Research and Analysis

Grading Competence-Based Assessments

Notes from a Small Literature

Paul E. Newton from Ofqual's Strategy, Risk and Research directorate

Acknowledgements

I am very grateful to Shelley Gillis and Leesa Wheelahan for very helpful conversations related to the implementation of Competence-Based Assessment in Australia. I would also like to thank Michelle Meadows, Phil Carr, and colleagues from Ofqual’s Standards Advisory Group, for their helpful comments on this report.

Contents

Executive Summary

Although England has a tradition of grading in Technical and Vocational Education and Training (TVET) contexts – albeit one that has waxed and waned over time – there is no identifiable literature, from the UK, on this topic. To develop a deeper understanding of this area, attention therefore needed to be focused further afield. The present report explores insights from a small literature, derived mainly from Australia, into the theory and practice of grading Competence-Based Assessments (CBAs). CBA reflects a particular approach to designing Vocational and Technical Qualifications (VTQs), which has become increasingly widespread since the early 1990s. Many regulated VTQs in England bear the hallmarks of CBA.

As originally conceived, the point of CBA was to certify that a learner had reached the appropriate level of competence to practise in a professional or occupational domain. This led to competence often being characterised as a dichotomous concept, amenable only to the award of a passing grade, and not to higher grades. During the 1990s, Australia initiated a 'Grade Debate' into the feasibility and desirability of grading CBAs. Concluding that grading is both feasible and desirable, the emerging literature explored a range of potential criteria for grading, a variety of alternative approaches, and a number of conceptual bases.

The most important insights from this small literature are analytical, concerning the conceptual bases for grading, and the legitimacy of alternative criteria. Drawing inspiration from the wide variety of grading practices that have been implemented across Australia (and New Zealand and England) since the early 1990s, contributors to the Grade Debate have attempted to unpick tacit assumptions, and to explore underpinning principles for CBA grading. Consequently, the literature is stronger on theory than on practice; although work by the Melbourne School during the 2000s has consistently endeavoured to translate theory into practice, leading to the development and trialling of a number of fairly sophisticated approaches.

At the heart of the Grade Debate lies the issue of what might count as a legitimate grading criterion, and why. The degree to which grading criteria ought to be clearly meaningful has been particularly controversial. Some authors have recognised widely used criteria such as meeting submission deadlines or consistently high motivation. Others have dismissed criteria like these as unmeaningful and therefore illegitimate. Such dismissals need to be interrogated, though. To reward consistently high motivation, or the meeting of submission deadlines, is to reward a learner's diligence, or effort. It is fair to say that this is to reward an input to learning, rather than an outcome from learning. Yet, is that necessarily grounds for dismissal as a legitimate grading criterion? Are inputs to learning *un*meaningful as grading criteria, or simply *differently* meaningful? Views differ on important questions like these.

A number of contributors to the Grade Debate have adopted the idea of Standards-Referenced Assessment (SRA), from Sadler (1987), as a theoretical foundation for developing CBA grading practices. The Melbourne School, in particular, has emphasised the importance of treating competence, not as a dichotomous concept, but as a developmental continuum of learning. Scholars working in this tradition have tended to argue most strongly against the use of unmeaningful grading criteria; tending to define meaningfulness in terms of the underlying proficiency continuum. From this perspective, the ideal is to specify domain-specific grading criteria, which articulate levels of proficiency in terms of acquired learning outcomes; eg grading criteria for a carpentry qualification would refer to elements of competence in carpentry.

Not all stakeholders would adopt this perspective, however; and some have argued for domain-general, ie generic grading criteria. Generic criteria are written to apply across multiple domains; eg criteria that refer to communication, or problem solving, skills. There are certainly advantages to generic criteria. In particular, because they can be applied across a range of domains, they are relatively cheap to develop. In addition, the use of generic criteria means that grades can be interpreted in the same way across domains, which is also attractive. Where generic criteria have been adopted in Australia, this has tended to be associated with a two-step, dual-outcome, approach to assessment and reporting. The first step is CBA, as traditionally practised in terms of: an atomistic (ie detailed but disconnected) specification of learning outcomes (LOs) and assessment criteria (AC); a mastery measurement model, ie all LOs and AC need to be met; and exhaustive sampling, ie all LOs and AC need to be assessed. This leads to a pass/fail judgement, ie Competent, or Not-yet-competent. The second step is grading, on the basis of generic criteria; often based upon the learner's performance across the course as a whole. This leads to a grading judgement, eg Merit. Dual-outcome reporting means presenting both outcomes alongside each other, eg Competent with Merit.

Debate over the use of specific versus generic criteria leads to a deeper question concerning whether there really are intrinsic differences between grading in TVET contexts and grading in other contexts. Bearing in mind that CBA – characterised by atomistic specification, mastery measurement, and exhaustive sampling – represents a relatively uncommon assessment approach, we might assume that such intrinsic differences do exist; particularly given the unique significance of the competence threshold in TVET contexts. Yet, the reconstruction of competence as a developmental continuum of learning, as set out by the Melbourne School, begins to challenge this view. It invites us to think of the competence threshold as just one point on a continuum of learning, which implies that there may be no intrinsic difference between higher grades and the passing grade. More importantly, it invites us to question whether CBA, as traditionally practised, is the optimum approach for determining *any* grade. SRA, as described by its originator, Royce Sadler, embodies a far more complex and holistic process than CBA, as traditionally practised. Yet, some would argue that this is closer to the model that was originally intended by key Australian stakeholders during the early 1990s. In short, the differences between assessment in TVET contexts and assessment in other contexts may not be quite so intrinsic as current practices might lead us to believe.

A particular feature of the Australian literature is its quest to identify underlying principles for grading in TVET contexts. Unfortunately, this literature does not actually develop the idea of TVET-specific grading principles at all well. Indeed, the majority of principles within most of the proposed sets are neither TVET-specific nor even grading-specific. Instead, they are simply basic principles of assessment. As such, the present report concludes with nothing more than a very high-level statement of principle for grading in TVET contexts:

1. The grounds for differentiating between candidates, via grades, must be defensible; that is, sufficiently **meaningful** and sufficiently **useful**, when judged in relation to a **profile of purposes**.

2. The grading process must be sufficiently **accurate**.

3. The **benefits** from implementing the grading process must, on balance, **outweigh** its **costs**.

In other words, it is not enough that any particular grading practice is capable of differentiating between candidates; nor even that it is capable of differentiating between candidates reliably. Instead, grading practices need to be capable of differentiating purposively, accurately, and in a manner that is economically, politically, and socially acceptable.

Background

In the accompanying report (Newton, 2018), approaches to grading regulated Vocational and Technical Qualifications (VTQs) in England were considered from two perspectives: policy and practice. The report concluded that current practice in grading VTQs is not underpinned by a straightforward, generally accepted, set of ‘good practice’ principles. This raised the question of what principles for grading VTQs might look like; which begged a further question concerning what the wider literature has to say on this matter.

Unfortunately, there is no neatly circumscribed body of work on grading, let alone grading in Technical and Vocational Education and Training (TVET) contexts. Of course, grading is prevalent in educational settings all over the world, and has remained so for well over a century. And, naturally, it has been subjected to a great deal of research and analysis over the decades too, although much of the published work focuses on grading in the context of North American schooling (eg Brookhart et al, 2016; Anderson, 2018). While textbooks on grading are not hard to find, often tailored to particular educational settings and to specific regional concerns – such as *Transforming Classroom Grading* (Marzano, 2000), or *Grading Student Achievement in Higher Education* (Yorke, 2008) – the full corpus of work on grading remains widely dispersed and unwieldy. Implications for grading in TVET contexts are far from clear.

Importantly, the task of grading VTQs – especially the kind of VTQs that are common in England – raises a number of fairly unique challenges. These stem, in particular, from the idea of assessing competence. The idea of Competence-Based Assessment (CBA) – which emerged from work in the USA on Criterion-Referenced Assessment (eg Glaser, 1963) and Mastery Learning (eg Bloom, 1968) – became closely associated with VTQs in England during the 1990s, with the introduction of National Vocational Qualifications (NVQs). NVQs were designed to certify competence, defined as: “the standard required successfully to perform an activity or function” which in employment areas, unlike general education, meant “performing to professional or occupational standards” (Jessup, 1991, p.25). The following characteristics came to be seen as hallmarks of CBA:

- the atomistic specification of measurement standards in terms of learning outcomes and assessment criteria;
- a mastery measurement model, meaning that a certificate of competence could be interpreted to mean competent across each and every learning outcome and assessment criterion;
- assessment based on the exhaustive sampling of learning outcomes and assessment criteria.

The first and foremost challenge, in a CBA context, is the question of whether grading – that is, the award of higher grades beyond the passing grade – is even compatible with the notion of assessing competence. From a pragmatic perspective, once a learner has reached the professional or occupational standard associated with a qualification, they are essentially ready for certification. Thus, strictly speaking, in Competence-Based Training (CBT) contexts, higher grades have no significance.

Consequently, NVQs were not graded higher than a Pass. However, in practice, the CBA model was extended to contexts that more closely resembled general education, where the arguments against awarding higher grades seemed far weaker. The development of General National Vocational Qualifications (GNVQs) in England during the early 1990s provides a clear example of this. Indeed, the GNVQ did award higher grades.

Over the years, since the 1990s, CBA has become increasingly influential; not only in England, but internationally, including in New Zealand and, in particular, Australia. Currently, in England, many regulated VTQs incorporate features and processes that resonate with the core characteristics of CBA; including atomistic specification, mastery measurement, and exhaustive sampling. Indeed, it seems likely that the core characteristics of CBA currently operate as something of a default template for designing (many although not all) VTQs in England (Newton, 2018).

