Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform

Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform

Val Klenowski and Claire Wyatt-Smith

ABSTRACT

This paper puts forward a proposal for reviewing the role and purpose of standards in the context of national curriculum and assessment reform more generally. It seeks to commence the much-needed conversation about standards in the work of teachers as distinct from large-scale testing companies and the policy personnel responsible for reporting. Four key conditions that relate to the effective use of standards to measure improvement and support learning are analysed: clarity about purpose and function; understanding of the representation of standards; moderation practice; and the assessment community. The Queensland experience of the use of standards, teacher judgement and moderation is offered to identify what is educationally preferable in terms of their use and their relationships to curriculum, improvement and accountability. The article illustrates how these practices have recently been challenged by emerging political constraints related to the Australian Government's implementation of national testing and national partnership funding arrangements tied to the performance of students at or below minimum standards.

Introduction

International measures of educational attainment, such as the Programme for International Student Assessment (PISA) developed by the Organisation for Economic Co-operation and Development (OECD) or the Trends in International Mathematics and Science Study (TIMSS) of the International Association for the Evaluation of Educational Achievement (IEA), have had a major impact on international curriculum and assessment reform. These drivers for educational reform have led

countries to introduce national educational standards, as Germany did for seven subjects in 2003 and 2004 (Kōller, 2009), or have prompted countries such as Australia and Norway to plan for and develop national curriculum and achievement standards. Other countries are also in the process of developing standards. For example, in New Zealand there is a focus on standards for literacy and numeracy (Crooks et al., 2009), and in the USA the Common Core State Standards Initiative (CCSSI) is developing common core state standards (National Governors Association Center for Best Practices, & Council of Chief State School Officers, 2010).

The various approaches adopted in the use of standards for accountability and learning purposes have been debated (Mansell, James, & Assessment Reform Group, 2009; Mortimore, 2008; Popham, 2008; Reid, 2009). Related issues of the political framing of standards (Price, O'Donovan, Rust, & Carroll, 2008; Zepke et al., 2005) and the politics of standard setting (Darling-Hammond, 2004; Sacks, 1999) have also been reviewed and analysed, indicating the need for clarity about the nature of standards, their purpose and the need for effective communication about standards between stakeholders (teachers, parents, students, employers, professional bodies and government). In this article we acknowledge that there are legitimate political concerns about standards and their role in informing progress over time. This observation holds for all levels of schooling, including higher education, as evidenced in the Australian context with the strengthening role of the Tertiary Education Quality and Standards Authority in 2010. However, we wish to make explicit how different sectors (political and educational) view standards in relation to schooling, and to emphasise how the competing beliefs about standards can result in unintended consequences for student learning.

The key questions that frame this article are:

- What are the conditions required for standards to not only be used to measure improvement but also to inform student learning and teaching for improvement purposes in the context of national curriculum and assessment reform?
- What evidence is there that teachers' use of standards for system reporting is a valid and reliable practice?

The Queensland experience of the use of standards, teacher judgement and moderation is offered in this article to identify what is educationally preferable in terms of their use and their relationship to curriculum, improvement and accountability. Recently emerging political constraints to these practices are discussed and highlighted as the Australian Government implements national testing and introduces the national partnership funding arrangements tied to the performance of students at or below minimum standards.

We begin by discussing how the term *standards* can stand for many different ideas and aspirations and how the use of the term will vary according to the context of its use. We then introduce a framework of the four conditions for the effective use of standards, followed by a discussion of the conditions as they relate to the recent experience in Australia of standard setting and implementation. In this section of the article we aim to analyse how interpretations of assessment scores in contexts of test-based educational accountability highlight the need to understand how measures of quality are communicated, especially when such measures are represented as standards. Of particular interest is the "fit" between how standards are formulated and how they are used in practice, by whom and for what purposes.

What are standards?

In pursuing the key questions related to the functions and purposes of "standards", and the evidence that supports teachers' use of standards for system reporting, we will begin by looking at the different definitions of standards. The meaning is found to vary according to the context in which the term is used and the purpose (goal) and function (role) it fulfils.

More specifically, the word *standard* is ubiquitous yet difficult to define, because its meaning is derived from its historical and social context, and so different countries and even different states or districts within countries can have varying views about what constitutes a standard (Goldstein & Heath, 2000). Dictionary definitions also illustrate that the term has different meanings and that they change with time and will continue to do so. The concept of standards is therefore elusive, and confusion can often occur when the term is used in official documents or when making comparative judgements, because it is not always clear what meaning is intended.

The distinction between content and achievement standards needs to be stated at the outset because they are both often referred to in the context of assessment. *Content standards* apply to schools and systems, and generally refer to the knowledge and/or processes that are taught. Maxwell (2009) emphasises that these standards help schools to develop their curriculum in relation to their local contexts. *Achievement standards* (also referred to as performance standards) apply to students, and refer to what they have learnt. They are usually embedded in the tasks the students need to complete by drawing on their knowledge and skills (Marsh, 2009). They are used for summative assessments and to report on the quality of the achievement or performance of the student. These standards can also be used formatively to inform students of their strengths and areas for development (Maxwell, 2009).

Lessons learnt from national curriculum and assessment reform include the need for certain conditions to be addressed when implementing standards (Goldstein & Heath, 2000; Mansell et al., 2009; Stobart, 2008). These conditions are now discussed.

