Setting standards: Fitting form to function

Graham Samuel Maxwell

Educational Consultant

Paper presented at the 34th IAEA Annual Conference
Cambridge UK, September 2008

Abstract

This paper explores some issues concerning the representation and application of standards. The term 'standard' has a variety of meanings, with different consequences for practice. A key distinction is 'content standards' versus 'performance standards'. Another distinction is a 'range of standards' versus a 'targeted or expected standard'. Standards can be represented by cut-scores or ordered categories (or a combination of these). The traditional psychometric approach sees standard setting as an empirical exercise dependent on the assessed cohort performance; the emergent decision-based assessment approach sees standard setting as a judgement process dependent on prior description and example. Also, standards representing comparative performance on a particular task or course can be different from standards representing developmental improvement over time. Clearly, standards need to be represented differently for different purposes—form fitted to function. There is also a need to invent new ways of representing and managing standards that fit a personalised approach to learning.

Types of standards

The term ‘standards’ figures frequently in discourse on educational assessment. However, one of the difficulties in discussing standards is that the term can have many different meanings. These different meanings are not pedantic and benign. Failure to clarify which meaning is intended often results in a patina of agreement or a Babel of confusion. Having some way of identifying different meanings can lead to better communication. Also, it allows us more easily to examine their hidden assumptions and implications. This may lead us to some changes of direction in educational practice as we clarify options and invent new possibilities.

There are at least five types of standards (Maxwell, 2002a; forthcoming 1):

1. Standards as moral or ethical imperatives (what someone should do)
2. Standards as legal or regulatory requirements (what someone must do)
3. Standards as target benchmarks (expected practice or performance)
4. Standards as arbiters of quality (relative success or merit)
5. Standards as milestones (progressive or developmental targets).

Standards as moral or ethical imperatives (type 1) indicate something that is desirable but lacking regulatory force. That is, they offer principles or guidelines. Examples are the assessment for learning guidelines of the UK Assessment Reform Group (2002) and the various standards for school subjects in the USA (for example, NCTM, 2000). These are standards for schools and school systems to adopt in constructing and delivering their curriculum. At the student level, standards as moral or ethical imperatives typically relate to their moral or ethical behaviour. While some of these, such as no cheating or plagiarism may be regulatory requirements, others such as being polite and conscientious are merely encouraged (though rewarded informally).

Standards as legal or regulatory requirements (type 2) involve some form of compulsion. There are consequences for failure to satisfy the requirements. At an institutional level, an example is the ISO Standards (see the Standards for Statistical Methods, ISO, 2008, though this is only one of thousands of such standards). In this case, the standards must be satisfied to gain the ISO imprimatur. At a student level, the (minimum) requirements for being awarded a certificate or degree are often referred to as the standards for gaining the award. Usually, these kinds of standards involve a checklist of all the things to be satisfied for being awarded the certificate.

Standards as target benchmarks (type 3) define an expected or typical outcome (for example, a particular level or quality of performance). These can be requirements for a ‘pass’ or ‘satisfactory’ or ‘sound’ (or in the training sphere ‘competent’). Typically, this goes beyond the checklist typical of the previous two types of standards; what is needed is some representation of the point along a continuum that defines a minimum acceptable level.[1]

The last two types of standards are concerned with differentiated levels or performance: the first in terms of levels of merit or quality and the second in terms of levels of development or progress. Both are expressed as ordered categories that represent the range of possible levels of performance. Typically, a judgment is made about which category best represents each student’s performance—a judgment that pitches the student against the standards and not (at least not directly) against other students; this produces comparative information as a by-product but comparison is not the primary purpose. The two types of standards differ in terms of focus and time frame. Merit standards (type 4) apply to a single assessment event, such as a completed task or a completed course, and are often tailored to that event; they allow a rating

[1] Competence in vocational training programs is often assessed by checklist. However, each item of the checklist requires a judgment of whether an acceptable level of performance has been demonstrated. So, we can just shift the focus: each separate item evidences a target benchmark; collectively they define the regulatory requirement for certification as competent.

of the quality of the performance. Developmental standards (type 5) allow a series of interim judgments against sequential stages of progress over time along a continuum of learning (therefore allowing periodic re-assessment to determine current status).