Since the introduction (and demise) of the GNVQ, many VTQs bearing the hallmarks of CBA have also been graded. Surprisingly, though, the UK has produced no identifiable literature on grading CBAs, or grading in TVET contexts more generally (cf. Johnson, 2008). Apparently, the only country to have produced anything resembling a literature on grading CBAs is Australia.[1] Influenced by developments in the UK, Australia also promulgated CBA across its states and territories during the 1990s. However, unlike the UK, it initiated a national debate on the subject of grading CBAs. This resulted in a number of research projects, position papers, and methodological innovations; which have spanned the past three decades. That said, even this literature is small. Furthermore, only some of it has officially been published, while the rest resides in the 'grey' zone, meaning that certain key reports are hard to locate. What follows is a series of notes from this small literature, focused specifically upon grading practices, ie the various approaches to grading CBAs that have been proposed and, in many instances, also trialled in Australia.[2]

The purpose of this review is to provide a broader horizon on approaches to grading CBAs (and CBA-influenced VTQs) than is provided by the accompanying survey of current grading practices within regulated VTQs in England. Unfortunately, neither the Australian literature, nor the accompanying survey, leads to simple conclusions concerning principles of good practice (or bad practice) for grading in TVET contexts. Consequently, like the accompanying survey, the Australian literature helps to provide a foundation for dialogue – helping us to deepen our engagement with issues of grading in TVET contexts in England – but it does not provide a straightforward resolution. Nor does it result in a detailed critique of specific practices.

[1] This is not to suggest that only Australia has extensive experience of grading in TVET contexts. For instance, a report by Cedefop (2015) on ensuring the quality of certification in TVET suggested that the use of 'grids' for grading vocational learning is common across many systems (see pp.53-4). However, although such practices might be widespread, they do not appear to have stimulated a great deal of scholarly work in this area.

[2] This literature is occasionally cited further afield, when similar issues arise, such as: Nurse Education in Australia (eg Andre, 2000); TVET in Ghanaian polytechnics (eg Boahin, 2018). The present report is restricted to the core literature, however, which enables a more focused and coherent analysis.

Notes from the Australian literature

The present report is not intended to provide a comprehensive review of all of the Australian papers that have discussed grading for CBAs over the years. Instead, it provides a more focused overview of papers that provide helpful insights into alternative grading practices and principles.

The papers discussed below explore the desirability and the feasibility of grading in TVET contexts. From the perspective of the present report, their specific conclusions tend to be less important than the evidence and analysis upon which those conclusions are based. Consequently, rather than reviewing each paper systematically, the following account extracts key insights, with common themes identified and discussed towards the end of the report.

The Australian literature can be divided into two phases, which correspond roughly to the 1990s (the Grade Debate) and to the 2000s (the Melbourne School).

The Grade Debate (1990s)

The *Grade Debate* was the title of a research report by Thomson, Mathers, and Quirk (1996). It arose from a 1993 paper, by the first author, which drew attention to the urgent need for research to inform policy in the area of CBA, including the appropriateness of assessing and reporting *levels of performance* in CBT, ie grading for CBAs. Sponsored by the National Centre for Vocational Education Research (NCVER), this research report, which also included a synthesis of existing contributions from the grey literature, set the scene for subsequent work on grading CBAs in Australia during the 1990s. Insights from key papers that emerged during the period of the Grade Debate are presented below, paper-by-paper.

Thomson et al (1996)

The Thomson report was subtitled: *Should we grade competency-based assessment?* It explored the pros and cons of grading CBAs, in the context of evidence concerning the prevalence of such practices. Although CBA had been promoted as national policy since the beginning of the 1990s, by the mid-1990s it was still not universal across Technical and Further Education (TAFE) institutions in Australia. The researchers identified considerable support for the policy of grading CBAs; particularly for grading learners in TAFE institutions, although less so for grading learners in the workplace. However, they also identified a significant element of resistance to grading, from stakeholders who firmly believed it to be antithetical to the principle of CBT. Hence, the Grade Debate.

Figure 1. Discriminators used by the lecturers when assigning grades.

Distinction

- Student's original ideas showing greater insight into topics discussed should lead logically to their conclusions. Answers should be close to required length with all key points covered.
- Student is clearly able to show evidence of judgement in applying the theory to the topic.
- Clear evidence of wider reading through sources referenced in answer and Harvard reference system correctly used; a bibliography with each answer.
- Very high standard of presentation.
- Work submitted by due date (unless prior arrangement made with lecturer).

Credit

- Student's original ideas should lead logically to their conclusions. Answers should be close to required length with all key points covered.
- Student is able to apply the theory to the topic.
- Evidence of wider reading through sources referenced in answer and Harvard reference system correctly used; bibliography with each answer.
- High standard of presentation.
- Work submitted by due date (unless prior arrangement made with lecturer).

Competency achieved

- Student's ideas should lead logically to their conclusions. Answers may deviate from required length.
- Student is able to apply the theory to the topic but not to the extent required for a higher grade.
- Harvard reference system correctly used, but less evidence of wider reading than that required for a higher grade; a bibliography with each answer.
- Acceptable standard of presentation.
- Work submitted by due date (unless prior arrangement made with lecturer).
- Some resubmission may be required for (part) answers not competent.

Note: Work submitted late without prior arrangement with lecturer will not be given a higher grading than *competency achieved*.

The report concluded:

> An absence of national co-ordination has resulted in the State and Territory TAFE systems and private training providers' determining their positions based on individual interpretations of underlying principles.
>
> Where grading is being implemented in Australia there appears to be a range of assessment approaches and assumptions operating. Of particular concern is the quality of the assessment instruments currently being used to arrive at grades. There is a need to develop assessment exemplars to encourage improvement in existing procedures.

(Thomson et al, 1996, p.viii)

The research undertaken for the Thomson report included five case studies of approaches to grading CBAs (pp.31-46). For example, Case Study 2 concerned National Accounting modules; which drew, in particular, upon the experiences of South Australian TAFE accounting lecturers, engaged in a pilot project on grading accounting students. A number of 'additional assessment indicators' were identified for the purpose of scaffolding grading decisions within this project, which included:

- ability to express ideas;
- evidence of wider reading;
- logical presentation of ideas; and
- appropriate presentation of work.

Reproduced in Figure 1, above, is an example of grading criteria relating to an assessment from the module on *Business planning and control* (see Thomson et al, 1996, p.36). This assessment covered only two of eight learning outcomes, and involved scoring unsupervised written essays or short answers prepared by the students for a particular task. Note that these 'discriminators' (generic grading criteria) reflect a 'best fit' approach to grading candidates' work.

From its five case studies, the Thomson report identified eight types of grading criteria in operation, which it summarised in a table (see Thomson et al, 1996, p.22). This is reproduced, below, as Table 1. All three columns (criterion, example, comment) are reproduced in full.

Table 1. Types of grading criteria: A summary

Criterion/indicator	**Example**	**Comment**
1 Number of attempts	Higher grades in Accounting are only available on first attempt. If standard met on second or later attempts, only 'competency achieved' grade is available.	It is clearly in interest of students and trainees to defer assessment until they believe they are ready. This discouragement of frivolous attempts also reduces pressure on teachers and lecturers.
2 Level of supervision	Level of supervision in the Australian defence organisation's Statement of Attainment form varies from *expert* (where no supervision is mentioned) to *skilled* (normal supervision), *trained* (close supervision) and *partially trained* (constant supervision).	It should also be noted that level of supervision is frequently used in the national industry standards to differentiate ASF levels. Critics of the grading process will sometimes claim that this criterion is, in fact, differentiating ASF levels not grades within a level.
3 Speed of performance	The ability to do things quickly (e.g. cook a meal) relates to industrial productivity in Tourism and Hospitality.	Many employers are interested in speed of performance as their profit margins are determined by the time taken to complete certain jobs.
4 Meeting deadlines	Grading in the Electrical industry case study is confined to theory modules (or theory sections of the modules) but not only must the learning outcomes be met but they must be met 'within the specified time'.	Meeting deadlines is similar to the 'speed of performance' criterion. However, in the cases cited one applies mainly to the workplace and the other to the classroom.
5 Consistency	Consistency of performance of most outcomes is a highly valued characteristic in industries such as Tourism and Hospitality.	Consistency is being introduced into new AVTS cooking assessment criteria. As a judgement about consistency requires repeated observation it will usually be appropriate to assess this in the workplace.

Criterion/indicator	Example	Comment
6 Accuracy	Among the examples provided by Edgecombe (1995) are Accounting grades given according to the number of errors made in calculations (e.g. Distinction = 0 errors, Credit = 1 error, Pass = no more than 3 errors).	Accuracy of performance is one of the easiest of criteria to devise and apply. Unfortunately it tends to be applied in an arbitrary fashion. Accuracy criterion should only be adopted after a period of trials and evaluation.
7 Profile (word picture, testimonial, reference)	The Australian defence organisation's Statement of Attainment form has provision for 'supervisors' to make a series of comments about the trainee such as recommendations for future postings, employment, future training or other developmental opportunities. Additional information gathered in this way goes under various other names, e.g. testimonials, references. Johnstone et al. (1995) provide an interesting discussion of the use of profiles.	Profiling provides valuable confirmatory evidence about the grade assignment process. The downside of profiling is the additional workload it places on teachers and trainers.
8 Complex traits: – artistry – creativity – flair – initiative – motivation – adaptability – efficiency	Complex human behaviours having to do with attitudes, values, interests and appreciation are frequently cited as important to the grading process. (For example, the artistic presentation of a plate of food or the creative ability to generate new ideas to solve a problem.) Furthermore, these complex traits are usually associated with grading at the higher ASF levels. This is why they have been separated from numbers 3-6 above.	Making judgements about complex traits of this sort is a challenge. Not only is it usually the province of experts, but also the experts themselves need some guidance to ensure their judgements are reliable.

The Thomson report also included a section on: *Issues related to the implementation of grading* (pp.12-16). This synthesised insights from a number of pre-existing reports, including contributions by Peddie, and Wilmut and Macintosh. These contributions were subsequently elaborated within a Special Issue of the *Queensland Journal of Education* (see Maxwell, 1997a).

