# Principles underpinning graded assessment in VET: a critique of prevailing perceptions

*Shelley Gillis and Patrick Griffin*

## Abstract

*To help achieve national consistency of assessment and reporting in the Australian Vocational Education and Training sector, there is a need to develop a set of national principles for graded performance assessment. This paper challenges a number of prevailing principles from both a theoretical and assessment perspective, namely that grades must be criterion referenced (Williams & Bateman, 2003), meaningful (Rumsey, 2003) and applied once competence has been achieved (Williams & Bateman, 2003). This paper argues that the use of generic criteria cannot be defended in terms of their validity and reliability and that a clear understanding of the underlying developmental continuum of learning is required to inform the development of meaningful and valid criteria and descriptors of quality performance. Finally, the paper proposes a set of principles that have been grounded in theory, have been put to the test in large-scale research, and are consistent with international literature on competence assessment.*

## Introduction

Since the introduction of competency based training and assessment into the Australian Vocational Education and Training (VET) sector in the early 1990s, the way in which the outcomes of assessment should be reported has been a contentious issue amongst researchers, policy makers and practitioners. In the early 1990s, debate typically focused on whether the principles that underpinned competency based assessment implied only one acceptable standard of performance or whether graded assessment was possible within a competency based assessment framework. During that period, most competency training and assessment arrangements reported on a dichotomous (two levels) scale (i.e., competent or not yet competent) (Rumsey 1997). Whilst Registered Training Organisations (RTOs) comply in that they assess and report competent/not yet competent decisions, there has recently been a movement to also assess and report varying levels of performance (Smith, 1996; Thomson, Mathers & Quirk, 1996; Williams & Bateman, 2003).

Although there is wide variation in the processes used to recognise and report levels of performance, the term "graded assessment" has been loosely used to encompass all practices and models currently used or proposed in the VET sector where differentiated levels of performance are recognized and reported (Schofield & McDonald, 2004; Williams & Bateman, 2003). What is missing in the debate and in the development of the grading approach is any recognition of the nature of knowledge or learning involved and how that knowledge is acquired and developed. The lack of any reference to the nature of learning in a vigorous educational debate, at times led by educators, is astonishing. The dogged arguments about grading or the dichotomy have treated it purely as a political or policy related matter, devoid of any consideration of the underpinning theories of learning and development or of educational assessment and measurement.

A wide variation in graded assessment and reporting practices has evolved in the Australian VET system (Griffin, Gillis, Keating & Fennessy, 2001; Schofield & McDonald, 2004; Smith, 2000; Williams & Bateman, 2003). Schofield and McDonald (2004), in the "high level review of training packages", recommended that policy be developed at the national level to address the issue of grading. Specifically, they argued that

> *Coordination and leadership on this issue at a national level is, [however], overdue, and we suggest that ANTA investigate the range of graded assessment models currently being implemented across Australia with a view to developing a model that allows for grading assessments to be provided with Training Packages as supplementary reports.*
>
> (Schofield and McDonald, 2004, p.19).

To help achieve consistency of assessment and reporting, some states have developed or are intending to develop grading models (QLD, WA, SA and NSW) whilst others leave it to the discretion of the RTO (ACT and VIC).

Grading is possible, and desirable, within a competency based vocational education and training system (Griffin et al., 2001; Smith, 2000; Williams & Bateman 2003). However, most debates of recent times have tended to focus on the practicability issues associated with grading such as:

- its purpose (e.g., whether it is appropriate to be used for recognition of prior learning or whether it should be limited to summative assessment?),

- the context (e.g., should it be restricted to off-the-job assessments?),

- the associated costs (e.g., development versus implementation),

- the number of performance levels to be assessed and reported (e.g., should this be standardized at a national, state and/or industry level?) and

- the nature of the grading criteria (e.g., content versus specific?) (Williams & Bateman, 2003).

Rumsey (2003) and Williams and Bateman (2003) proposed sets of principles that they considered should underpin graded assessment models for vocational education and training. Rumsey (1997) was the first to attempt the articulation of principles. He argued that in competence assessment:

- there must be a clearly identified ***need*** and ***purpose*** for the reports;
- the grading criteria must be ***defined*** and ***meaningful***;
- the assessment data collected and used for grading must be ***measurable***;
- the assessment process involved must be ***feasible, valid, reliable*** and ***fair***;
- the overall assessment and related reporting processes (including both on- and off-the-job aspects) must be ***cost-effective***;
- 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;
- 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; and
- supplementary grading/reporting processes must ***not compromise*** or ***confuse*** the competency based reporting of assessment outcomes (i.e., qualifications and lists of competency units achieved).(Rumsey, 1997, p.6)

Similarly, Williams and Bateman (2003) developed a set of eclectic principles that were derived from an analysis of current practice and dominant perceptions. They argued that grading should have specific characteristics, which appealed to ideological stances current at the time of their paper perhaps as a diplomatic and compromise position. Grading was required to be:

- criterion referenced;
- applied once competence is determined;
- transparent; and
- discretionary.