Content standards versus performance standards

Content standards

A distinction can be drawn between content standards and performance standards. Content standards are what some might call a syllabus of study, detailing and sequencing the content to be learned. In recent years, content standards have figured large in the USA, mainly as a way of giving guidance to schools and teachers about what they should be teaching. As such, they offer moral or ethical imperatives concerning the curriculum (type 1 standards).[2]

Typically, content standards have two important characteristics: first, they provide a mapping of the knowledge and skills a student should acquire in a field (for example, science);[3] second, the subject matter is apportioned to school years (or grades), moving from simpler concepts and skills to more complex ones.[4] Two important factors are ignored: pedagogical considerations about how to help students develop from novices to experts in the subject; and the individuality of each student's learning journey. The assumption is that a single 'road map' and 'travel schedule' will suit all students—that all learners can travel the same route and at the same pace. However, in practice, students do not progress sequentially and inexorably through the subject matter (by the most direct route), all students learn partially (miss bits and misconstrue other bits), and some students get left behind. Content standards therefore have some limitations as a guide for student learning and assessment.

Performance standards

Whereas content standards focus on inputs, performance standards focus on outcomes. Performance standards are about how well something has been done. The 'something' could be, for example, a test, task, portfolio, semester, course or certificate. Their main purpose is essentially retrospective—once something has been done, how well was it done. This is so even if the assessment is also used formatively. In most cases, the 'something' being assessed is non-repeatable, once completed it will not be done again, at least in that form. The student and the assessments move on. What remains is a record of what was done and how well it was done.

Necessarily, there has to be a context within which the performance occurs and from which the performance standards are derived. Even where the performance standards are criterion-referenced rather than norm-referenced, the context includes a reference group of some kind. In the norm-referenced situation, this is usually all those who take the test or produce the performance, though this can be extended by linkage to previous cohorts. In the criterion-referenced situation, it is usually a notional target group to which the standards are referenced (for example, what is expected of a Year 5 student and a Year 12 student is different even if the same labels are used for the standards). If any standard is potentially unachievable by someone in the notional target group then that standard is irrelevant; this helps define the range of useful standards for that context (Sadler, 1987).

---

[2] There is much facile thinking about such documents, both the presumption that because something is listed it will be taught and learned, and also the presumption that because something is omitted it will not be taught or learned. The taught curriculum and the experienced curriculum are both dynamic constructions to which any official document is only one input (albeit an important one).

[3] See, for example, Science standards by the National Research Council (1996), English, Mathematics and Science standards by New Standards (1997) and Mathematics standards by the National Council of Teachers of Mathematics (2000).

[4] Content standards can take other forms, such as lists of 'essential learnings'. These tend to be less elaborate and deal with concepts rather than detailed content. These might be organised into stages (covering several years) rather than single years. See, for example, Queensland Studies Authority <http://www.qsa.qld.edu.au/assessment/3160.html>.

Sometimes a single standard suffices. A single standard could be a competency standard (such as is common in vocational training programs) or a passing or satisfactory standard (such as for certification or for progression from one stage of a program to the next). Such a standard is a target benchmark, required outcome or expected performance (type 3 standard). Performance of insufficient quality to satisfy the benchmark is considered (and often labelled) as failing or unsatisfactory or inadequate. Sometimes there are sanctions (such as denial of a certificate or denial of opportunity to proceed to the next stage of schooling); sometimes sanctions are mitigated by opportunity to try again (or repeat); sometimes there are no sanctions and the assessment is treated merely as indicative (essentially a type 1 standard).

In the latter case, if no action is taken to help the student remediate the assessed deficiency, it is to be wondered that there is much value in this kind of assessment. Incitement to 'try harder' might work for some but for others the challenge can be daunting. Repeated failure can have serious and long-term consequences in terms of self-image. We ought to think carefully about labels that have no apparent positive benefits.[5]

How are performance standards determined? There are basically two approaches and these can be called the psychometric approach and the decision-making (or interpretive or hermeneutic) approach. The key characteristic of the psychometric approach is the cut-score, a point on a continuous scale that defines the minimum score required for achieving a particular standard. The key characteristic of the decision-making approach is the holistic rating by an assessor of the student performance against performance levels. Typically these are ordered categories rather than scaled quantities, though they are sometimes subsequently mapped onto a numerical scale, in which case the decision about which category (level) applies comes first and mapping onto a numerical scale follows, not the other way around.