Maxwell (1997a)

Maxwell (1997a) introduced three contributions to the Grade Debate – from New Zealand (Peddie, 1997), England (Wilmut and Macintosh, 1997), and Australia (Maxwell, 1997b), respectively – each of which interrogated the following questions concerning competence-based education and training and subsequent competitive selection:

1. Is it possible to provide some means of differentiation among students under a competency based assessment system without threatening the coherence of competency based assessment?
2. Even if appropriate differentiation is possible, is it desirable, or can the problem be satisfactorily resolved in some other way?

Affirming that it is both possible and desirable to grade CBAs, the Special Issue was notable for its breadth of thinking on alternative approaches.

Peddie (1997)

Peddie began by emphasising that competence-based learning outcomes that require totally accurate performance are not amenable to grading above the competence threshold (eg where health and safety is involved). However, for outcomes that are amenable to grading, he distinguished between methods that grade on the basis of qualification-specific learning outcomes, and methods that grade on the basis of factors beyond them.

Methods for grading on the basis of specified learning outcomes included:

1. achievement of standards required for the next framework level;
2. achievement at a standard well beyond the competency/credit standard;
3. speed – attaining learning outcomes at a faster rate; and
4. consistency of performance.

Peddie noted the similarity between Methods 1 and 2, contextualising both in relation to qualifications that represented similar learning outcomes at multiple qualification levels, with each level reflecting a higher plane of professional practice. Method 2 suggested the existence of significant achievement gaps between competence standards at adjacent levels, within which higher grade standards could be located. Method 1 suggested simply that where some (but not all) learning outcomes have been achieved at the higher level, this might be recognised with a higher grade (at the lower level).

Peddie considered speed as a grading criterion in relation: either to speed of acquiring learning outcomes (ie time taken to complete a unit); or to speed of demonstrating the attainment of a learning outcome (ie time to complete an assessment task). He suggested that speed of learning was potentially admissible as a grading criterion in the sense of having public credibility as an indicator of merit; while speed of performing was potentially admissible in the sense of indicating superior expertise. He distinguished both from instances whereby speed is a basic formal requirement of a learning outcome, eg speed of service.

The criterion of consistency in achieving learning outcomes was intended to capture the idea that learners who consistently need only one opportunity to demonstrate competence might be considered superior to those who need to attempt assessment tasks repeatedly, in order to pass. Peddie suggested that this criterion might be more acceptable within certain non-European cultures (and that it had recently been rejected by the New Zealand Qualifications Authority).

Peddie's remaining four methods were intended to capture learner characteristics that lay beyond qualification-specific learning outcomes, but that were still somehow competence-related. Methods for grading on the basis of factors beyond specified learning outcomes included:

5. transfer of skills to new situations;
6. achievement of additional learning outcomes;
7. originality, creativity, flair; and
8. outstanding attitudes, approach to learning, motivation.

He discussed transfer mainly in relation to the challenge of judging it, eg what if one learner demonstrated transfer of one skill to a variety of contexts, while another learner demonstrated transfer of several skills into one new context – should we judge them equally meritorious? A more general problem, which he did not explicitly consider, is whether the idea of a non-transferrable learning outcome is even credible, ie whether a purely context-bound performance could be considered sufficient to meet even the basic competence standard.

Whilst recognising the achievement of additional learning outcomes as a potential criterion, Peddie acknowledged that this was at least somewhat paradoxical in relation to CBT: if the additional learning outcomes are considered important enough for the award of a Merit grade, then why are they not part of the set required for competence? Noting a parallel between additional learning outcomes, as a criterion, and speed of learning, he suggested that: for time-bounded courses, they might amount to the same criterion; whereas, for non-time-bounded courses, there might be reason to question whether the 'plodder' who eventually achieves additional learning outcomes is genuinely worthy of a Merit grade.

Although creativity might seem attractive as a grading criterion – where creativity is not specifically articulated as a learning outcome – Peddie observed that 'creatives' are not always assumed to be amongst the highest achievers. More importantly, creativity is often not expected, and sometimes not desired, of learners during the early stages of learning within a domain.

Finally, just as for the consistency criterion, Peddie suggested that the 'good learner' criterion might be considered more acceptable in certain cultures than in others. The fact that it determines Merit purely on the basis of the learner – entirely divorced from the actual learning that has occurred – renders it potentially problematic.

Wilmut and Macintosh (1997)

Wilmut and Macintosh identified four approaches to differentiating learners, at the end of a phase of competence-based education or training, for the purpose of competitive selection:

1. using sources of information from outside the learning outcomes;
2. generating grades from ungraded units;
3. grading the units; and
4. grading the whole qualification.

The first of these – which might involve a record of achievement, or a selection test – does not involve grading, per se, and will not be considered further.

While Peddie's analysis was essentially conceptual, Wilmut and Macintosh's was more pragmatic; informed by recent experiences in the UK, particularly the GNVQ experience, where grades were awarded at the overall qualification level.

Wilmut and Macintosh identified two distinct approaches to grading the whole qualification:

4a. applying overall grading criteria related to the competency base but applied overall; and

4b. grading on consistency or speed of performance on an overall assignment or project.

In the first approach, generic grading criteria are used to judge the complete work of the candidate across all units of the qualification, as presented within a portfolio, or suchlike. The criteria would need to be applied holistically, suggesting that they might be satisfied across a substantial portion of the evidence in a candidate's portfolio, but that the candidate need only satisfy some aspects of each grading criterion in each unit. The criteria specified for each grade, which would necessarily be framed in context-free language, might (as in the case of the GNVQ) include attributes such: as planning, information handling, or evaluation; or (as suggested by Peddie) creativity, or originality.

Wilmut and Macintosh also referenced Peddie's identification of consistency and speed as potential generic criteria for grading a whole qualification. Perhaps more plausibly, though, they noted the possibility that grades might be determined from performance on a specially designed integrating project or assignment; nowadays (in England) typically referred to as a 'synoptic' assessment. Again, this would require the specification of a set of grading criteria; although, these could be tailored more closely to the particular demands of the assessment.

The obvious alternative to grading at the qualification level is to grade at the unit level. An overall qualification grade could then be determined, for instance, on the basis of an aggregate of the unit grades. Three approaches to unit grading were identified by Wilmut and Macintosh:

3a. applying overall criteria to each unit;

3b. using unit-specific criteria; and

3c. adding unit tests.

The first option would, once again, require the specification of generic grading criteria: underlying attributes, such as quality of communication; or highly regarded skills, such as planning, information handling, or evaluation. The second option would require the specification of unit-specific grading criteria, tailored to the particular learning outcomes in question. The third option would require the development of unit-specific tests, specifically designed to contribute information for the purpose of grading; ie teacher assessment for the competence judgement, in combination with testing for the grading judgement.

Instead of deriving grades from the whole qualification, or from unit grades, alternative methods ('for generating grades from ungraded units') might include:

2a. using combinations of vocational units;

2b. using core skill units; and

2c. adding tests or assignments.[3]

The key differentiator, within this category of approaches, is the particular combination of units offered for certification; with the award of higher grades being contingent upon achieving specified additional units, either generic or domain-specific.

Maxwell (1997b)

Whereas Peddie's analysis was essentially conceptual, and Wilmut and Macintosh's more pragmatic, Maxwell's was more theoretical; extending the idea of Standards-Referenced Assessment (SRA) from Sadler (1987). Sadler has written extensively on the specification and promulgation of assessment standards. SRA is a model of assessment, which is premised upon the effective calibration of assessors' professional judgements to a common set of standards. In other words, assessors' judgements are referenced to descriptions and exemplifications of the standards associated with specified learning outcomes. Thus, Maxwell emphasised the centrality of professional judgement to CBA; in contrast to an impoverished but widespread view, in which professional judgement is trivialised, as though assessment could be reduced to an unproblematic, observational, list-ticking, process (see also Wolf, 1993; Gonczi, 1994; Hager, Athanasou, and Gonczi, 1994; Hager and Gillis, 1995).

Introducing an idea from Wolf (1993), which the Melbourne School would also soon develop, Maxwell argued that the competence standard associated with any CBA constitutes just one point on a continuum of proficiency associated with the domain of practice in question. By characterising proficiency as a continuum, not a binary construct, the idea of higher grades above the passing grade acquires greater theoretical legitimacy. Revisiting ideas from Peddie, and Wilmut and Macintosh, Maxwell proposed that there are four basic types of approaches that can be implemented for differentiating learners:

[3] It was not clear how this differed from methods described within other approaches, so will not be discussed further.

1. preparation of a (processed) portfolio;
2. standards-based assessment of the whole course;
3. standards-based assessment of some or all units; and
4. standards-based assessment within some units.

The idea of a processed portfolio implies that the evidence that it contains is made more digestible – either as a 'record of achievement', or a 'summary report', or a more detailed 'principal's report' – although not specifically in the form of an overall grade. If an overall grade were to be awarded, then this would transform the idea of a processed portfolio into the second approach, which Maxwell subdivided into:

2a. standards-based assessment of a portfolio;

2b. standards-based assessment of a special project; and

2c. standards-based assessment of a special module.

The idea of a special project, in 2b, is what Wilmut and Macintosh referred to as a specially designed integrating project or assignment (a.k.a. a 'synoptic' assessment). The idea of a special module, in 2c, extends this, to include an explicit framework of support for acquiring the necessary skills, ie a structured sequence of learning activities to develop and practise them.

Maxwell's discussion of standards-based assessment of some or all units essentially reproduced ideas from Wilmut and Macintosh. However, his discussion of standards-based assessment within units went further by formally recognising grading at the level of particular learning outcomes or competencies:

4a. standards-based assessment of some learning outcomes; and

4b. assessment of desirable versus essential competencies.

The first option, here, recognises Peddie's opening remark, that certain learning outcomes are not amenable to grading; suggesting that grading might be best operationalised at the level of individual learning outcomes, rather than units, or qualifications. The second option recognises the possibility of being able to distinguish between essential competencies within a domain (required for the passing grade) and desirable competencies (required for higher grades); desirable competencies being beneficial but not critical.