First condition: The purposes and functions of standards

The first condition is to be clear about the *purposes* of standards and their *functions*. This is important in a context where there is a growing global trend for using standards not just for accountability but also for the purpose of improving learning (Popham, 2008; Stobart, 2008; Wyatt-Smith, Klenowski, & Gunn, 2010).

Standards for improving learning

Achievement standards are intended to indicate the quality of achievement that is expected and to provide the basis for judgements about the quality of students' work. The purpose is to use the standards to improve student learning. Research indicates that standards are useful for the purpose of informing teachers' work and in contributing to quality teaching and learning experiences (Sadler, 2005; Stanley, MacCann, Gardner, Reynolds, & Wild, 2009; Wyatt-Smith & Castleton, 2004). Standards, as

descriptors of student achievement, function by monitoring the growth in student learning and providing information about the quality of student achievement to fulfil the purpose of improving student learning. Standards used to assess the quality of learning help teachers to identify areas for improvement in teaching, curriculum design or development.

Wyatt-Smith and Gunn (2009) have explored the pedagogical potential of standards and have proposed a critical inquiry approach to assessment, as suited to learning in the 21st century. They argue that a system of standards can indicate what teachers are expected to teach and the level of performance expected for a particular age group. Further, mandated standards and the application of standards in the assessment activities that teachers design can serve to focus teacher and student attention on the expected features of quality (Klenowski, 2006, 2007). As such, standards can help to meet the demand for public accountability at the local professional level of the teacher (Harlen, 2005; Wilson, 2004).

In Australia, an example of standards for the improvement of student learning is Queensland's standards for the Essential Learnings, which provide a generic description of the expected quality of student work (using an A–E framework), and provide a common language for teachers to use in discussing student work (Queensland Studies Authority, 2007). The standards used to assess the Queensland Comparable Assessment Tasks are a related example. These standards are intended to promote teachers' professional learning, focusing on learning-supportive assessment practices and judgement of the quality of student achievement against system-level benchmarks or referents. It is also expected that teachers using the standards will present more meaningful reports and show better engagement with assessment as a learning process.

Standards for accountability

For the purposes of accountability, standards defined as "quality benchmarks" (expected practice or performance), "arbiters of quality" (relative success or merit) and "standards as milestones" (progressive or developmental targets) (Maxwell, 2002, p. 1) seem most appropriate. Standards as quality benchmarks describe "an expected or typical outcome" and require representation on a continuum that defines a

minimum acceptable level (Maxwell, 2008, p. 2). Standards as arbiters of quality and standards as milestones represent differentiated levels of performance. These representations fulfil the purpose of accountability by acting as benchmarks, arbiters or milestones. They differ in terms of focus and time frame, so that, typically, standards as arbiters of quality function by focusing on a single assessment event, while standards as milestones function by providing judgements over time along a continuum of learning. Standards defined in these ways provide a common frame of reference and a shared language for communicating student achievement. Standards need to be described in such a way that schools can relate to them.

Governments and policy makers enact high-stakes assessments and aim to set high standards of achievement to improve education and inspire greater effort on the part of students, teachers and principals. However, the inadequacy of high-stakes assessments, in terms of their lack of sufficient reliability or validity for their intended purposes, can result in unintended consequences. For example, improvements in assessment results might not relate to improved learning; students might be placed at increased risk of failure or disengagement from schooling; teachers might be blamed or punished for inequitable resources that are beyond their control; and curriculum and teaching can become distorted if high grades per se become the overriding goal (Klenowski, 2008). The No Child Left Behind Act[1] in the USA is an example where the push to raise standards has led to enormous pressure on teachers and distortions in the teaching of a holistic curriculum, and a reduction in authentic and challenging learning experiences for students (Marsh, 2009; McCarty, 2009). In fear of job losses and school closures, teachers have resorted to test irregularities such as providing answers to exam questions and reducing Native-language and culture-responsive teaching (McCarty, 2009).

Second condition: Understanding the representations of standards

This second condition relates to the teacher's understanding of the representation of the standards. What becomes apparent is that irrespective of the context or purpose of standards, professional judgements are needed to describe and maintain standards, and this implies a degree of trust of the professionals. As described by Goldstein and Heath:

> It is difficult, if not impossible, to arrive at an 'objective' definition of educational standards. Despite claims to the contrary, ultimately the final appeal is to human judgement and no amount of technical sophistication can alter this. (2000, p. 8)

Trust between educators and the public is therefore a recurring topic in relation to the use of standards in curriculum reform. Policies based on comparisons of examinations, tests or other devices should be seen for what they really are: human judgements. Even multiple-choice tests are reliant on human (that is, subjective) judgement of choices of question design in terms of what is included (and excluded). However conscientiously pursued, assessments reliant on human decision making are ultimately subjective and are influenced by culture, personality and general perceptions of the external world (Goldstein & Heath, 2000, p. 8).

Defining examination or assessment standards requires interpretation and inference, so fundamentally they too are subjective or reflect the individual teacher's perception. The interpretation of high-stakes tests or examination results should be that they are an indication of what students can do, but not an exact specification (Cresswell, 2000). What should be assessed, and the levels of attainment that are comparable to those represented by each grade in other examinations or assessments in the same family (Cresswell, 2000), should be defined by the standards as used in examination and assessment systems for public reporting. However, to compare attainment in different subjects we can only use indirect bases for comparison, and for this we rely on statistics and expert judgement (Cresswell, 2000).