In the psychometric approach, descriptions of the standards categories can be either an input or an output. As an input they guide an expert standard-setting team in determining appropriate cut-scores (for example, using Angoff processes, see Cisek, 2001). As an output they describe typical performance within each category (or alternatively at the cut-score) after dividing the continuum into categorical levels (by quotas or by intuitive judgment).[6]

In the decision-making approach, the standards categories are seen as conceptual categories that are supported and explained by the descriptions (more appropriately, therefore, referred to as 'descriptors'). In this sense, the descriptors are not the standards themselves (which are conceptual constructs) but merely representations of those standards. The descriptors serve to explain or illustrate the standard and also to make assessor judgment more objective and accountable, by providing concepts to guide the judgment and language to justify it (particularly to students). The descriptors can also be used formatively to help students target their learning and to acquire understandings of quality performance. Descriptors are usefully supplemented by exemplars, that is, examples of performances for each standard (Sadler, 1987).[7] Common understanding of the standards among assessors requires moderation processes among assessors (Maxwell, 2001; 2002b; forthcoming 2).

A common way of representing performance standards is through a rubric, or in the language of Sadler (1987) a criteria-and-standards matrix. The criteria are the dimensions (rows) and the standards are the levels (columns). Sometimes rubrics are not laid out as a matrix but rather with a single holistic or aggregate statement under each standard; typically, these statements could be disaggregated into matrix form because they string together a set of

[5] It could be argued (see Wiggins, 1998) that any level less than the top standard falls short of what is desired (exemplary). In that case, most students could be considered as failing.

[6] An example of the latter is the set of standards descriptions developed for Higher School Certificate in New South Wales (Bennett, 1998).

[7] Wiggins (1998) uses the term 'exemplar' to refer to an example of the highest standard ('exemplary') and uses 'work samples' to refer to what Sadler (1987) calls exemplars.

statements that each reference one criterion. Reading across a row in the matrix, we should be able to recognise successive increases in quality from one level to the next.[8]

How explicit should a rubric be? That depends on the circumstances of its use. The point is to provide sufficient detail about the desired performance characteristics to be able to make a consistent judgment about which level best characterises the performance.[9] Necessarily, levels are broad categories and therefore somewhat fuzzy and imprecise. Yet, any degree of explicitness about the nature of performances typical of each standard sharpens the focus, fosters consistency and improves communication. Sometimes finer distinctions are made, for example, high, middle and low within each level, though usually without specific descriptors.

A criteria-and-standards matrix produces a performance profile: a specific level on each criterion. Sometimes, an overall level is reported, combining the performance across all criteria. There are two ways this can be done, by aggregating scores or by judgment. With score aggregation, a further decision is needed about cut-scores for the overall standards. With judgment, where performance differs across criteria, a best-fit judgment is required that allows trade-offs across the criteria. In both cases, the meaning of the overall standard is rather ambiguous, at least in the middle categories, because the trade-offs differ across students and the aggregate grade descriptions only characterise typical performance. This may be adequate for certification, where usually only the standards labels (e.g., A–E) are reported. However, for feedback (formative) purposes, the detailed profile of performance on separate dimensions is essential and the overall performance level is too vague and general to be useful.

How many criteria should a rubric have? This can be approached epistemologically, that is, through consideration of the inherent dimensions of the subject matter and/or the nature of the task (or portfolio, etc.) being assessed. However, there are pragmatic considerations too. It is difficult to keep very many characteristics in mind at the same time (Miller, 1956). For more than five criteria it is best to develop a hierarchical structure (sub-criteria embedded within main criteria) but the more the detail the more the cognitive demand anyway.