Table 2. Strengths and weaknesses of alternative approaches to grading

Criteria for grading assessment outcomes	Description	Strengths	Weaknesses
1. Additional performance criteria (e.g. competent with merit)	A set of additional performance criteria are developed for which students may prepare and be assessed. The performance criteria are developed in conjunction with the industry concerned. Only students who opt to be assessed are candidates for the merit grade	Assessment against the merit criteria is independent of the unit assessment	Requires development of additional criteria Could involve significant costs in additional resource and assessor training
2. Hierarchy of performance criteria within unit	An existing competency unit is reviewed to establish higher levels of performance for each performance criteria in the unit. These are then defined and become the merit grade criteria as in (1)	Uses existing performance criteria	May be difficult to establish meaningful higher levels of performance Definition between grades may be too fine for valid and reliable assessment Could involve significant costs in additional resources and assessor training
3. Measure of underpinning knowledge (theory)	The underpinning theoretical knowledge for a unit is defined more thoroughly and criterion are established for the separate criterion referenced assessment of the knowledge. The assessment outcomes are still used as evidence for the overall unit assessment but a separate supplementary graded report is provided on the knowledge component. The grade is only provided if the candidate has demonstrated competence in the unit concerned	Assesses and reports on an important component of competence needed for transfer of skills Is similar to criterion-referenced grading used in other education sectors	May be difficult to identify the required hierarchy of knowledge and develop appropriate criteria and assessment instruments Could significantly add to the costs of assessment and reporting

Criteria for grading assessment outcomes	Description	Strengths	Weaknesses
4. Numbers of assessments before competence	Merit grades are only awarded on the first attempt. If the performance criteria are achieved on the second or later attempts, only the competent grade is awarded	Relatively simple to monitor and implement Encourages candidates to defer assessment until they were confident that they are ready	Would tend to penalise candidates who are nervous under assessment conditions Raises the question of what the purpose of the merit grade is and what it actually indicates
5. Measures of key competencies	Key competencies are separately assessed against defined criteria. If the candidate demonstrates sufficient performance against defined criteria then a merit grade is awarded. The grade is only provided if the candidate has demonstrated competence in the competency unit concerned	Reports on important competencies that are relevant across different contexts	May be difficult to establish meaningful levels of performance Assessment is likely to be context dependent Could involve significance costs in additional resources and assessor training
6. Time to achieve competence	The time spent by students completing a training module is monitored. Students completing the module in a specified time or less are awarded a merit grade. If the performance criteria are achieved on the second or later attempts, only the competent grade is awarded.	Time to complete tasks is seen to be important dimension of workplace performance by many employers	May be difficult to track and measure accurately Extraneous factors may affect the time taken to complete a module affecting the validity of the measure

Criteria for grading assessment outcomes	**Description**	**Strengths**	**Weaknesses**
7. Degree of supervision	The degree of supervision required by the apprentice/trainee during on-job training/assessment is monitored. The degree of supervision is rated against two or more defined grades. The grade is only provided if the candidate has demonstrated competence in the competency unit concerned.	The ability to work with minimal supervision or guidance is seen to be important dimension of workplace performance by many employers	Judgement of performance for this criterion could be highly subjective affecting reliability and/or fairness Assessment in the training environment may not predict similar performance in the workplace
8. Speed of performance	The speed with which the apprentice/trainee completes workplace tasks during on-job training/assessment is monitored. The speed of performance is rated against two or more defined grades. The grade is only provided if the candidate has demonstrated competence in the competency unit concerned.	Speed of performance of workplace tasks is seen to be important dimension of workplace performance by many employers	Assessment in the training environment may not predict similar performance in the workplace
9. Measure of adaptability	The degree to which the apprentice/trainee is able to adapt to new and unusual workplace tasks during on-job training/assessment is monitored. The assessed adaptability is rated against one or more defined grades. The grade is only provided if the candidate has demonstrated competence in the competency unit concerned.	The ability to adapt to new work tasks or workplace contexts is seen to be important dimension of workplace performance by many employers	Judgement of performance for this criterion could be highly subjective affecting reliability and/or fairness May be difficult to establish meaningful levels of performance Assessment in the training environment may not predict similar performance in the workplace Could involve significant costs in additional resources and assessor training

Criteria for grading assessment outcomes	Description	Strengths	Weaknesses
10. Problem-solving ability	The degree to which the apprentice/trainee is able to solve problems during on-job training/assessment is monitored. The ability to solve problems is rated against one or more defined grades. The grade is only provided if the candidate has demonstrated competence in the competency unit concerned.	The ability to solve workplace problems is seen to be important dimension of workplace performance by many employers	Judgement of performance for this criterion could be highly subjective affecting reliability and/or fairness May be difficult to establish meaningful levels of performance Assessment in the training environment may not predict similar performance in the workplace Could involve significant costs in additional resources and assessor training
11. Measures of various traits, e.g.: • motivation • initiative • flair • artistry • creativity • accuracy • efficiency	The degree to which the apprentice/trainee is able to demonstrate the trait during on-job training/assessment is monitored. The ability to demonstrate the trait is rated against one or more defined grades. The grade is only provided if the candidate has demonstrated competence in the competency unit concerned.	Dependent on the industry and occupation, these traits may be seen to be important dimensions of workplace performance by the employers concerned	Valid and reliable judgement of performance for these traits is extremely difficult and could be highly subjective, affecting validity, reliability, cost effectiveness and/or fairness May be difficult to establish meaningful levels of performance Assessment in the training environment may not predict similar performance in the workplace Could involve significant costs in additional resources and assessor training

Rumsey (1997)

Writing for the Australian National Training Authority (ANTA), Rumsey discussed his research into the suitability of alternative approaches to reporting assessment outcomes under New Apprenticeships, which adopted a competence-based approach to training and assessment. He identified five broad types of assessment reports, of which the third is of most relevance to the present review: criterion-referenced graded assessment reports. In a very useful table, Rumsey summarised the relative strengths and weaknesses of a variety of approaches to grading CBAs. Table 2, above, reproduces this table (see Rumsey, 1997, pp.40-43). All four columns (criteria, descriptions, strengths, weaknesses) are reproduced in full.

Williams and Bateman (2003)

The research undertaken by Williams and Bateman – produced by the NCVER with funding from ANTA – provides a useful overview of progress during the Grade Debate phase. Via literature review and stakeholder consultation, it sought to update the research originally undertaken by Thomson et al (1996), to take account of changes in the TVET environment.

Their report began by observing that, in the absence of clear policy on graded assessment in TVET, a range of practices had evolved. Most significantly, it concluded: "that 'good' practice in competency-based assessment itself, let alone graded competency-based assessment, is still not fully understood nor universally implemented across the national training system" (Williams and Bateman, 2003, p.5). Key findings from their research (see pp.5-6) included:

- Few policies or guidelines exist to assist registered training organisations in implementing graded assessment in a valid and consistent manner.
- Even where policy guidelines exist, there is variation in the way graded assessment is carried out.
- Instances of 'good' practice in graded assessment were identified. These incorporated features such as professional development of assessors, provision of policy and/or guidelines, provision of exemplars of assessment tools and grading schemas as well as validation processes.
- Limited information is available and findings are mixed regarding the additional costs that may be incurred in implementing a graded assessment system. Indeed, there appears to be little will to explore this issue at either registered training organisation, state or national level.
- The lack of transparency in reporting is of major concern. The wide variation in grading methodologies employed by registered training organisations leads to significant discrepancies in what the grades represent. Transparency in reporting is essential to make the grades meaningful to stakeholders.

Four models of grading were presented in the Appendix to the report, to illustrate approaches deemed to be compatible with the idea of CBA. In addition to examples from Western Australia, Queensland, and the University of Ballarat, Model 4 introduced approaches pioneered by researchers at the University of Melbourne.

The Melbourne School (2000s)

Maxwell (1997a) introduced his Special Issue of the *Queensland Journal of Education* by contrasting a complex, holistic, conception of competence – as originally intended by key Australian stakeholders during the early 1990s – with the simpler, more atomistic, conception that had eventually come to dominate CBA practices. The Melbourne School – led by Patrick Griffin and Shelley Gillis from the Assessment Research Centre at the University of Melbourne – took issue specifically with the reduction of competence to a dichotomous concept, as though competence is a quality that is either present or absent. Conversely, they insisted that competence must be reconceptualised as a developmental continuum of learning. This provided the basis for a theoretical model of grading in TVET, which Gillis and Griffin pioneered throughout the 2000s.

Setting out their stall in 2005, Gillis and Griffin argued that the extant literature on grading CBAs had been characterised by three major misperceptions, which they countered as follows. First, CBA is not antithetical to the idea of grading. This misperception, they argued, was based on a naïve view of criterion-referenced assessment, which reduces the assessment of competence to a series of binary, pass-fail judgements. Instead, as proposed by Glaser, the originator of criterion-referencing, assessment judgements ought to be referenced to “stages along progressions of increasing competence” (Glaser, 1981, p.935).

Second, contrary to the impression given by the promotion of unmeaningful criteria for grading CBAs, such as ‘speed of completion’ or ‘number of attempts’, it is quite possible to develop meaningful grading criteria. However, generic criteria, designed to be applied to performance in general regardless of content or context, are unsatisfactory (according to the Melbourne School). Instead, candidates’ performances need to be evaluated against criteria that are both content- and context-specific. Such criteria will clearly require the exercise of professional judgement.

Finally, grading does not need to be conceptualised as somehow ‘bolted-on’ to CBA. During the Grade Debate, writers tended to assume that the primary task of CBA is to assess competence, as an either/or concept. Grading, from this perspective, is undertaken as an entirely separate activity, based upon supplementary criteria. Conversely, once the pass-fail cut-off is recognised as simply one way to partition a proficiency continuum, the special status of the passing grade vanishes, becoming just one of multiple, sequential cut-off points along a developmental continuum of learning.