This article sets out to discuss a number of these principles from both a theoretical and assessment perspective. In particular, the following three principles are discussed:

- grades must be criterion referenced (Williams & Bateman, 2003);
- the grading criterion must be defined and meaningful (Rumsey, 2003); and
- applied once competence has been achieved (Williams & Bateman, 2003)

Rumsey's (2003) measurement principle (i.e., the assessment data collected and used for grading must be measurable) will not be reviewed in this paper, as his intent for this principle is uncertain given that data cannot be measured. Furthermore, as the remaining principles are not unique to the graded assessment debate in the sense that they are perceived as important and relevant to all assessments (e.g., transparency, fairness, validity, reliability and cost effectiveness), they are less contentious, and in need of less urgent debate, than the three principles listed above. In exploring these principles, the article reviews a number of models that are currently being implemented or proposed for implementation in competence assessment. Finally, the paper proposes a set of principles that have been grounded in theory, have been put to the test in large-scale research, and are consistent with international literature on competence assessment.

## Criterion referenced (Williams & Bateman, #1)

Whilst all forms of assessment, whether based on competence or curriculum models, employ similar techniques to gather evidence (Hager, Athanasou & Gonczi, 1994; Hall & Saunders 1991), differences emerge in the way evidence is interpreted. It is argued that competence assessment is different to other forms of assessment. For example, in curriculum models the evidence is typically interpreted in either normative or criterion referenced frameworks (Foyster, 1990). However, in competence assessment, it is common to argue that interpretation should be limited to a criterion referenced framework (CSB Assessors & Workplace Trainers, 1993; NTB, 1992). Rarely is any rationale for this distinction defended in any way other than a statement that competence based education is curriculum free in that it focuses only on the demonstration of the competence, not the process of acquisition. Apart from this argument being a non sequitur, it underlines the persistent pattern in the debate to ignore the nature of learning involved or the interpretation frameworks for assessment evidence.

Historically, norm referenced frameworks have been the predominant means of interpreting the result of educational assessments and, in particular, those that have reported varying levels of achievement in the form of grades (William, 2000). When using a norm referenced framework, an individual's performance is compared to the average or expected performance of a more or less well defined group of individuals (Griffin & Nix, 1991). Letter grades (e.g., A, B, C, D etc) are often determined by the standardization of scores to represent a “bell curve” distribution. They are used to represent the nature of the group and to differentiate between members of sub groups. In every case, it demands a definition of the group before any sense can be made of the normative assessment or of the grades that go with it. It provides opportunities for relative (as opposed to absolute) interpretations of individuals’, sub groups’ and whole group performances, but it does not allow for any substantive interpretation of the grade (Griffin, Gillis & Calvitto, 2004). Consequently, there is typically no direct indication of the kinds of knowledge, skills and understandings that have been acquired by an individual (Hager et al., 1994). Knowledge type, skill definition and contextual understandings are ignored in a norm referenced graded approach to reporting performances, and no single population can be regarded as the definitive normative group (Murphy & Davidshofer, 1989). The same raw score or letter grade can produce a wide range of interpretations, depending on which group is chosen and how the sub groups are divided and labeled with the letter codes to identify the sub groups. Normative scores or grades *cannot* be used to:

- establish and test substantive benchmarks;
- provide an adequate basis for monitoring individual growth or development;
- identify learning difficulties (i.e., for purposes of diagnostic assessment) and hence help in developing training or intervention plans; and
- identify areas where improvements in learning are required (i.e., for purposes of formative assessment).

(Masters, 1993; William, 2000).

Normative scores have no substantive or absolute interpretive value. They can help to sort candidates for purposes of selection or any other differentiating purpose. For this reason, norm referencing is often the default interpretation framework for differentiating among candidates and the simplest form of this is to divide the distribution into intervals and call the labels assigned to those intervals - ’grades’. In a competence system, it is widely held that this form of recording and reporting is inappropriate. There are few legitimate reasons for the use of normative grading in a competence system; unless it is first derived from a criterion referenced interpretive system. But the normative interpretation cannot be dismissed and certainly cannot be banned. Expectations (or norms) have a central role to play in establishing standards (Peddie, 1997). If normative grading is replaced by a criterion referenced or standards referenced framework for interpreting performance data, a new and illuminating approach to competency reporting emerges.