In the context of examinations for high-stakes testing, or the use of standards for improving learning, the teacher has an important role in a community of judgement practice. This is because standards-referenced assessment relies on teacher judgement that can be made dependable if standards are promulgated in appropriate forms and teachers have the requisite conceptual tools and professional training. Teacher judgement is central to the use of standards and moderation.

Standards are understood differently depending on their *context* and their *purpose*. The different representations and models of standards therefore need to be defined and understood in relation to the context and the purpose

for which they are used. The main methods include numerical cut-offs, tacit knowledge, exemplars and verbal descriptors (Sadler, 1987).

Artefacts such as exemplars or model answers can also represent standards. Exemplars help to explicate judgement practice and form one part of a comprehensive approach to moderation. Not only are annotated samples of each standard (A–E) required, but also an overall commentary for each, detailing the approach used to reach the judgement (i.e., holistic, analytic, trade-offs etc.). To improve and support judgement practice, exemplars need to be provided. However, we acknowledge that exemplars:

> use only a small sample of possibilities. There is a danger that they can narrow assessment and curriculum by encouraging teaching to the exemplars and the particular contexts in which necessarily they are embedded. (Goldstein, 2010, personal communication)

How the features of the standard are communicated can have an important effect on teaching and student learning. It is therefore important that exemplars be used in conjunction with achievement standards and moderation practice.

Third condition: Moderation practice

Our third condition relates to opportunities for teachers to share their interpretations of assessment criteria and standards by participating in moderation to reach consensus. Our focus is on "social moderation" (Linn, 1993) or "consensus moderation", as distinct from the more widely known statistical moderation. The former involves groups of teachers meeting to discuss and negotiate assigned gradings of student work with the aim of reaching consensus and a common understanding of the quality of work (Gipps, 1994).

Social moderation processes are a necessary part of school-based assessment. The purpose of social moderation is to produce valid and reliable judgements that are consistent with one another and with stated standards of performance (Wilson, 2004). In effect, it is through moderation that teachers develop a shared understanding of the meaning of standards and how to apply them in a range of cases. More than this—and arguably

more importantly—it is in the context of moderation that teachers act as a community of assessors: they talk about actual student work examples and examine how the work matches the expected features as specified in stated standards. And it is through such talk and the classification of the work against the standards that teacher judgement becomes "tuned in" or calibrated to achieve high levels of reliability or inter-rater consistency.

Teacher-based assessment is usually deemed to have high validity but questionable reliability. Others have argued that to achieve consistency in judgement involves assessors developing a common understanding of the standards as well as similar recognition of performances that demonstrate those standards (Maxwell, 2009). As discussed in what follows, the search for consistency of judgement calls for greater recognition of the dynamic and "faceted" nature of judgement processes and practices than previously.

Social or consensus moderation is not an optional extra in assessment systems: it is essential, on two fronts. First, moderation can provide the necessary checks and balances on teacher judgement, acting as a form of quality assurance for delivering comparability in evidence-based judgements of student achievement. Quality assurance refers to the methods for establishing confidence in the quality of procedures and outcomes. Comparability requires assessment against common characteristics or criteria, as provided by a subject syllabus or other frame of reference. It requires consistency in the application of common standards so that all achievements that have been given the same grade or level of achievement have reached the same standard (Maxwell, 2007). Further, efforts to use teachers' judgements of student achievement for the purposes of local-level assessment and system-level accountability necessarily require "a way to integrate teachers' judgements of students' responses to the various assessment modes with those of other teachers" (Wilson, 2004, p. 11). Second, moderation benefits curriculum design and delivery in the classroom. Once teachers buy in to moderation, either as a system initiative or local practice, it has been shown to build teacher assessment capacity, as well as teacher confidence in the judgements they make of student work.

Several writers (Harlen, 2005; Sadler, 1987; Wyatt-Smith et al., 2010) have emphasised how common standards provide external

reference points for informing judgement and are pivotal for achieving comparability. Consensus moderation means that the frames of reference (standards, scoring guidelines, assessment criteria etc.) must be defined and disseminated to allow for a common interpretation (Maxwell, 2007, 2009). This highlights the social nature of moderation, whereby teachers interact with one another as they make explicit their judgements of student work samples. When participating in these meetings, teachers express their interpretations of the evidence in relation to the standards and explicitly state their justifications for their judgements. Such disclosures would otherwise remain private and unarticulated.

In addressing ways to achieve high reliability while preserving validity, several writers have argued that it is important for teacher assessors to develop a common understanding of mandated standards and reach "similar recognition of performances that demonstrate those standards" (Maxwell, 2001, p. 6). However, clear communication about the nature of standards and the levels they seek to specify is not necessarily achieved by simply publishing stated standards.

This is illustrated by considering Sadler's (1998, p. 80) description of three elements that make up teacher judgement of student work: the teacher attending to the learner's production; appraising this against some background, or reference framework; and making an explicit response, such as assigning the learner's work to a class (as in grading). Sadler (1998, pp. 80–82) also identified some of the intellectual and experiential resources teachers must be able to draw upon when making a judgement of student work:

- superior knowledge of the content or substance of what is to be learnt
- deep knowledge of criteria and standards (or performance expectations) appropriate to the assessment task
- evaluative skill or expertise in having made judgements about students' efforts on similar tasks in the past
- a set of attitudes or dispositions towards teaching as an activity, and towards learners, including their own ability to empathise with students who are learning, their desire to help students develop, improve and do better, their personal concern for the feedback and veracity of their own judgements, and their patterns in offering help.