Like Maxwell, Griffin and Gillis acknowledged the idea of SRA, from Sadler (1987), as a basis for theorising the assignment of candidates to proficiency bands – and, by extension, to meaningful grades – premised on the idea of a developmental continuum of learning. They described, explained and illustrated their approach across a variety of reports and presentations (eg Griffin and Gillis, 2000; Griffin, 2001; Bateman and Griffin, 2003; Gillis and Griffin, 2004; Gillis and Griffin, 2005; Griffin, Gillis, and Cavitto, 2007; Griffin, 2007). Their model acknowledges the centrality of professional judgement to CBA.

Griffin (2001) proposed that, within a standards-referenced framework, there are eight (somewhat iterative) steps in the process of defining proficiency levels:

1. **Define the proficiency that is to be assessed (within a unit).** What does the high end look like? What does the low end look like? Does it have only one dimension, or more?

2. **Develop tasks to assess this proficiency.** As a set, these tasks should allow us to differentiate between candidates with differing proficiency levels.

3. **Develop rubrics, ie a marking/scoring scheme for each critical indicator/task.** These rubrics define distinguishable levels of performance quality.

4. **Assign each performance quality level, for each critical indicator/task, to a relative position on the proficiency scale, via a calibration process.** This might be achieved either via professional judgement or on the basis of statistical analysis. Qualities located at the same relative position on the proficiency scale help to elucidate the proficiency continuum (see 6).

5. **Locate cut-points on the scale in such a way that levels of proficiency are interpretable, separable, and distinct.** This is the most technically complex of steps, and its nature will depend on the approach adopted in the previous step.

6. **Interpret the proficiency scale.** The partitioning of the proficiency scale can be given meaning by synthesising qualities located at the same relative position. The continuum itself is a synthesis of the hierarchy of levels.

7. **Refine the scale.** Does the proficiency scale appear to be coherent and complete? Do any of the tasks/rubrics need fine-tuning, or removing? Are additional tasks/rubrics needed?

8. **Evaluate the model.** To what extent does the empirically-derived scale reflect the original conception of the proficiency continuum? Which of the proficiency levels ought to be selected as the competence threshold?

Table 3. Quality Definition Matrix for the Public Safety competency unit.

	Performance Quality Level		
Item	**Low**	**Medium**	**High**
The need for exercise is identified in consultation with stakeholders	Identify which risk management strategies will require testing by exercise.	Demonstrate communication and consultation skill with stakeholders. Determine roles, responsibilities and resource implications of involvement exercise.	Achieve and foster commitment from relevant stakeholders (financial and human resources) of involvement in exercise.
Objectives of the exercise which meet the identified need are determined	Determine the objectives of the exercise. Document objectives in clear, simple and measurable terms.	Determine pathways to achieve those objectives.	Determine context evidence required to evaluate stated objectives.
Exercise style, consistent with the objectives, is selected in consultation with stakeholders	Select the exercise style to meet stated objectives in consultation with stakeholders.	Justify the selection of exercise style to stakeholder groups	Examine the strengths and weaknesses of a range of alternative exercise styles Review and modify.
Exercise design team is assembled	Identify appropriate personnel to design and write exercise.	Assemble and brief exercise writing team and allocate tasks.	Evaluate and provide guidance to meet stated objectives.
Design exercise	Implement existing exercise formats.	Customise existing exercise formats to suit stated objectives.	Design innovative exercises to meet objectives.
Resource allocation	Identify required resources.	Justify resource allocation to stakeholder groups.	Secure resources required to implement exercise in consultation with stakeholders.
Manage exercise	Communicate aims, objectives, expectations and activity outcomes to personnel involved in exercise.	Initiate and facilitate exercise. Consult with participating personnel and relevant stakeholders on evaluation of exercise.	Monitor and review exercise plan. Provide feedback to participating personnel and stakeholders.

A critical requirement of the Melbourne School approach is that domain specialists should be responsible for producing content- and context-specific criteria for each performance quality level. Table 3, above, reproduced from Griffin (2001, p.12), illustrates just one way in which such criteria might be specified. In this example, a Public Safety unit is divided into seven items: where each item is an indicator/task associated with a particular learning outcome; and where each item is specified in terms of criteria at three quality levels: low, medium, and high.

Gillis and Griffin (2004, citing their earlier work) recommended ten rules for defining quality level rubrics, which should:

1. Reflect the **quality** of performance, ranging from low to high. They should reflect the quality of cognitive, affective or psychomotor learning that is demonstrated in the student's performance. These indicators should be ordered in terms of increasing proficiency.

2. Enable an **inference** to be made about the developmental learning. They should not simply be counts of the number of things right or wrong.

3. **Discriminate** between levels of learning and performance quality. Each recognisable different level of quality performance should be written as a quality indicator.

4. Be based on an analysis of samples of performance and the samples should cover a **diverse** range of levels of performance.

5. Be written in a language that is **unambiguous** and easily understood by all appropriate assessors. The quality indicators should be transparent and explicit in their descriptions of what is meant by increasing complexity. They should be expressed in observable terms and the use of comparative language or adjectives to differentiate between levels should be avoided.

6. Be written such that students can **verify** their own performance and understand how their performance does not fit into other coding categories.

7. Be defined by a set of quality indicators that are **developmental** in that each successive level implies a higher level of performance quality.

8. Be internally **coherent** such that they should consistently describe performance within the same developmental learning sequence.

9. Enable comparisons to be made of the performance quality **relative** to other rubrics and codes.

10. Lead to **reliable** and consistent judgements across assessors.

Finally, a critical feature of the Melbourne School approach is that the centrality of professional judgement does not exclude the implementation of complex statistical modelling. Griffin and Gillis have made extensive use of item response modelling to transform quality level judgements; enabling a degree of calibration that would otherwise require social moderation procedures (eg Bateman and Griffin, 2003; Griffin et al, 2007).

Extending their thinking on rubrics, Gillis and Griffin (2005) concluded their discussion of graded assessment in TVET by proposing a set of 10 underpinning principles:

1. the system of assessment and reporting must be situated in a **theory** of learning and assessment;
2. the procedures and assessment must satisfy both a normed and **criterion referenced** interpretation;
3. the model, approach used, assessment method, materials and decisions must be **transparent** and externally **verifiable** through a formal audit process;
4. the assessment procedure and model must be **resource-sensitive** in both development and application;
5. the model and the approach to assessment and reporting must **accommodate** the existing assessment procedures that workplace assessors have been trained to use with minimal change;
6. the rubrics, procedures and methods of design should be **accessible** to subject matter experts and not the domain of a small group of statistical experts;
7. the procedure must have both face and construct **validity**;
8. the procedure must be demonstrably fair, equitable and **unbiased**;
9. the model must be **communicative** and satisfy the information needs of stakeholders in a quality assurance context that must be accommodated; and
10. the scores and assessments are amenable to statistical and/or consensus moderation to ensure **consistency** of decisions and accuracy of score.

Comparing notes

It seems fair to conclude that the Grade Debate remained essentially unresolved throughout the 2000s, with different states and territories adopting different policies and practices (see Gillis, Clayton, and Bateman, 2009):

- system-wide grading models were introduced within Queensland, New South Wales and Western Australia;
- the Australian Capital Territory, South Australia and Victoria left grading to the discretion of individual training providers, many of which did trial grading in one form or another; yet

- graded assessment was not adopted within either Tasmania or the Northern Territory.

As these various policies and practices were rolled out, it became possible to compare notes on their relative strengths and weaknesses. For instance, Western Australia published a major evaluation of its Graded Performance Assessment model in 2002 (see Western Australian Department of Training, 2002), and another evaluation was conducted in 2005 (see Bateman and Gillis, 2005). Likewise, Queensland published a discussion paper on graded assessment to support state policy making in this area (Queensland Department of Employment and Training, 2005). Subsequently, the Australian National Quality Council commissioned Gillis et al (2009) to produce a similar report to support national policy making on grading in TVET. The two most salient, and controversial, issues to arise from these comparative exercises concerned grading criteria and principles for grading.

Grading criteria

Although the Melbourne School clearly favoured the use of specific grading criteria, others have favoured the use of generic criteria. As Bateman and Gillis put it: generic criteria "require the candidate's performance to be evaluated against a set of criteria that can be applied to performance in general regardless of the competency and/or learning area in which they are to be applied" whereas specific criteria "require the candidate's performance to be evaluated against a set of criteria that are thought to define the underlying learning or competency domain" (Bateman and Gillis, 2005, p.10). They classified major recent developments in either TVET or Senior Secondary Education as follows:

Use of generic criteria –

- WA Graded Performance Assessment model (Western Australia);
- QLD Performance Level Assessment model (Queensland); and
- VIC Certificate of Education VET (Victoria).

Use of specific criteria –

- NSW Higher School Certificate (New South Wales);
- WA Certificate of Education (Western Australia); and
- Scored Assessment approach (Australian National Training Authority).

It is important to note how the use of generic criteria, here, tends to be associated with a two-step grading process; such that only the higher grades are based upon satisfying generic criteria. The passing grade – the competence threshold – still requires the satisfaction of specific criteria, relating to professional or occupational standards. Consequently, this tends to be associated with a dual-outcome reporting process, where a candidate might be awarded 'Competent with Merit', or 'Competent with Distinction'. This disassociation between the passing grade and higher grades can be seen as a unique selling point in TVET contexts, helping to preserve the traditional characteristics of CBA, whilst also accommodating the desire for additional information concerning important differences between candidates. The 2002 WA Department of Training report reached a very positive conclusion concerning its approach to grading, based upon the use of generic criteria:

> The environmental scan clearly illustrates that the model being implemented in Western Australia is well-grounded in theory and in this respect is not dissimilar from other models both nationally and internationally. The model, like others identifies supplementary criteria which highlight underpinning knowledge together with the generic and employability skills so sought after by industry. One of the particular strengths of the model is the fact that the supplementary criteria provide greater focus on the development and reporting of the key competencies that are integral in all Training Packages. Another positive aspect is that the model appears to require fewer resources than required by alternative models.