It is a widely held misconception that grading is only possible in a norm-referenced system, but this only applies to grading governed by the distribution of scores. It is possible to recognise varying levels of performance within a standards or criterion referenced framework without relying on a norm referenced system. However, standards referenced approaches demand firstly that we recognize competence and report it as performance beyond the minimum “standard of performance required in employment” (NTB, 1992, p. 10). Standards referencing also demands that criterion referencing is technically and correctly understood and applied in a competence assessment. Criterion referencing is ".. the development of procedures whereby assessments of proficiency could be referred to stages along progressions of increasing competence" (Glaser, 1981; p.935). In this approach to interpretation, an individual's performance is compared with descriptions of stages on a scale of increasing competence, thus allowing the performance to be positioned along a developmental continuum. Wolf argued that,

> *…there is nothing about criterion referenced testing which ties it to a pass/fail, on-off approach. Criterion referenced assessment produces a distribution of performance…a single pass-fail is ONE way to partition that distribution but only one.*
>
> (Wolf, 1993, p.13)

A criterion referencing framework differs from norm referencing frameworks in at least the following ways:

- interpretation of the performance can only be carried out in a criterion referenced framework. It cannot be interpreted in a norm referenced framework. Absolute measures are used to interpret performance in criterion referencing as opposed to relative measures in norm referencing;
- there is no a priori distribution of scores across the grade levels and it is possible for *all* students to be performing at the highest possible level in a criterion referenced system; and
- the grade or score has meaning in the sense that it can be directly linked to a description of the specific skills and knowledge that the student has demonstrated. This is not possible in norm referencing.

Despite the VET sector's claim to adhere to the principles of criterion referencing, much of the research and development activities pertinent to competence assessment have not reflected a criterion referencing approach largely as a result of the failure to recognise an underlying developmental continuum (Gillis, 2003). That is, possession of the competency tends to have been determined by the direct observation of performance, where each performance criterion is treated as an activity or task to be observed, with no notion of a developmental continuum. This is not a technically correct interpretation of criterion referencing. In fact, such an approach led to the demise of criterion referencing in the 1970s (Griffin, 1995).

The current approach to interpreting competence assessments may be a result of the definition of assessment in the Australian Vocational Education and Training sector, where competency based assessment (CBA) was defined as:

> *…the process of collecting evidence and making judgments on whether competency has been achieved to confirm that an individual can perform to the standard expected in the workplace as expressed in the relevant endorsed industry/enterprise competency standards or the learning outcomes of an accredited course.*
>
> (ANTA, 2001, p.5).

This definition was limited to an assessment of competence in which there were two levels of performance to be reported (i.e., competent and not yet competent). Gillis (2003) found that assessors adopted an evidence-driven approach to assessment and defined competence assessment as:

> *… a purposeful and rational process of systematically gathering, interpreting, recording and communicating to stakeholders, information on candidate development against industry competency standards.*
>
> (Gillis, 2003, p. 263)

Because it incorporated the notion of development, this definition enabled a range of assessment outcomes to be reported against industry competency standards, including those above and below the "threshold" of competence. It was also consistent with the original intent of the implementation of competency based assessment in the Australian VET sector, where CBA was defined as:

> *… the process of collecting evidence and making judgments on the extent and nature of progress toward the performance requirements as set out in a standard, or a learning outcome, and at the appropriate point making the judgment whether competency has been achieved.*
>
> (NTB, 1992, p. 57).

This original definition of competence assessment captured the essence of criterion referencing, but subsequent developments appeared to degrade to the practices and understandings that were more representative of the behaviourist objective movement of the 1970s. The dangers of repeating history if a technically correct criterion referenced approach to competence assessment was not properly implemented have been set out in a number of papers (e.g., Bowden & Masters, 1993; Griffin, 1995; Griffin et al., 2001; Wolf 1993).

With a broader definition of competence assessment, the process is not limited to a fixed number of achievement levels. This helps to eliminate the misconception that graded performance assessment is different to assessment practices and processes which report the dichotomy (competent or not yet competent). Graded performance assessment simply refers to an alternative reporting strategy. But if it is criterion referenced, rather than normative, it requires a clear understanding of an underlying developmental continuum, in which levels of performance can be defined and used for interpretation purposes.

### The grading criteria must be defined and meaningful (Rumsey #2)

Whilst the importance of establishing explicit criteria to grade performance has been widely recognized among RTOs (Griffin et al. 2001; Rumsey, 1997; Thomson, Mathers & Quirk, 1996), the meaningful nature of such criteria has itself become a contentious issue. Initial attempts to establish grading criteria in the VET sector were associated with the specification of criteria that were thought to be easily quantifiable, such as the "number of attempts in the assessment" or the "speed of performance" (Rumsey; Thomson et al.). Such criteria required minimal, if any, need to exercise professional judgment. However, the meaningful nature of such criteria, in terms of differentiating performance levels, was difficult to defend within a competence assessment framework. Fortunately, the importance of professional judgment in assessment has now been recognised in the VET sector, largely due to attempts to measure higher order competencies, and particularly those delivered at the higher level of the Australian Qualifications Framework (Connally, et al., 2003; Foreman, Davis & Bone, 2003; Johnstone & Evans, 2001).