These insights make clear how teachers are viewed as the primary change agents who, through judgement practices that are integral to the requirements of assessment tasks and expectations of quality performance (Sadler, 1989), are best placed to identify important steps for students to improve in their learning and to develop useful insights about how best to change pedagogy to meet students' particular learning needs. Such classroom assessment to promote learning requires the use of assessment data by the teacher to inform teaching and to facilitate students' learning (Hattie, 2005). What is fundamental to attaining greater coherence between system-level accountability and local-level practice are teachers' judgements and informed interpretations of assessment data.

Further, it is widely recognised that standards written as verbal descriptors necessarily remain open to interpretation and call for qualitative judgements. To address this, Sadler (1998) argued that exemplars or samples of student work provide concrete referents for illustrating standards that otherwise remain tacit (unarticulated, "in the head") knowledge. He made the point that the stated standards and exemplars work together to show different ways of satisfying the requirements of say, an A or C standard.

Given that standards require interpretation, moderation provides the means through which teachers meet to review how they have interpreted and used given standards, and in this way moderation is a vital part of efforts to promote a more consistent use of standards over time and across sites. It is this understanding that has motivated countries such as Scotland, England and Wales, where there have been many years of high-stakes testing, to value moderation as an important practice worthy of implementation. Most recently, the Cabinet Secretary for Education and Lifelong Learning in Scotland stated:

> A national system of quality assurance and moderation of 3–18 will be developed to support teachers in achieving greater consistency and confidence in their professional judgements (Hyslop, 2009).

In Wales, national tests were abandoned and the value of school-based assessment and teacher moderation practice was recognised. The Report on *Future Assessment Arrangements for Key Stages (KS) 2 and 3* (Department for Education, Lifelong Learning and Skills, 2007) recommended

cluster-group moderation for transition links with Key Stages 2 and 3 schools. Assessment at the end of Key Stage 3 is strengthened by means of external moderation of sample evidence of teachers' understanding and application of the national curriculum-level descriptions, and verification of school-based systems and recognition of the quality of teacher assessment by awarding schools "accredited centre" status. Since September 2007 primary school teachers have also used school-based moderation, involving suitably robust systems and procedures to ensure they have appropriate opportunities to discuss their pupils' work and act on an agreed, shared understanding of standards.

Similarly, in England there has been a move to privilege teacher assessment through it being the only form of judgement for students aged 5, 7 and 14, and it is also becoming more significant in the assessment of 7–11-year-olds. The recent move away from testing is evident from the changes to the use of tests at Key Stage 1 in 2005, the abolition of national tests for 14-year-olds at Key Stage 3 in October 2008 and the removal of the science tests at Key Stage 2 in 2010.

In Queensland, moderation practice occurs via the system of externally moderated standards-based assessment in senior schooling[2] and the Queensland Curriculum Assessment and Reform initiative,[3] which has recently attempted standards-referenced moderation in Years 1–9. Although the details of these two approaches necessarily differ, a common feature is the understanding that system-level support can ensure teachers reach judgements with high validity and high reliability levels.

This is not to suggest that the function of moderation should be narrowly understood as serving accountability alone. Indeed, as evidenced by the recent uptake of moderation practice in England, Scotland and Wales, we are proposing that moderation practices where teachers come together to assess and judge student work against stated standards can have a direct flow on to and benefit for efforts to improve curriculum design and development in the classroom. Specifically, it is in the context of standards-based moderation talk that teachers can explore the meaning and use of standards as they relate to construct validity, achieving clarity of expectations for themselves and their students in relation to task design. Further, moderation can function as the main means through which teachers reach agreement on the qualities of the learning being assessed.

Fourth condition: The assessment community

Our fourth condition relates to the need to recognise influences on judgement and how these influences depend on assessment purposes. For example, there are assessment purposes that involve judgement for accountability, and reporting of student achievement more specifically. These can be readily distinguished from assessment and judgements that have as their primary aim the improvement of learning. For example, a teacher might adjust the curriculum and related assessments for students with learning difficulties, providing opportunities for those students to achieve goals that are realistically attainable for them, though below the goals of other students. Such adjustments permit learning and assessment to be customised to the interests and needs of individual students. As such, they do not necessarily reflect standards expected of students at a given year level. They do, however, serve the purpose of promoting and monitoring learning for individuals who, over time, may well achieve those year-level standards.

A word of caution should therefore be applied to the notion that standards should necessarily have a regulatory influence over teaching and learning. As mentioned above, standards can serve to inform teachers about curricular intent and the demands of assessment tasks relative to that intent. This observation holds most powerfully where teachers are directly involved in using centrally developed standards for judging student work, be this in the form of individual tasks or portfolio collections of work collected over time.