Despite the apparent benefits (and increasing popularity) of rubrics,[10] there are some difficulties. Some of these have already been mentioned: they can only be interpreted in relation to a specific context; standards descriptors are necessarily fuzzy and imprecise; aggregate standards are ambiguous; lower standards tend to signify failure or deficiency; standards descriptors, even when accompanied by exemplars, are insufficient to ensure common interpretation and usage, which requires training and moderation.

There are some additional difficulties. First, there is a tension between generic and specific descriptors. Generic descriptors maintain consistency of language across different contexts (including different years); this creates interpretive difficulties, with the labels and descriptors referring to different observable features of performance (for example, ‘excellent’ refers to quite different performance in Year 5 and Year 12, or at the beginning and end of a course). Specific descriptors are tailored to the specific assessment event; this clarifies their meaning for that context but makes them one-off wonders.

Second, descriptors are often tautological, that is, they merely repeat the qualitative language of the standards labels (for example, limited, sound, high); alternatively, they are often vaguely quantitative (for example, few, some, many, all; or moderately, generally, very).

---

[8] Wiggins (1998) distinguishes between holistic and trait-analytic rubrics. The former correspond with what is called here an aggregated statement (what Wiggins calls one general descriptor). The latter leads to a criteria-and-standards matrix but Wiggins calls each row, rather than the whole matrix, a rubric.

[9] Where strong traditions exist, it is possible to dispense with descriptors and depend on the common tacit knowledge and experience of the assessors as connoisseurs (Sadler, 1987).

[10] Examples of rubrics are plentiful. One website <http://www.rcampus.com/indexrubric.cfm> provides a tool for developing rubrics and claims to have some 30 000 ‘ready to use’ rubrics.

Some degree of generality is needed since some of the characteristics of the performance are (hopefully) unpredictable. But many descriptors add little to the labels.

Third, preferably, higher standards mean more elaborate (qualitatively different) knowledge, understanding and production, not just a bit more of the same. Yet the criteria-and-matrix layout encourages tautological and quantitative language. More elaboration should mean bringing in *more criteria* at higher levels as the performance acquires more breadth and depth and sophistication (and therefore additional dimensions).

Fourth, are these kinds of standards anything more than arbitrarily pragmatic? While criteria can conceivably be based on theories of knowledge and theories of learning, performance standards seem to lack any such theories. Rather, they seem to be driven by practical considerations derived from the way education is packaged (into years of schooling or courses) and the consequent desire to compare and discriminate the different learning outcomes this necessarily produces in a student cohort. A revisiting of the theories of John Carroll in the 1960s (Carroll, 1963; 1989) which suggested that time be allowed to vary in pursuit of learning goals, would seem worthwhile.

Merit standards versus developmental standards

Performance standards of the kind just described are merit standards (type 4). They differentiate levels of quality on something assessable at a particular point in time. There is a comparative edge to this assessment since the levels are expected to span the range of possible performance in the target group of students (usually a year or course cohort). There are several situations where this may be useful, for example, where the assessments are used for selection or for system monitoring rather than individual reporting. It may (perhaps) be useful also where the assessments are used formatively, but typically only with further specific elaboration, to help students appreciate the difference between their own performance and better performance and therefore where to put their effort for future improvement. However, where the assessments are purely summative (reportative) *and no action follows*, the effects may be more damaging than beneficial.[11] This applies as much to end-of-semester reporting as anything else. Repeated reporting (over 24 semesters) of the same grade can stereotype students in their own and others' perception. For some, it is the great turnoff and image destroyer. If the purpose of education is personal development and advancement, this is not an appropriate outcome.

Developmental standards (type 5) offer a different approach to differentiated standards where we zoom out to see the current assessment in the context of a longer trajectory of learning. In other words, we depict the current assessment in relation to progress towards a longer-term goal. Developmental standards provide a series of levels, steps or stages that map progression from novice to expert performance. Whereas with merit standards, the goal posts are continually shifting as the same labels are applied to new situations, the goal posts for developmental standards are fixed. Progress over time can be tracked against a constant scale (analogous to measuring increasing height during childhood and adolescence).

Some examples include: the progress levels for the national curriculum in England; the Queensland Studies Authority (QSA) writing scale for Years 3, 5 and 7 testing; and the standards and progression points for the Victorian Essential Learning Standards (VELS).

