchool website. The authors agree with APPA that an opportunity has been missed tollow the advice by the MySchool Working Party. The APPA President, a member of the MySchool Working Party, has indicated that she is "particularly disappointed with the decision to maintain the privileged status of NAPLAN data on the website... [rather than] allow a school story to be told... [allowing] a school's NAPLAN results to be placed in their proper context" (Available at: http://www.appa.asn.au/index/php/appa-business/news-items/1145). This stance is also clear in Wu's message to teachers that NAPLAN results cannot reflect teacher and school performance. She emphasises that NAPLAN results should never be published and that parents should not be encouraged to use the results to judge schools (Wu, 2010).

In conclusion, we put to readers the need to move beyond the single indicator for success and recognise the multifaceted nature of learning and achievement over time. For intelligent accountability, the teacher's role remains central. This is particularly the case in times of curriculum and assessment reform. Given the considerable investment in testing initiatives in so many countries, with conflicting evidence of benefit, it is timely to begin the international conversation about national investment in continuing professional capacity building teachers and school leaders.

References

Adie, L., Klenowski, V., & Wyatt-Smith, C. (submitted September 2010). Towards an understanding of teacher judgement in the context of social moderation. *Oxford Review of Education*.

Amrein-Beardsley, A., Berliner, D., & Rideau, S. (2010). Breaking professional law: Degrees of cheating on high stakes tests. *Education Policy Analysis Archives, 18*(14), 2-33.

Assessment Reform Group (2002a). *Assessment for Learning: 10 principles research-based principles to guide classroom practice*, Assessment Reform Group, London, United Kingdom.

Assessment Reform Group (2002b). *Testing, Motivation and Learning*, Assessment Reform Group, London, United Kingdom.

Australian Primary Principals Association (APPA). (2010). *The reporting and use of NAPLAN*. Kathleen, Australian Capital Territory: Author. Available at: www.appa.asn.au

Australian Primary Principals Association (APPA). (2009). *Australian Primary Principals Association position paper on the publication of nationally comparable school performance information*. Kathleen, Australian Capital Territory: Author. Available at: www.appa.asn.au

Baker, E. L., & Stites, R. (1990). Trends in testing in the USA. *Journal of Education Policy*, *5*(5), 139-157.

Nichols, S.L., & Berliner, D.C. (2010). How high-stakes testing corrupts America's schools. Cambridge, Massachusetts: Harvard Education Press.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347-364.

Bligh, A. (2009). *Letter to Parent*. Brisbane: Queensland Government.

Broadfoot, P., & Black, P. (2004). Refining assessment? The first ten years of assessment in education. *Assessment in Education: Principles, Policy and Practice*, *11*(1), 7–27.

Cumming, J. J., Wyatt-Smith, C. M., Elkins, J., & Neville, M. (2006). *Teacher judgment: Building an evidentiary base for quality literacy and numeracy education.* Brisbane, QLD: Queensland Studies Authority. Available at: www.qsa.qld.edu.au/downloads/publications/research_qsa_teacher_judgment.pdf

Darling-Hammond, L. (2010). *The flat world and education: How America's commitment to equity will determine our future.* New York: Teachers' College Record.

Dawkins, P. (2007). *Federalist Paper No 2: The future of schooling in Australia.* Victoria, Australia: The Council for The Australian Federation.

Delandshere, G. (2002). Assessment as inquiry. *Teachers' College Record, 104*(7), 1461-1484.

Department of Employment, Education, Training and Youth Affairs. (1998). *Literacy for All: The Challenge for Australian Schools*, Australian Schooling Monograph Series No. 1. Canberra: Australian Government Printing Service.

Dunleavy, G. (2009). *Flying squads to rescue struggling schools*. Accessed 15 October, 2009 from www.bribanetimes.com.au/queensland/flying-squads-to-rescue-struggling-schools-20090908-ffco.html

Dwyer, C. A. (1998). Testing and affirmative action: Reflections in a time of turmoil. *Educational Researcher, 27*(9), 17-18.

Earl, L. M. (2005). *Thinking about purpose in classroom assessment: assessment for, as and of learning.* Deakin West, ACT: Australian Curriculum Studies Association.

Freebody, P., & Wyatt-Smith, C. (2004). The assessment of literacy: Exploring the tension between 'system' and 'site' validity. *Journal of Educational Enquiry*, *5*(2), 30-49.

Goldstein, H. (2010) *School League Tables: Who is accountable?* Presented at the Queensland University of Technology, 25 October, 2010.

Harlen, W., & James, M. (1996) Creating a positive impact of assessment on learning, paper presented at the Annual Meeting of the American Educational Research Association, New York, United States, Education Resources Information Center, viewed July 2008, http://eric.ed.gov.