(WA Department of Training, 2002, p.19)

Critics, on the other hand, have questioned the validity of judgements made on the basis of generic criteria. Gillis et al (2009) claimed that it was tricky to develop unambiguous and clearly differentiating generic criteria; and that this tends to result in reliance upon comparative terms (eg good vs excellent), which tends to be associated with higher levels of judgemental inconsistency. To guard against inconsistency, generic approaches may therefore require a significant investment in professional development programmes, moderation meetings, exemplar materials, etc. In other words, these expensive *maintenance* costs need to be considered alongside the relatively cheap *development* costs associated with generic criteria. Having said that, maintenance costs associated with the use of specific criteria cannot be ignored, either; and the relative cost-effectiveness of these two approaches remains to be established.

Beyond these more pragmatic criticisms, Gillis et al (2009) also identified the interpretational challenge associated with a two-step, dual-outcome, approach. Because the higher grades convey a different *kind* of information from the passing grade, it would be quite possible for a candidate awarded ‘Competent with Distinction’ to be less competent (in terms of their domain-specific learning trajectory) than a candidate awarded ‘Competent with Merit’. That is, the Distinction candidate might have stronger generic competencies, but weaker specific competencies. This presents a significant communication challenge, and presents a major threat to the accurate interpretation of higher grades, derived from generic criteria.

Grading principles

One of the most interesting features of the Australian literature is its quest to identify underlying principles. As noted above, Gillis and Griffin (2005) identified 10 such principles, starting from their premise that any system of assessment and reporting must be situated in a theory of learning and assessment. Yet, quite different sets of principles have been proposed. Indeed, Bateman and Gillis (2005) included a detailed comparative analysis of such principles within their evaluation report. Table 4 illustrates three sets of principles; from Rumsey (1997), Williams and Bateman (2003), and the Queensland Department of Employment and Training (2005).

There are plenty of similarities between the sets of principles presented in Table 4. For instance, they all agree that graded assessment must be: criterion-referenced (R1, WB1, Q2); transparent (R6, WB3, Q2); and two-stepped (R8, WB2, Q5). Yet, it is interesting to note how each successive contributor disagreed sufficiently with previous sets, to feel the need to propose a new one.

An important distinction can be drawn, here, between TVET-specific grading principles and those that might apply to grading in any context; indeed, to assessment in any context. Presumably, expectations related to purpose, cost-effectiveness, transparency, consistency, validity, reliability, fairness, feasibility, and quality assurance would all apply to any assessment context. And, clearly, the majority of principles from Table 4 fall into this category. Criterion-referencing, on the other hand, is not a necessary feature of educational assessment, per se; so this would seem to be a TVET-specific grading principle, albeit not limited to TVET contexts. The idea of implementing a two-step grading process, and the idea of making grading optional, are also TVET-specific, and more unique to TVET contexts.

Table 4. Examples of proposed grading principles

Rumsey	Williams and Bateman	Queensland DET
1. There must be a clearly identified need and purpose for the reports.	1. Grading must be criterion-referenced, as opposed to norm-referenced.	1. There must be a clear need and purpose for graded assessment.
2. The grading criteria must be defined and meaningful.	2. Grading must be applied only once competence is determined.	2. Graded assessment must be criterion-referenced, reflect good assessment practice and align with existing Standards: based on public and transparent criteria; consistent with the principles of validity, reliability, flexibility and fairness; utilise measurable assessment data; transparent.
3. The assessment data collected and used for any grading must be measurable.	3. The grading system must be transparent to all participants and stakeholders.	3. Graded assessment must be cost effective.
4. The assessment process involved must be feasible, valid, reliable and fair.	4. Grading must be discretionary, allowing candidates to opt out.	4. Graded assessment must be applied once competency is determined.
5. The overall assessment and related reporting processes must be cost-effective.		5. Graded assessment must be conducted and reported consistently.
6. The assessment and related reporting process must be transparent to all involved, including students, employers, trainers, assessors and others with an interest in the assessment outcomes.		6. Grading must be underpinned by quality assurance measures designed to produce quality and consistent assessment outcomes.

Rumsey	Williams and Bateman	Queensland DET
7. There must be consistency in the way the grading and reporting is conducted across the relevant, enterprise(s), industry, multiple industries, or client groups involved.		7. Graded assessment should be available to all learners, in all qualifications but be optional for the learner.
8. Supplementary grading/reporting processes must not compromise or confuse the competency-based reporting of assessment reporting.		

Reviewing the sets presented in Table 4, alongside other sets, Bateman and Gillis (2005) suggested that certain of them – including implementing a two-step process and making grading optional – were better described as 'business rules' rather than principles. According to this analysis, only one of the three TVET-specific principles mooted above would be considered a genuine principle – criterion-referencing. Of course, criterion-referencing is also promoted in many non-TVET contexts. So this raises the question of whether the search for TVET-specific grading principles is even appropriate (Gillis et al, 2009).

Discussion

Despite having remained fairly positively disposed to grading CBAs since the early 1990s, even Australia is yet to establish a coherent national policy, let alone to identify and promulgate a generally accepted set of 'good practice' principles.

Toward the middle of the 1990s, the Thomson report identified a wide variety of practices in operation, tending to be based upon idiosyncratic interpretations of underlying principles and assumptions. Williams and Bateman reached essentially the same conclusion during the early 2000s. Just recently, a report from the Australian Government Productivity Commission (AGPC, 2017) observed that, although state-wide systems had been piloted in the early 2000s, the policy focus had then shifted to ensuring the quality of vocational courses and teaching, with grading no longer an emphasis. Having said that, the same report concluded that grading in TVET contexts had the potential for enhancing economic performance, and therefore identified it as a priority concern.[4]

Although still essentially unresolved, the Grade Debate in Australia has resulted in a small, but important, body of work on grading CBAs, from which other countries can benefit. In the following sections, I attempt to synthesise some it its most important insights under four headings – criteria, evidence, theory, and principle – offering a brief comment upon each theme.

Criteria

The Australian literature identifies a wide variety of potential bases for awarding grades, classified according to a number of different schemes. It seems possible to distil those schemes into the following six meta-criteria of grade-worthiness:

1. **diligence** during course of **learning** (eg wide reading, outstanding attitudes, consistently high motivation);
2. **diligence** during **assessment** process (eg quality of presentation, meeting submission deadlines, commitment to passing first time);
3. **calibre of performance** during course of **learning** (eg speed of acquisition of learning outcomes, number of additional learning outcomes acquired);
4. **calibre of performance** during **assessment** process (eg level of supervision/autonomy, speed of performance, consistency of performance, number of errors made);
5. **level of proficiency** in relation to **generic learning outcomes** acquired (eg originality, creativity, adaptability, planning, information handling); and
6. **level of proficiency** in relation to **specific learning outcomes** acquired (ie in relation to the learning outcomes specified for each unit).

[4] See Recommendation 3.2 – the Australian Government should develop tools for proficiency-based assessment.

Presented thus, the first two are essentially inputs to learning and assessment; whereas the last two are explicitly outcomes from learning. The middle two lie somewhere in between. These six criteria resonate strongly with the classic folk psychology equation: **attainment = effort x ability**.

In terms of the above criteria: diligence resonates with effort; calibre resonates with both ability and attainment; while proficiency resonates with attainment. In fact, attainment is synonymous with proficiency, as this term is used above; and effort seems to be tantamount to diligence. Ability seems to reflect certain of the examples located under the calibre criteria. For instance, the idea that speed of acquisition might constitute a learning calibre criterion seems to be based on the assumption that faster learning implies a greater ability to learn in the domain in question. Yet, for other examples located under the calibre criteria, the comparison seems closer to attainment than ability; particularly the examples within Criterion 4, which identify features of performance that might be associated with superior expertise. Finally, although Criterion 5 is defined, here, in terms of attainment, ie acquired proficiency, it may be a moot point as to whether characteristics such as originality, creativity, and adaptability are better understood in terms of ability. In other words, might they be better understood as personality traits rather than as learning outcomes?

As noted earlier, the Melbourne School has claimed that it is difficult to defend, within a competence assessment framework, the use of unmeaningful grading criteria, such as meeting (or failing to meet) submission deadlines. From this perspective, only criteria with the potential to differentiate between learners in terms of their underlying levels of competence can be considered meaningful, and therefore defensible (Gillis and Griffin, 2005). Yet, it is quite possible to adopt an alternative perspective, by associating higher grades with a quite different kind of meaning. Indeed, it seems quite possible to argue in favour of awarding higher grades on the basis of any of the six meta-criteria presented above; particularly when operating the kind of two-step, dual-outcome, model discussed earlier.

For instance, it does not seem unreasonable to suggest that the primary concern of an employer, when selecting between applicants for a job, is to screen-out anyone who lacks sufficient competence to practise. This kind of information is still provided by the passing grade, under a two-step, dual-outcome, model. However, once the screening purpose has been achieved, that same employer might prefer to achieve the selection purpose on some basis other than proficiency. For instance, when required to choose between one competent applicant and another, they might prefer to do so on the basis of information concerning their diligence; perhaps on the 'tortoise vs hare' assumption that a diligent plodder will eventually acquire greater competence than an indifferent whizz.

To facilitate this, the Pass grade, for the occupational or professional qualification in question, might be determined on the basis of a level of proficiency in relation to specific learning outcomes (Criterion 6), while the Merit grade might be determined on the basis of a level of diligence during the course of learning (Criterion 1). The key point, here, is that information provided by the Merit grade (regarding diligence) would be of a different *kind* from information provided by the Pass grade (regarding proficiency). As such, diligence would not be an *un*meaningful criterion, per se, but a *differently* meaningful criterion.