Further, although standards and exemplars together can serve to clarify the desired characteristics of quality, they do not necessarily fully account for the factors that shape teacher judgement. In a three-year large-scale Australian study of teacher judgement in middle schooling, Cooksey, Freebody, and Wyatt-Smith (2007) reported high levels of variability in teachers' notions of quality and also unearthed the range of factors that shape how judgements are reached (Wyatt-Smith, 1999; Wyatt-Smith & Castleton, 2004). Similarly, in a recent study of standards, teacher judgement and moderation, Wyatt-Smith et al. (2010) found that although teachers take account of centrally developed and mandated standards, their judgement acts, as displayed in recorded moderation sessions, go well

beyond the use of standards. Specifically, these researchers identified that although teachers did use stated standards and related textual resources (e.g., sample responses and the *Guide to Making Judgements* issued by the Queensland Studies Authority), they also actively referred to other tacit knowledge (e.g., teachers' personal knowledge of students, knowledge of curriculum and teaching contexts where they have delivered the curriculum, prior evaluative experience and individual tacit knowledge of standards not elsewhere specified) in arriving at judgements. Parts of this second category of resources were often used in combination with—and sometimes in opposition to—the stated standards. They reported how, at times, the other knowledge was used as a reason for discounting or even subverting stated standards.

Given this, it is crucial that guidelines and professional development opportunities be provided to teachers about desired judgement practice and the legitimacy (or otherwise) of the various resources available for teachers to draw upon. This observation is consistent with the results of other studies of judgement and teacher use of standards, some dating back more than a decade (Smith, 1994; Wyatt-Smith & Castleton, 2005).

Research evidence

In seeking to understand the processes and interactions of teachers in moderation meetings, we have found Cook and Brown's (1999, p. 54) concept of bridging epistemologies instructive. These researchers view "the interplay of knowledge and knowing as a potentially generative phenomenon". From this perspective, as teachers meet in their moderation groups to develop and articulate their understanding of the standards—their purpose, function, nature and meaning—there is an interplay of individuals' explicit and tacit knowledge of the standards with the explicit and tacit knowledge of the understanding of the group. For those participating in the moderation group, new knowledge and knowing lies in "the use of knowledge as a tool of knowing within situated interaction with the social and physical world" (Cook & Brown, 1999, p. 54).

In other words, moderation as social practice provides an opportunity for teachers as assessors to develop and articulate their understanding of

the standards as used in the assessment of student work. The moderation meeting itself also provides an opportunity to generate new knowledge and new ways of knowing as teachers draw on their individual tacit and individual explicit knowledge and the group's tacit and explicit knowledge, and use this knowledge as a tool of knowing within a situated interaction with the social and physical world. This is what Cook and Brown (1999) term *knowing in action*, and it leads to the production of new knowledge and knowing about standards in relation to their judgements in the practice of moderation.

Teachers' engagement in moderation practice, and the new knowledge and ways of knowing that are generated, include the following:

- Teachers are able to check that similar skills and levels of skills are taught, and that similar learning outcomes are assessed as equitable and of a comparable quality.
- Fairness for all students is extended beyond the classroom or school to other schools, as teachers focus on quality and how judgements of quality are arrived at.
- There is increased confidence for teachers, parents, students and other staff members in that common standards are expected and being achieved by a particular year group of students.
- Teaching and assessment practices are made transparent in that teachers' work is made public, open to scrutiny and critique, which helps to address accountability and quality assurance demands. Gaps or omissions in the teaching programme can be identified, particularly if the director of curriculum or head of department participates in the moderation meetings.
- A sense of community develops as teachers negotiate their understanding and seek clarification and advice when they are unsure of the standard or the standard of work. There is a shift from individual practice to shared practice and the improvement of that practice.
- Engaging in moderation focuses teachers' attention on assessment and its place in the teaching and learning programme. Teachers seem motivated to teach a topic when they realise the results achieved by other teachers using different approaches. In this way, teachers learn new ways of teaching and are diversifying their practice to meet the needs of individuals, and are also possibly improving their practice

(Klenowski & Adie, 2009).

In this context, standards are intended to be used as the basis for judgements of student achievement, while the results from assessment tasks are meant to both inform the teaching/learning process and report and track student progress. In such a system, the role and reliability of teacher judgement take centre stage. Further, recent work on efforts to support students at educational risk suggests that they can be supported to achieve at higher levels when information about standards and the expected features of quality are central to how learning and teaching occur in the classroom (Wyatt-Smith & Bridges, 2008).

With the Australian Federal Government's accountability and transparency agenda based on failed overseas models (Reid, 2009), political constraints have emerged for such preferred practice of standards, judgement and moderation. The push to publish the performance of individual schools and to target schools that are underperforming has the potential for the mistakes experienced in England (Broadfoot, 2007; Stobart, 2008) and by the No Child Left Behind policy in the USA to be repeated in Australia. There is the danger that performance pay will be introduced, and where principals fail to demonstrate improvement over time they could be replaced or the school could be closed. As Reid (2009) has argued, this approach to accountability does not address issues of equity and fails to recognise, trust and include the professional judgement of teachers.

Emerging political constraints to preferred educational practice

At this point it might be helpful to look at some background to the current situation in Australia and the context for these emerging political constraints. In 2007 the six states and two territories developed individual approaches to the use of grades and standards in the implementation of curriculum, assessment and reporting. This occurred because the then federal Liberal Minister of Education (Dr Brendon Nelson) threatened to withhold funds from the states and territories unless they agreed to implement A–E reporting and benchmark testing for literacy and numeracy, as well as several other "disparate curriculum initiatives" (Reid, 2009, p. 2). While there was a requirement to report student

achievement across the country using A–E standards, the standards were not defined. Therefore states and territories formulated different versions of these standards. Such an approach is in contrast to what occurred in England, where the Task Group on Assessment and Testing (TGAT) was set up in 1986 to advise the Government on assessment and testing in the national curriculum. The TGAT recommended progressive levels so that even those students who were making limited progress would still be able to be recognised as making some progress in terms of the levels. The grades system was rejected because of concerns related to the damage to students and schools of diminished learner identity for the student who continues to be awarded an E grade throughout their years of schooling.