The national curriculum in England

Although the term 'standards' is not explicitly used, in England each national curriculum subject charts progress across nine levels (1–8 plus exceptional) along several attainment targets (strands). The levels are represented through paragraph-length level descriptions (LDs) that summarise the characteristics of performance typical of each level. Progress against the levels is assessed at the end of each key stage (Stage 1: Year 2, Age 7; Stage 2: Year 6, Age

[11] Some systems respond to imputed failure by forcing students to repeat a year. This rarely leads to improvement and can lead to worse performance (Heubert & Hauser, 1999; Shepard & Smith, 1989).

11; Stage 3: Year 9, Age 14). A holistic on-balance judgment is made of which level best fits each student's performance.

An example is the following LD for level 4 of the writing strand in English:

> Pupils' writing in a range of forms is lively and thoughtful. Ideas are often sustained and developed in interesting ways and organised appropriately for the purpose of the reader. Vocabulary choices are often adventurous and words are used for effect. Pupils are beginning to use grammatically complex sentences, extending meaning. Spelling, including that of polysyllabic words that conform to regular patterns, is generally accurate. Full stops, capital letters and question marks are used correctly, and pupils are beginning to use punctuation within the sentence. Handwriting style is fluent, joined and legible.

The complete attainment targets and levels are found on the Qualifications and Curriculum Authority website <http://curriculum.qca.org.uk>.

There are some challenges with this scheme: the multidimensionality of each statement and the dependence for interpretive meaning on professional understandings (Sainsbury & Sizmur, 1998); and their apparent use only for key stage reporting and the apparent absence of moderation processes to assure consistency (Hall & Harding, 2002). However, there are several potential benefits, including the efficiency of having one set of benchmarks across all years, progress depicted as movement along a continuum; focus on achievable progress rather than fixed ability; and 'natural' differentiation at each age or year level (Green, 2002). Additionally, they provide feed-forward opportunities, that is, higher levels as targets for learning (Sadler, 1989) and the potentially motivating effects on students of experiencing growth and success rather than receiving the same grade year on year (Dweck, 1986).

QSA writing scale for Years 3, 5 and 7 testing

An exemplary single scale was developed by the Queensland Studies Authority for the writing component of the Queensland Years 3, 5 and 7: Literacy and Numeracy Tests <http://www.qsa.qld.edu.au/downloads/assessment/3579_handbook_reporting_07.pdf>.[12] This had four dimensions (Contextual Factors; Structure; Grammar, Vocabulary, Comprehension and Punctuation; Spelling) and twelve levels (O, N, A–J). For manageability, the scale was divided into three overlapping sections (one section for each year level) but in exceptional cases students might perform outside those sections. Three mid-level standards for Contextual Factors were:

> **F. Planned response that uses and elaborates on ideas from the stimulus to meet task demands:** begins to elaborate the subject matter to connect and explain subject matter; may have a personal organisation; some lapses in the development of the thinking pattern.
>
> **E. A planned response to the task with an awareness of the reader:** responds to similarities and differences in demand as a top-level thinking pattern; begins to explain subject matter,; lacks connections between ideas; may use an informal, chatty voice often modelled in junior texts — *I'm going to tell you.*
>
> **D. Response to the task shows some planning and sequencing:** identifies particular pieces of knowledge which they tell randomly; recognises the task demand to explain similarities and differences of single concepts.

Standards and progression points for VELS

The Victorian Essential Learnings Standards (VELS) builds on and incorporates the previous Curriculum Standards Framework (CSF) <http://vels.cvaa.vic.edu.au/>.

VELS has three strands (Physical, Personal and Social Learning; Discipline-based Learning; and Interdisciplinary Learning), interrelated through what is characterized as a triple-helix. Each strand has several domains, which are split further into several dimensions. For each domain there is a table of 'standards and progression points' that describes six developmental levels over the eleven years of compulsory schooling together with three progression points

[12] This scale is unfortunately no longer in use because national testing has replaced state testing.

between each level (that is, 24 categories overall). The levels represent typical progress at two-year intervals from Preparatory to Year 10.