Evidence

When higher grades are to be determined on the basis of acquired learning outcomes, eg Criterion 5 or 6 above, the most important practical consideration is whether this will be based upon evidence already collated for the passing grade, or whether it will require a bespoke evidence-gathering exercise. As explained by contributors to the Grade Debate, separately gathered evidence might include work deemed particularly relevant to the grading criteria in question, such as:

- **core skills** units, designed to assess generic learning outcomes; or
- **integrating** units, designed to assess specific learning outcomes synoptically.

Alternatively, it might take the form of work that is easier to grade, such as:

- conventional **knowledge tests**.

When higher grades are to be determined on the basis of evidence already collated for the passing grade, this still leaves open two important questions concerning whether that evidence is to be judged: holistically, or atomistically; at unit level, or at qualification level. A variety of alternatives might be envisaged, for instance:

1. **unit grading**, with an **atomistic** grading judgement for every single assessment criterion (judgements somehow aggregated to the unit grade);
2. **unit grading**, with an **atomistic** grading judgement for any assessment criterion deemed amenable to grading, excluding other assessment criteria (judgements somehow aggregated to the unit grade);
3. **unit grading**, with an **holistic** grading judgement concerning the corpus of evidence produced for the unit, based upon **unit-specific** grading criteria;
4. **unit grading**, with an **holistic** grading judgement concerning the corpus of evidence produced for the unit, based upon **unit-generic (qualification-specific)** grading criteria;
5. **unit grading**, with an **holistic** grading judgement concerning the corpus of evidence produced for the unit, based upon **unit-generic (qualification-generic)** grading criteria;
6. **qualification grading**, with an **holistic** grading judgement concerning the corpus of evidence produced for the qualification, based upon **qualification-specific (unit-generic)** grading criteria; or
7. **qualification grading**, with an **holistic** grading judgement concerning the corpus of evidence produced for the qualification, based upon **qualification-generic (unit-generic)** grading criteria.

The Melbourne School has tended to shun generic criteria, questioning the consistency of grading judgements based upon them; and even questioning the very idea of attainment/proficiency (cf. ability/aptitude) in relation to generic learning outcomes (Gillis et al, 2009; McCurry, 2003). Yet, the idea of generic criteria remains at least superficially attractive, when considered in relation to a suite, or system, of qualifications of a certain type: potentially drastically reducing development and implementation costs, if only a single set of criteria is required; as well as potentially allowing equivalent grades, from different qualifications, to be interpreted in the same way.

Theory

The fact that Royce Sadler has spent his professional career working on assessment challenges in Queensland helps to explain the impact that his idea of Standards-Referenced Assessment (SRA) has had on the literature on grading CBAs. As Maxwell and the Melbourne School have emphasised, thinking about learning in TVET as a trajectory, or continuum, rather than as a series of binary acquisitions, helps to provide a plausible theoretical basis for grading.[5] Standards, which can be described and exemplified, provide a common yardstick against which to classify learners' performances, and hence their degree of progression from relative novice to relative expert within the domain of learning. Consequently, this theoretical perspective strongly recommends Criterion 6, level of proficiency in relation to specific learning outcomes acquired.

Although Maxwell and the Melbourne School both recommended SRA as a theoretical basis for grading practices, they differed somewhat in their characterisation of grading judgements. Maxwell's repeated reference to standards-based assessment seemed to foreground the centrality of holistic judgement against holistic criteria, which relates directly to the idea of competence as a complex, holistic, characteristic. This is entirely in keeping with Sadler's model of SRA (eg Sadler, 1987; Sadler, 2009a). The Melbourne School, however, seems more open to approaches that reflect the manner in which competence has traditionally been assessed via CBA, ie criterion/indicator-by-criterion/indicator. As such, learners might still be assessed atomistically, as long as multiple quality levels are specified for each criterion/indicator. This approach does not preclude the specification of holistic criteria. However, they would be developed on the back of atomistic criteria, to synthesise them; and they would not constitute the primary basis for assessment judgements.

Returning to the debate over the legitimacy of 'unmeaningful' criteria, Sadler, in his work on grading in higher education institutions, has taken a strong line on the necessity of restricting grading judgements purely to evidence concerning acquired learning outcomes:

> If a grade is to be trusted as an authentic representation of a student's level of academic achievement, one of the requirements is that all the elements that contribute to that grade must qualify as achievement, and not be something else.

[5] It also raises an interesting, and potentially consequential, question of whether the competence threshold – traditionally the passing grade – ought necessarily to be specified as the lowest grade. If the competence threshold is conceived as nothing more than a point along a continuum of proficiency, then there is no logical reason to preclude grades below that competence threshold, each representing a significant amount of progress towards competence (see Maxwell, 2010, for instance).

(Sadler, 2009b, p.727)

Conversely, he has identified all sorts of 'unmeaningful' factors that often contribute to grading judgements in higher education, which may even be specified in regulations as a matter of institutional policy. These include factors that operate either via:

- **transactional credits**, eg marks added for the completion of specified activities (practice exercises, log books, reflective journals, interim drafts, etc.); or via
- **transactional debits**, eg marks deducted for the late submission of assignments, or for exceeding maximum word length.

He, too, has dismissed the relevance of effort to grading, on the basis that effort is an input to learning, not an outcome from learning, and must therefore be excluded from consideration. His preference is clearly for Criterion 6.

Of course, as Sadler formulates the grading challenge, he is entirely correct: if a grade is to be trusted as an authentic representation of a student's level of proficiency, then it ought to be based exclusively upon proficiency-related evidence. However, as noted earlier, the possibility still exists that users might be interested in a different kind of information, perhaps related to an entirely different characteristic, such as diligence; particularly when competence is already indicated by the passing grade.

In addition, it is worth noting that Sadler advocates a 'purist' approach to assessment design, which respects only to the **information perspective** on assessment purposes. In other words, he assumes that the sole purpose of the assessment is to provide a certain kind of information; namely, information concerning each candidate's level of acquired proficiency. An alternative perspective on assessment purposes, the **engagement perspective**, recognises that educational assessments are typically designed at least partly to promote engagement with a course of learning – to motive and direct both learners and their teachers – and that assessment design decisions ought to be made with engagement-related purposes in mind, as well as information-related ones.[6] Sometimes, assessment policy makers are prepared to tolerate a certain amount of information contamination – allowing grades to be swayed by factors other than acquired proficiency – for the 'greater good' of enhancing teaching and learning.

Finally, the suggestion that grades ought to be based purely upon level of proficiency in relation to acquired learning outcomes is potentially challenging in the context of CBA; specifically, in the context of learning that is not circumscribed by a course of fixed duration. We can unpack this challenge using the classic folk psychology equation mentioned earlier, which might be expressed, more comprehensively, as: **attainment = effort x ability x time**.

When time is held constant – as it is for qualifications that presume a course of fixed duration – attainment is tantamount to effort x ability. From a folk psychology perspective, this combination of effort x ability is often presumed to capture an individual's **aptitude** for learning. This helps to explain why examination results,

[6] See Newton (2017) for a thorough discussion of 'purpose purism' versus 'purpose pluralism' in assessment design.

which ostensibly represent a level of already-acquired proficiency, are often used for selective purposes. In other words, result users will often treat a certain level of attainment as though it indicated a certain level of aptitude, which is presumed to indicate a learner's potential for acquiring proficiency in the future.

However, when time is not held constant, attainment is no longer tantamount to effort x ability (ie aptitude) according to this folk psychological analysis. For example, imagine that two apprentices certificated at the same time. The first one 'cashed in' for a Pass after two yeas of work-based learning; which, for the sake of this argument, was the conventional duration of an apprenticeship in her occupational field. Conversely, the second one, who was a slower learner, 'cashed in' for a Merit after four years of work-based learning.[7] In one sense, of course, their grades are directly comparable; the slower learner has genuinely attained a higher level of proficiency. In another sense, though – the sense given by our folk psychological analysis – the two grades are less comparable than if the two apprentices had studied for the same amount of time. Our folk psychologist might argue that the slower learner, despite her higher grade, demonstrated less aptitude.

In fact, time is not necessarily the only additional factor that an employer might wish to introduce to the folk psychology equation. They might also like to bear in mind attainment-at-time-zero; that is, level of attainment when the period of study commenced. Imagine, for instance, two 21-year-old carpentry apprentices, who both 'cashed in' for a Merit after two years of work-based learning. One had started the apprenticeship with academic A levels, but with no experience of carpentry. The other had started after already having completed a two year diploma in carpentry. At the end of their apprenticeships, were they both equally meritorious, and both equally worthy of selection for a carpentry job? This is a rhetorical question, intended to unpick unstated assumptions regarding the meaning and utility of grading. However, exactly this situation arises in relation to the Accreditation of Prior Learning (APL), perhaps raising questions of fairness concerning the integration of APL and grading (Williams and Bateman, 2003).

Again, questions like these are rhetorical, but they are far from trivial. Grading only makes sense when there is some likelihood that groups of candidates within a qualification cohort will differ significantly, in one way or another, so as to justify the award of different grades. Understanding the ways in which groups of candidates can be expected to differ – including the variety of possible explanations for, and implications of, those differences – is fundamental to establishing a defensible approach to grading. A defensible grading approach, from this perspective, is one for which it is possible to construct a strong argument that it is both sufficiently meaningful and sufficiently useful.

Principle

Although many of the Australian reports have considered principles of good practice explicitly, this literature does not actually develop the idea of TVET-specific grading principles at all well. Indeed, the majority of principles within most of the proposed sets are neither TVET-specific nor even grading-specific. Instead, they are simply basic principles of assessment; like validity, consistency, and cost-effectiveness.

[7] Assuming, here, that Merit indicates a higher level of *proficiency* than Pass.

Similarly, although this literature generally recognises criterion-referencing as a core principle for grading CBAs, criterion-referencing is hardly unique to TVET contexts.

In terms of underlying principles, there does appear to be a significant difference between the stance adopted by the Melbourne School and that adopted by various other Australian commentators. This relates to the idea of whether grading ought to be conceptualised as somehow 'bolted-on' to CBA. The principle of a two-step grading process, and the principle of making grading optional, both embody this 'bolted-on' conception. Consequently, these principles (or rules) are actually independent of grading practices; in the sense that they say nothing about how higher grades ought to be awarded. They simply require that, whatever grading approach is adopted, it should not interfere with nor detract from the primacy of the competency judgement that precedes it.