In Australia, the establishment of the National Curriculum Board (NCB) (www.ncb.org.au) in February 2008 led to plans to develop the "core content and achievement standards" in mathematics, science, history and English, from kindergarten to Year 12, with a national curriculum to be available in 2010 and to be extended in a second phase to include languages and geography in 2011. In May 2009 the newly established Australian Curriculum, Assessment and Reporting Authority (ACARA), an independent statutory authority, took over the work of the NCB. ACARA now manages the implementation of the national curriculum, national student assessment and reporting of school education outcomes. To "invigorate a national effort to improve student learning in the selected subjects" (National Curriculum Board, 2008, p. 3), a standards-referenced framework was developed. However, what has been missing is any identification of the assessment evidence that was used to inform the development of the achievement standards. There has been very little public information made available about how the achievement standards were developed, who was involved in their development and how they are to be used in practice. More than this, the fit across the three elements—standards, the assessment evidence to which they relate and the curriculum—has not been looked at as part of the development work.

Despite this notable omission, the achievement standards are put forward as providing:

> an expectation of the quality of learning that students should typically demonstrate by a particular point in their schooling (that is, the depth of

> their understanding, the extent of their knowledge and the sophistication of their skills). (ACARA, 2009, p. 20)

The achievement standards for kindergarten to Year 10 are represented at every year of schooling by a statement of learning typically expected for the year, a set of generic grade descriptors and a set of work samples illustrating the quality of expected learning. The use of annotated student work samples aims to illustrate the differences in the quality of student work.

The purposes of the "achievement standards" then includes, first, to make clear the expected quality of learning (knowledge, understanding and skills) to be achieved; second, to provide language with which teachers can discuss with students and their parents the student's current achievement level, progress to date and what should come next; and third, to help identify students whose rate of progress puts them at risk of being unable to reach satisfactory achievement levels in later years (National Curriculum Board, 2008). These standards are intended to fulfil the purpose of improving student learning and accountability.

At this point we revisit the competing uses of the term *standard* in the current political-educational contexts in Australia. In the following example, the term *standard* is used in the context of the National Assessment Program Literacy and Numeracy (NAPLAN) testing and fulfils a particular role in this context:

> For each year level a *national minimum standard* is located on the scale. For Year 3 Band 2 is the national minimum standard, for Year 5 Band 4 is the national minimum standard, for Year 7 Band 5 is the national minimum standard and for Year 9 Band 6 is the national minimum standard. The skills that students are typically required to demonstrate for the minimum standard at each year level are described on the back page of the student report.
>
> These *standards* represent increasingly challenging skills and require higher scores on the national scale. (Ministerial Council on Education, Employment, Training and Youth Affairs, 2009, our emphasis)

In Queensland, the State Government is keen to raise standards as represented by the results of NAPLAN testing, as evident in 2009 when

the Queensland Premier advised schools to sit practice NAPLAN tests in Years 3, 5, 7 and 9 because she was disappointed by the overall results of the 2008 tests, which she indicated were designed to assess whether students were meeting "national *standards* in numeracy, reading, writing, spelling, punctuation and grammar" (Bligh, 2009, our emphasis). At the time, nationally, there were no officially sanctioned statements about the expected learning of literacy and numeracy at particular year levels.[4] Further, there were no published standards to inform teachers about the expectations of quality, except those produced after the testing is complete. For parents, there are summary statements of the skills assessed to inform them about their child's report. In the case of NAPLAN, the standard given primacy is referred to as the minimum or benchmark standard, taken to represent a level below which a student is at educational risk. This being the case, it is somewhat surprising that the benchmark is identified after the test data are available. This means that it is developed to "fit" the data and is not tied directly to statements of expected quality, communicated to stakeholders in advance of the testing.

Distinctions such as these in the uses of the term *standards* need to be made explicit. In the first example, the term is used in the context of ACARA to mean achievement standards. The notion of a standard in this context is as a measure or yardstick for judging achievement. In the second example, the term is used in reference to national minimum standards, and Queensland's Premier's response to the NAPLAN testing programme highlights how the meaning of the term *standard* differs in that it is used as a level of attainment or point of reference as measured by some yardstick (in this case, band levels on a scale). The concern for teachers is that by emphasising that the NAPLAN test is the measure or reference point, teachers will react by narrowing their focus to what is tested or measured. In other words, in the absence of prescribed indicators of quality, it is reasonable to expect that teachers will emphasise in their teaching whatever has been specified in the test (as distinct from the curriculum).

As is evident from the Queensland Government's response to the NAPLAN results, governments are increasingly anxious about education standards, particularly as reflected in national or international comparisons of student achievement (Goldstein & Heath, 2000; Stobart, 2008). In such an intense education policy environment, standards can assume an importance that

is detached from the evidence base from which they derive their meaning (Darling-Hammond, 2004). This is despite the fact that the meaning of standards is always and inevitably tied to the nature of the assessment to be judged, the conditions and contexts in which the evidence was collected and, more specifically, the judgement practices applied to arriving at decisions of quality. In short, standards, evidence, assessment conditions and judgement practices are necessarily tied together. This applies irrespective of how standards are represented (as alphabetic grades or numeric scores).