VELS uses term 'standards' in three different ways: content standards—the knowledge and skills expected to be taught in each of the strands; developmental standards—the levels and progression points for assessing progress; and expected standards—the typical or targeted level for each year level. As a further complication, the Australian Government now requires all schools to report student performance to parents each semester on an A–E scale (Commonwealth of Australia, 2005). Under VELS, Victoria maintains an expectation that schools will continue to assess the standard (level) and progression point reached by each student, with computerised conversion to an A–E grade appropriate for each year level (and representation of the levels in terms of their year of typical attainment). These characteristics of VELS are both visionary and realistic, adhering to the benefits of charting student progress developmentally but acceding to governmental and parental expectations of merit grading within year cohorts. Whether this will be successful or confusing remains to be seen.[13]

In general, there are some clear benefits in using developmental standards:

- They provide explicit steps and targets for developmental progress.
- There is a language and expectation of progress.
- They make evident to students the progress they have made.
- They provide clear targets for further learning
- Student spurts and plateaus can be seen as natural and expected.

There are a couple of caveats. First, as for content standards, developmental levels descriptors depict typical performance but will not fit every student. Second, as for merit standards, levels can be holistic (cover several dimensions) with similar problems of best fit (tradeoffs) and imprecise meaning. Third, how slower progress is handled will affect student self-perceptions. There is a clear need for flexibility in using developmental standards.

There are also some challenges for developmental standards: how to promote acceptance of a new and different framework for reporting progress that breaks with traditional concepts of grading ; how to combine developmental levels with expected levels without reverting to a language of failure; how to talk about slower progress without creating negative self-perceptions; and how to develop school structures to support developmental progression.

Conclusion

This paper has explored some different meanings of the term 'standards' in educational assessment. It indicates a variety of ways in which we currently talk about and represent standards, each serving a different purpose and having different strengths and limitations. We should not confuse one meaning and purpose with another. We can reduce confusion and improve communication by being clear about the type of standard to which we are referring. This is a matter of fitting form to function. Rather than attempt to shoehorn one type of standard into all situations, that is, assume that 'one size fits all', we should recognise the strengths and limitations of each type of standard and tailor our practice accordingly.

However, that is not the end of the story. The analysis in this paper also suggests that there are some critical issues to address in relation to the way we talk about and frame educational standards. In particular, standards of any kind assume a 'typical student' (to set the pace and the expectations) and a 'typical range of students' (to represent different degrees of coping with the standard pace and expectations). The consequence is that we force-fit students to

[13] Referents for A–E in Victoria are defined relative to the expected level for each year: well above, above, at, below, well below. Other Australian states and territories have adopted similar generic descriptors (for example, excellent, good, satisfactory, limited and poor) that offer crude comparative indicators (almost certainly inconsistently applied by different teachers and schools) but convey no information about what the student actually knows or can do. This may be sufficient for some purposes but not others.

essentially arbitrary expectations and ranges. All students are expected to progress linearly and at the same rate through a common curriculum. And yet they don't. So we adjust to that by having merit standards that allow some students to do well and other students to fail. Developmental standards may rescue us (and students) to some extent from this assembly-line thinking but it is still a 'monolithic batch system' (Christensen, Horn & Johnson, 2008) and there are many casualties.

Increasingly, what seems to be needed are systems of *personalised (or customised) learning*. Personalised learning has already captured considerable interest around the world as a key concept in future delivery of educational services (see Keamy, Nicholas, Mahar & Herrick, 2007; OECD, 2006). But to realise this properly requires that we go beyond requiring all students to fit the same mold and instead introduce flexibility and adaptability. We have hardly begun to think about we might do that.

Performance and developmental standards can serve adequately for certification and accountability but not for personalising learning. Placing greater emphasis on the personal advancement of students against targets that are tailored to their circumstances, needs, interests and stage of development means attending more carefully and deliberately to the detail of each student's learning. This would be fitting a new form to a new function.