The position advocated by the Melbourne School directly challenges the 'bolted-on' conception. It suggests that the competency judgement is no different, in kind, from the judgement required for higher grades. Instead, all of the grades awarded for a qualification should be understood in terms of a set of proficiency bands, arranged hierarchically to represent positions along a learning continuum.

In terms of grading practices, different implications might be drawn from this position, depending upon how radically the Melbourne School critique is interpreted. The least radical interpretation would assume that grading practices for higher grades ought to be continuous with grading practices for the passing grade, whilst modelling higher grade practices upon practices already in operation for the passing grade. Thus, if CBA requires atomistic specification, mastery measurement, and exhaustive sampling, for the award of the passing grade, then it should also require exactly the same for the award of higher grades. The most obvious way to achieve this would be to specify Merit and Distinction criteria for every single Pass criterion.[8] The award of Merit (or Distinction) overall might then be contingent upon having achieved at least a Merit (or Distinction) on every single criterion.

The most radical interpretation would assume, once again, that grading practices for higher grades ought to be continuous with grading practices for the passing grade. However, interpreted more radically, the idea of breaking down the 'bolted on' conception might be treated as a critique of the traditional CBA model itself. Recall the distinction that Maxwell (1997a) drew between CBA as a complex, holistic, ideal (as originally intended by key Australian stakeholders), and CBA as a simpler, more atomistic, reality (as it ultimately came to be practised). The most radical interpretation, therefore, might argue for a more complex, holistic, approach to judging both higher grades and the passing grade. This would certainly be in keeping with the spirit of SRA, as described by Sadler (1987; 2009a), as well as being in keeping with the spirit of CBA, as described by Hager and Gillis (1995).

In practice, the Melbourne School seems to have left open a third approach: accepting that grading practices might operate quite differently between the passing grade and higher grades; yet, still insisting that the grades should be interpreted as points along a trajectory of learning. This approach seems to reflect Principle 5, from Gillis and Griffin (2005): the approach must accommodate existing assessment

[8] Or, at least, for every criterion that was amenable to grading above the competence threshold.

procedures that workplace assessors have been trained to use with minimal change. (Note that this claim would seem to be more pragmatic than principled.)

Bearing in mind the lack of genuinely TVET-specific grading principles arising from the Australian literature, the present report will conclude with nothing more than a very high-level statement of principle for grading in TVET contexts:

1. The grounds for differentiating between candidates, via grades, must be defensible; that is, sufficiently **meaningful** and sufficiently **useful**, when judged in relation to a **profile of purposes**.[9]

2. The grading process must be sufficiently **accurate**.

3. The **benefits** from implementing the grading process must, on balance, **outweigh** its **costs**.

These are very general principles, which could be applied to any educational assessment process or procedure. In the present context, they specify that it is not enough that any particular grading practice is capable of differentiating between candidates; nor even that it is capable of differentiating between candidates reliably. Instead, grading practices need to be capable of differentiating purposively, accurately, and in a manner that is economically, politically, and socially acceptable.

[9] Information-related purposes, engagement-related purposes, etc. (see Newton, 2017).

References

Anderson, L.W. (2018). A critique of grading: Policies, practices, and technical matters. *Education Policy Analysis Archives*, 26 (49), 1–31.

Andre, K. (2000). Grading student clinical practice performance: the Australian perspective. *Nurse Education Today*, 20 (8), 672–679.

Australian Government Productivity Commission (2017). *Shifting the Dial: 5 Year Productivity Review. Inquiry Report No. 84*. Canberra, ACT: Commonwealth of Australia.

Bateman, A. and Gillis, S. (2005). *Review of the Graded Performance Assessment Model*. Report submitted to the Western Australian Department of Education and Training.

Bateman, A. and Griffin, P. (2003). The appropriateness of professional judgement to determine performance rubrics in a graded competency based assessment framework. Paper presented at the Joint AARE/NZARE Conference. Auckland, New Zealand. November 29 – December 3.

Bloom, B.S. (1968). Learning for mastery. *UCLA Center for the Study of Evaluation of Instructional Programs: Evaluation Comment*, 1 (2), 1–12.

Boahin, P. (2018). Competency-based assessment and reporting in Ghanaian polytechnics: A critique of the prevailing perceptions. *Journal of Education and Practice*, 9 (5), 131–140.

Brookhart, S.M., Guskey, T.R., Bowers, A.J., McMillan, J.H., Smith, J.K., Smith, L.F., Stevens, M.T., and Welsh, M.E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, 86 (4), 803–848.

Cedefop (2015). *Ensuring the Quality of Certification in Vocational Education and Training. Cedefop research paper, No 51*. Luxembourg: Publications Office.

Gillis, S. and Griffin, P. (2004). Using rubrics to recognise varying levels of performance. *Training Agenda: A Journal of Vocational Education and Training*, 12 (2), 22–24.

Gillis, S. and Griffin, P. (2005). Principles underpinning graded assessment in VET: A critique of prevailing perceptions. *International Journal of Training Research*, 3 (1), 53–78.

Gillis, S., Clayton, B. and Bateman, A. (2009). *Scoping Current Research and Practices in Graded Assessments*. Report submitted to the Australian National Quality Council.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18 (8), 519–521.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist*, 36 (9), 923–936.

Gonczi, A. (1994). Competency based assessment in the professions in Australia. *Assessment in Education: Principles, Policy & Practice*, 1 (1), 27–44.

Griffin, P. (2001). Using indicators of quality to infer competence. Paper presented at the National ACACA Conference. Sydney, Australia. July 25 – July 27.

Griffin, P. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation*, 33 (1), 87–99.

Griffin, P. and Gillis, S. (2000). A multi source measurement approach to assessment of higher order competencies. Paper presented at the BERA Conference. Cardiff, UK. September 7 – September 10.

Griffin, P., Gillis, S. and Calvitto, L. (2007). Standards-referenced assessment for vocational education and training in schools. *Australian Journal of Education*, 51 (1), 19–38.

Hager, P. and Gillis, S. (1995). Assessment at higher levels. In W.C. Hall (Ed.). *Key Aspects of Competency-Based Assessment* (pp.59–71). Adelaide, SA: National Centre for Vocational Education Research.

Hager, P., Athanasou, J. and Gonczi, A. (1994). *Assessment Technical Manual*. Canberra, ACT: Australian Government Publishing Service.

Jessup, G. (1991). *Outcomes: NVQs and the Emerging Model of Education and Training*. London: The Falmer Press.

Johnson, M. (2008). Grading in competence-based qualifications – is it desirable and how might it affect validity? *Journal of Further and Higher Education*, 32 (2), 175–184.

Marzano, R.J. (2000). *Transforming Classroom Grading*. Alexandria, VA: Association for Supervision and Curriculum Development.

Maxwell, G.S. (1997a). Competency based assessment and tertiary selection: Background context and issues. *Queensland Journal of Educational Research*, 13 (3), 4–15.

Maxwell, G.S. (1997b). Future directions for competency based assessment. *Queensland Journal of Educational Research*, 13 (3), 71–84.

Maxwell, S. (2010). *Using Rubrics to Support Graded Assessment in a Competency-Based Environment.* Adelaide, SA: National Centre for Vocational Education Research.

McCurry, D. (2003). But will it work in theory? Theory, empiricism, pragmatics and the Key Competencies: The place of theory and research in the development of a notion of work related skills and the Whole School Assessment of Generic Skills. Melbourne, VIC: Australian Council for Educational Research.

Newton, P.E. (2017). There is more to educational measurement than measuring: The importance of embracing purpose pluralism. *Educational Measurement: Issues and Practice*, 36 (2), 5–15.

Newton, P.E. (2018). *Grading Vocational & Technical Qualifications: Recent policies and current practices*. Coventry: Office of Qualifications and Examinations Regulation.

Peddie, R.A. (1997). Some issues in using competency based assessment in selection decisions. *Queensland Journal of Educational Research*, 13 (3), 16–45.

Queensland Department of Employment and Training (2005). *Graded Assessment in Queensland: A discussion paper*. Brisbane, QLD: Queensland Department of Employment and Training

Rumsey, D. and Associates (1997). *Reporting of Assessment Outcomes within Competency Based Training and Assessment Programs under New Apprenticeships (formerly Modern Australian Apprenticeship and Traineeship System)*. Brisbane, QLD: Australian National Training Authority.

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

Sadler, D.R. (2009a). Indeterminacy in the use of preset criteria for assessment and grading. Assessment & Evaluation in Higher Education, 34 (2), 159–179.

Sadler, D.R. (2009b). Grade integrity and the representation of academic achievement. *Studies in Higher Education*, 34 (7), 807–826.

Thomson, P., Mathers, R. and Quirk, R. (1996). *The Grade Debate: Should we grade competency-based assessment?* Adelaide, SA: National Centre for Vocational Education Research.

Western Australian Department of Training (2002). *Graded Performance Assessment In a Competency Based Environment: An environmental scan and analysis*. Perth, WA: Western Australian Department of Training.

Williams, M. and Bateman, A. (2003). *Graded Assessment in Vocational Education and Training: An analysis of national practice, drivers and areas for policy development.* Adelaide, SA: National Centre for Vocational Education Research.

Wilmut, J. and Macintosh, H.G. (1997). Possible options for differentiation in competency based assessment. *Queensland Journal of Educational Research*, 13 (3), 46–70.

Wolf, A. (1993). *Assessment Issues and Problems in a Criterion-Based System*. London: Further Education Unit.

Yorke, M. (2008). *Grading Student Achievement in Higher Education: Signals and shortcomings*. Oxford: Routledge.

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

0300 303 3344
public.enquiries@ofqual.gov.uk
www.gov.uk/ofqual

November 2018 **Ofqual/18/6441/2**