This observation is timely given the observable shifts in national reporting practices discussed. In the case of Australia, for example, there is a stated government position on ensuring transparency in the reporting of student achievement to parents and the wider community. This includes the reporting of school performance data on the Web to help inform parents' school selection. However, such a principled position does not in itself ensure transparency and informed interpretation of test data. For example, although currently mandated national tests of literacy and numeracy present point-in-time literacy and numeracy data, they do not represent learning in the curriculum more generally. Similarly, the mandated use of A to E grading for reports does not ensure that teachers are receiving reports that can claim high levels of validity and reliability. These examples illustrate the need to ensure that the evidence base from which standards derive their proper meaning is presented, if misinterpretation of reported standards is to be avoided.

Conclusion

The aim of the article is to bring to centre stage the role of standards in teachers' work and in meeting accountability demands. It highlights how the term *standards* can be used in various ways and calls for clarity around the meaning of the term in policy and curriculum contexts**.** It is timely to consider the importance of understanding the roles and purposes of standards and the conditions needed for their effective use by teachers to measure improvement and support learning. This article takes the position that teachers' use of standards is fundamental in large-scale efforts to improve learning and achieve consistency for reporting.

Although there has been considerable investment of energy in discussions and forums about the national curriculum in Australia, there has been

a striking silence about how assessment of student achievement in the national curriculum will occur. Also striking has been the limited public attention given to the standards that might apply for gauging student achievement in the new curriculum. Yet, at another level, this is perhaps consistent with other curriculum policy initiatives where assessment remains unaddressed until after curricular decisions are taken. This approach continues the long and unhelpful tradition of separating curriculum and related teaching and learning activities from assessment.

The overview of the conditions presented in this article points clearly to the need to build the capability of the workforce if educational assessment policy is to succeed in getting the profession to realise the potential of standards to inform teacher judgement and, in turn, improve student learning and outcomes. As a corollary, we suggest that improvement will not come from curriculum reform in and of itself, and that it is timely to review the role of teachers as the primary assessors of student learning. In this role, we suggest that national prospects for achieving improved learning, and indeed, greater equity opportunities in schooling, should be directly tied to efforts to achieve improved assessment literacy on the part of policy makers, teachers, principals and educators in general. In part, this can be achieved through preservice and inservice development, with a focus on quality assessment practices, including the use of standards and evidence from case studies of informed practice. It could also be achieved through a greater balance in the policy direction to promote the improvement function of standards, a focus that can often be lost in the intense policy interest in standards for reporting purposes alone.

References

ACARA (Australian Curriculum, Assessment and Reporting Authority). (2009). *Curriculum design paper*. Retrieved 16 June 2009, from http://www.acara.edu.au/verve/_resources/Curriculum_Design_Paper_.pdf

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

Broadfoot, P. (2007). *An introduction to assessment*. New York: Continuum.

Cook, S. D. N., & Brown, J. S. (1999). Bridging epistemologies: The generative dance between organizational knowledge and organizational knowing. *Organization Science, 10*(4), 381–400. (Edited version).

Cooksey, R., Freebody, P., & Wyatt-Smith, C. M. (2007). Writing. *Educational Research and Evaluation*, *13*(5), 401–434.

Cresswell, M. (2000). The role of public examinations in defining and monitoring standards. In H. Goldstein & A. Heath (Eds.), *Educational standards* (pp. 69–104). Oxford: Oxford University Press.

Crooks, T., Darr, C., Gilmore, A., Hall, C., Hattie, J., Smith, J., et al. (2009). *Towards defining, assessing and reporting against national standards for literacy and numeracy in New Zealand*. Christchurch: The New Zealand Assessment Academy, University of Canterbury.

Darling-Hammond, L. (2004). Standards, accountability and school reform. *Teachers College Record, 106*(6), 1047–1085.

Department for Education, Lifelong Learning and Skills (2007) *Future assessment arrangements for Key Stages 2 and 3 – Report on the findings and outcomes of the national consultation, held 31 October 2006 to 12 January 2007*. Retrieved 15 November 2009, from http://wales.gov.uk/consultations/education/1463776/?lang=en

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

Goldstein, H., & Heath, A. (Eds.). (2000). *Educational standards*. *Proceedings of the British Academy 102*. Oxford: Oxford University Press.

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

Hyslop, F. (2009). *Assessment for Curriculum for Excellence*. Retrieved 15 November 2009, from http://www.ltscotland.org.uk/Images/AssessmentforCfE_tcm4-565505.pdf

Klenowski, V. (2006). *Evaluation report of the pilot of the 2005 Queensland Assessment Task (QAT)*. Brisbane: Queensland Studies Authority. Retrieved from http://www.qsa.qld.edu.au/research/reports.html

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

Klenowski, V. (2008, November). *The changing demands of assessment policy: Sustaining confidence in teacher assessment*. Paper presented at the Australian Association of Research in Education conference, Brisbane.

Klenowski, V., & Adie, L. E. (2009). Moderation as judgement practice: Reconciling system level accountability and local level practice. *Curriculum Perspectives, 29*(1), 10–28.

Köller, O. (2009, November). *Using computer-based assessment to test listening and visual comprehension in large-scale assessments of foreign languages*. Keynote presentation at the 10th annual Association for Educational Assessment—Europe conference, Malta.

Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6*(1), 83–102.

Mansell, W., James, M., & Assessment Reform Group. (2009). *Assessment in schools: Fit for purpose? A commentary by the Teaching and Learning Research Programme*. London: Economic and Social Research Council, Teaching and Learning Programme.