References

Assessment Reform Group (2002). *Assessment for learning: Ten principles*. <http://k1.ioe.ac.uk/tlrp/arg/CIE3.pdf> (accessed 6 July 2008)

Bennett, J. (1998). *Setting standards and applying them across different administrations of large-scale, high-stakes, curriculum-based public examinations*. Sydney: New South Wales Board of Studies. <www.boardofstudies.nsw.edu.au/archives/occasional_papers/occasionalp1_assess.htm> (accessed 18 June 2008)

Carroll, J. B. (1963). A model of school learning, *Teachers College Record*, 64 (8), 723–723.

Carroll, J. B. (1989). The Carroll model: A twenty-five year retrospective and prospective view, *Educational Researcher,* 18 (1), 26–31.

Christensen, C. M., Horn, M. B. & Johnson, C. W. (2008). *Disrupting class: How disruptive innovation will change the way the world learns*. New York: McGraw Hill.

Cizek, G. J. (ed.) (2001) *Setting performance standards: Concepts, methods and perspectives*. Mahwah, New Jersey: Lawrence Erlbaum.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist* (special issue: *psychological science and education*), 41 (10), 1040–1048.

Green, S. (2002). *Criterion referenced assessment as a guide to learning: The importance of progression and reliability*. A paper presented at the Association for the Study of Evaluation in Education in South Africa International Conference, Johannesburg. Hall, K. & Harding, A. (2002). Level descriptions and teacher assessments in England: Towards a community of practice, *Education Research*, 44 (1), 1–16.

Heubert, J. P. & Hauser, R. M. (eds) (1999). *High stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academies Press.

International Organization for Standardization (ISO) (2008). Standards for Statistical Methods. Geneva: Switzerland. <http://www.iso.org/iso/> (accessed 6 July 2008)

Keamy, R. K., Nicholas, H., Mahar, S. & Herrick, C. (2007). *Personalising education: From research to policy and practice*. Melbourne, Victoria: Department of Education and Early Childhood Development. <http://www.eduweb.vic.gov.au/edulibrary/public/publ/research/publ/personalising-education-report.pdf> (accessed 24 June 2008)

Maxwell, G. S. (2001). *Moderation of assessments in vocational education and training*. Brisbane: Queensland Department of Employment and Training. <http://www.trainandemploy.qld.gov.au/resources/about_us/pdf/moderation_report.pdf> (accessed 24 June 2008).

Maxwell, G. S. (2002a). *Are core learning outcomes standards?* Brisbane: Queensland Studies Authority (now Queensland Studies Authority). <http://www.qsa.qld.edu.au/downloads/publications/research_qscc_assess_report_1.pdf> (accessed 10 Jun 2008)

Maxwell, G. S. (2002b). *Moderation of teacher judgments in student assessment*. Brisbane: Queensland School Curriculum Council (now Queensland Studies Authority). <http://www.qsa.qld.edu.au/downloads/publications/research_qscc_assess_report_2.pdf> accessed 24 June 2008

Maxwell, G. S. (forthcoming 1). Defining standards for the 21st century. In C. Wyatt-Smith & J. J. Cumming (eds), *Assessment issues of the 21st century*. Springer.

Maxwell, G. S. (forthcoming 2). Moderation of student assessments by teachers. In B. McGaw, E. Baker & P. P. Petersen (eds), *International Encyclopedia of Education*. Oxford: Elsevier.

Miller, G. A. (1956). The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review*, 63, 81–97.

National Council for Teachers of Mathematics (NCTM) (2000). *Principles and standards for school mathematics*. Reston, Virginia: NCTM.

National Research Council (NRC) (1996). *National Science Education Standards*. Washington, DC: National Academy Press. <http://www.nap.edu/html/nses/> (accessed 19 June 2008).

New Standards (1997). *New Standards Performance Standards: Elementary, middle school and high school*. Rochester. Washington, DC: National Center on Education and the Economy.

Organisation for Economic Co-operation and Development (OECD) (2006). *Personalising education*. Paris: OECD.

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

Sainsbury, M. & Sizmur, S. (1998). Level descriptions in the National Curriculum: What kind of criterion referencing is this? *Oxford Review of Education*, 24 (2), 181–93.

Shepard, L. A. & Smith, M. L. (eds) (1989). *Flunking grades: Research and policies on retention*. London: The Falmer Press.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve performance*. San Francisco: Jossey-Bass.