Hall, K., & Ozerk, K. (2008). *Research briefing: Primary curriculum and assessment: England and other countries* (Primary Review Research Survey 3/1). Cambridge: University of Cambridge, Faculty of Education.

Hutchinson, C., & Hayward, L. (2005). The journey so far: Assessment is for learning in Scotland. *The Curriculum Journal*, *16*(2), 225–248.

House of Commons, Children, Schools and Families Committee S&FNC (2008) *Testing and Assessment Third Report* London: UK Parliament.

Klenowski, V. (2007). Evaluation of the effectiveness of the consensus-based standards validation process. Available at: http://education.qld.gov.au/corporate/newbasics/html/lce_eval.html

Klenowski, V., & Wyatt-Smith, C.M. (2010). Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform. *Assessment Matters*, *2*, 107-131.

Kramer-Dahl, A. (2008). Still and examination culture - for most: Singapore literacy education in transition. Point and Counterpoint, *Curriculum Perspective,* 28 (3), 82-89.

Lee, J., & Wong, K. K. (2004). The impact of accountability on racial and socioeconomic equity: Considering both school resources and achievement outcomes. *American Educational Research Journal, 41,* 797-832.

Linn, R. (2003). Accountability: Responsibility and reasonable expectations. *Educational Researcher, 32*(7), 3-13.

Looney, A., & Klenowski, V. (2008). Curriculum and assessment for the knowledge society: Interrogating experiences in the Republic of Ireland and Queensland, Australia. *The Curriculum Journal*, *19*(3), 177–192.

Masters, G. (2009). *A shared challenge: Improving literacy*, *numeracy and science learning in Queensland primary schools*. Camberwell, Victoria: Australian Council for Educational Research.

McCarty, T. (2009). The impact of high-stakes accountability policies on Native American learners: Evidence from research. *Teaching Education*, *20*(1), 7-29.

Meisels, S., Atkins-Burnett, S., Xue, Y., & Bickel, DD (2003). 'Creating a System of Accountability: The impact of instructional assessment on elementary children's achievement scores', *Educational Policy Analysis Archives*, 11(9), Arizona, United States.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral damage: How high stakes testing corrupts America's schools.* Cambridge, Massachusetts: Harvard Education Press.

Patrick, R. (2008). Perspectives on change: A continued struggle for academic success and cultural relevancy at an American Indian school in the midst of No Child Left Behind. *Journal of American Indian Education*, *47*(1), 65-81.

SFS Group. (2007). UK pupils facing exam stress. Available at: www.sfs-groupco.uk/uk-pupils-facing-stress/?parent=11

Shavelson, R.J., Black, P.J., Wiliam, D., & Coffey, J. (2004). *On linking formative and summative functions in the design of large-scale assessment systems.* Retrieved 3 May, 2007, from www.stanford.edu/dept/SUSE/SEAL/Reports_Papers/On%20Aligning%20Formative%20and%20Summative%20Functions_Submit.doc

Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. London: Routledge.

Wearmouth, J. (2008). Testing, assessment and literacy learning in schools: A view from England. Point and Counterpoint, *Curriculum Perspective,* 28 (3), 77-81.

Willett, L., & Gardiner, A. (2009). *A critical analysis of the NAPLAN spelling test*. Paper presented at the International Association for Educational Assessment, 35th Annual conference, September, 2009, Brisbane.

Wu, M.L. (2010). The (In)Appropriate Use of NAPLAN Data. Available at:http://www.qtu.asn.au/830626.html

Wyatt-Smith, C. M., & Cumming, J. J. (2003). Curriculum literacies: Expanding domains of assessment. *Assessment in Education: Principles, Policy and Practice*, 10(1), 47-59.

Wyatt-Smith, C.M., & Klenowski, V. (2010). The role and purpose of standards in the context of national curriculum and assessment reform for accountability, improvement and equity in student learning. *Curriculum Perspectives 30*(3), 37-47.

Wyatt-Smith, C. M., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 59-75.

Wyatt-Smith, C. M., & Bridges, S. (2008). *Meeting in the middle – assessment, pedagogy, learning and students at educational disadvantage*. Final Evaluation Report for the Department of Education, Science and Training on Literacy and Numeracy in the Middle years of Schooling. Available at: http://education.qld.gov.au/literacy/docs/deewr-myp-final-report.pdf

[i] Readers are referred to www.naplan.edu.au/

[ii] Readers are referred to www.myschool.edu.au/

[iii] Now called Department Education, Employment and Workplace Relations

[iv] Readers are referred to the Queensland Studies Authority website for further information https://naplan.qsa.qld.edu.au/naplan/Usermanual.pdf;jsessionid=21FC15D5461E939D620064E81613F2F1