Marsh, C. (2009). *Key concepts for understanding curriculum*. London: Routledge.

Maxwell, G. S. (2001). *Are core learning outcomes standards?* Brisbane: Queensland Studies Authority.

Maxwell, G. S. (2002). *Are core learning outcomes standards?* Brisbane: Queensland Studies Authority. Retrieved 23 November 2008, from www.qsa.qld.edu.au/downloads/publications/research_qscc_assess_report_1.pdf

Maxwell, G. S. (2007). *Implications for moderation of proposed changes to senior secondary school syllabuses*. Brisbane: Queensland Studies Authority.

Maxwell, G. S. (2008, September). *Setting standards: Fitting form to function*. Paper presented at the 34th International Assocation for Educational Assessment annual conference, Cambridge.

Maxwell, G. S. (2009). Defining standards for the 21st century. In C. M. Wyatt-Smith & J. J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 263–286). Dordrecht: The Netherlands. Springer International.

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

Ministerial Council on Education, Employment, Training and Youth Affairs. (2009). *National Assessment Program Literacy and Numeracy*. Retrieved 12 June 2009, from http://www.naplan.edu.au/frequently_asked_questions/frequently_asked_questions.html

Mortimore, P. (2008). Testing times. *Ed. Lines, 7*(4), 6–7.

National Curriculum Board. (2008). *The shape of the national curriculum: A proposal for discussion*. Retrieved 5 November 2008, from http://www.ncb.org.au/our_work/preparing_for_2009.html

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common core state standards initiative*. Retrieved 11 March 2010, from http://www.corestandards.org

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Price, M., O'Donovan, B., Rust, C., & Carroll, J. (2008). Assessment standards: A manifesto for change. *The Brookes eJournal of Learning and Teaching, 2*(3), 1–8.

Queensland Studies Authority. (2007). *Information statement April 2007: Standards draft 2*. Brisbane: Author.

Reid, A. (2009). Is this a revolution?: A critical analysis of the Rudd government's national education agenda. *Curriculum Perspectives, 29*(3), 1–13.

Sacks, P. (1999). *Standardized minds*. Cambridge, MA: Perseus Books.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, *13*(2), 191–209.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy & Practice, 5*(1), 77–84.

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment and Evaluation in Higher Education, 30*(2), 175–194.

Stanley, G., MacCann, R., Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of teacher assessment: Evidence of what works best and issues for development*. Oxford: Oxford University Centre for Educational Assessment, Qualifications and Curriculum Authority.

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

Wilson, M. (Ed.). (2004). *Towards coherence between classroom assessment and accountability*. Chicago: University of Chicago Press.

Wyatt-Smith, C., & Castleton, G. (2004). Factors affecting writing achievement: Mapping teacher beliefs. *English in Education, 38*(1), 37–61.

Wyatt-Smith, C., & Castleton, G. (2005). Examining how teachers judge student writing: An Australian case study. *Journal of Curriculum Studies, 37*(2), 131–154.

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

Wyatt-Smith, C. M. (1999). Reading for assessment: How teachers ascribe meaning and value to student writing. *Assessment in Education: Principles, Policy & Practice, 6*(2), 195–224.

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. Retrieved from http://education.qld.gov.au/literacy/docs/deewr-myp-final-report.pdf

Wyatt-Smith, C. M., & Gunn, S. (2009). Towards theorising assessment as critical inquiry. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century: Connecting theory and practice* (pp. 83–102). Dordrecht, The Netherlands: Springer International.

Zepke, N., Leach, L., Brandon, J., Chapman, J., Neutze, G., Rawlins, P., et al. (2005). *Standards-based assessment in the senior secondary school: A research synthesis*. Palmerston North: Massey University College of Education.

Notes

1 No Child Left Behind Act. Reauthorisation of the Elementary and Secondary Education Act. Pub. L. 107-110 2102(4) (2001).
2 For more information, visit the QSA website: http://www.qsa.qld.edu.au/assessment/2130.html
3 For more information, visit the QSA website: http://www.qsa.qld.edu.au/assessment/qcar.html
4 In December 2009 the Queensland Studies Authority and the Queensland Government published information related to the P–9 Literacy and Numeracy Indicators at year levels P–3 and 4–9 to guide teacher and school practice.

Acknowledgements

The authors wish to acknowledge the involvement and support provided by the Australian Research Council and our industry partners, The Queensland Studies Authority (QSA) and the National Council for Curriculum and Assessment (NCCA) of the Republic of Ireland. Other members of the original research team include: L. Adie, P. Colbert, Professor J. Elwood and Dr A. Looney.

Authors

Val Klenowski is Professor of Education in the School of Learning and Professional Studies at the Queensland University of Technology. She has researched curriculum development and assessment practice internationally at all levels, from early childhood through to higher education. She is particularly interested in teachers' classroom assessment practices and their interrelationship with learning, curriculum and pedagogy.

Email: val.klenowski@qut.edu.au

Claire Wyatt-Smith is Professor and Dean of the Faculty of Education at Griffith University. She has been a sole or chief investigator on a significant number of Australian Research Council and government-funded projects over the past decade. These have been primarily in the fields of literacy and assessment, with particular focus on teacher judgement, evaluative frameworks and the literacy–curriculum–assessment interface.

Email: c.wyatt-smith@griffith.edu.au