Two predominant ways of defining grading criteria have now emerged: generic and specific. 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 context in which they are to be applied (McCurry, 2003). Alternatively, 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 and, therefore, are content and context specific. Each approach is considered next.

### *Generic Criteria*

Generic criteria refer to statements of achievement levels that have been designed to form the foundation for all assessment of all candidates regardless of context (Tognolini, 2001). They tend to be couched at a general level in an attempt to enhance applicability to the broad range of industry contexts and/or disciplinary fields. Examples of such criteria include "underpinning knowledge", "communication skills", "work organization" and "creativity" (Rumsey, 1997; Thomson et al., 1996).

Candidates who have undertaken different training programs within different industries are judged using the common criteria. Whilst this reduces the development costs, a set of common criteria does not ensure valid comparability. In many high stakes assessment programs, statistical moderation is used to control systematic extraneous influences, such as gender and discipline area, on assessment performances (Tognolini, 2001). This should also apply in high stakes competence assessments, such as assessments of VET in School subjects that are conducted as part of senior secondary certificates of education. Any attempts to compare candidate performances across locations within an industry sector also need to be moderated, preferably statistically, to control the influence of industry context and location (such as on-the-job versus off-the-job), irrespective of the type of criteria used.

A number of state systems have introduced assessment procedures that are based on the use of generic scoring criteria that reflect the Mayer Key Competencies (e.g., Western Australian and Queensland Departments of Training and the Victorian Curriculum and Assessment Authority). For example, the Western Australian approach to grading (Western Australian Department of Training and Employment, 1999) uses the following generic criteria to define the broad parameters on which performance will be based:

1. underpinning knowledge;

2. communication, networking, language and interpersonal skills;

3. techniques and processes;

4. work organization; and

5. level of independence and performance of work tasks.

These broad parameters, referred to as “scoring criteria” (WA Department of Training, 2001), were derived from the Mayer Key Competencies (Mayer, 1992), the Australian Qualification Framework descriptor (MCEETYA, 1995) and the four components of competency (i.e., perform task skills, task management skills, contingency management skills and job/role environment skills) (ANTA, 2004). Performance against each criterion was rated using a five point scale, with performance descriptors being provided for levels 1, 3 and 5. The criteria and descriptor statements were constant across industry and across competencies assessed. An example of the performance level descriptors for one of the five criteria in the Western Australian Department of Training’s (2001) Graded Performance Assessment System is presented below, where the assessor had to record the candidate’s performance against the criterion “*Techniques and Processes*” using the following five-point scale.

**Figure 1. Sample generic scoring criteria and performance level: Techniques and Processes**

| **Scoring Criterion***:* **Techniques and Processes** | |
|---|---|
| RATING | LEVEL OF PERFORMANCE |
| 5<br>(high) | ❖ Displays excellent technical skills/procedures to the standard exceeding organizational expectations |
| 4 | |
| 3<br>(medium) | ❖ Effectively performs all technical skills/procedures to the standard higher than required by the workplace, including correct use of any equipment. |
| 2 | |
| 1<br>(low) | ❖ Performs all technical skills/procedures to the standard required by the workplace, including correct use of any equipment. |

*( Source: Western Australian Department of Training, 2004).*

Figure 1 illustrates that the levels of performance for the criterion (*Techniques and Processes*) are determined by *how well* the candidate applies techniques and procedures to the workplace context. As the criterion is generic, there is no specific reference to guide the assessor as to the nature of the techniques and processes nor their context for application. There is also a reliance on comparative terms (e.g., *effectively, excellent…*) to differentiate each level, which, in turn, create uneasy self referenced (ipsative – i.e., intra-personal ) interpretations of what each level descriptor means in terms of the assessment of the general aptitude. This reduces the grading system to a point where assessors have to use an ipsative interpretation framework. As individuals have their own interpretation of such relative terms, the consistency of interpretation at an individual level, across assessors and across assessments is destroyed. Competence assessment is impossible with an ipsative frame of reference. The ambiguity is amplified at the intermediate [blank] levels, where assessors are advised that these performance levels should be judged by comparing to adjacent levels. This further undermines the consistency of interpretation of the levels across different assessors (i.e, inter-rater reliability).

McCurry (2003) demonstrated that classical reliability of approaches to judging performance against generic competencies (using criteria similar to those displayed in Figure 1) was dependent on aggregated data. But even this breaks down when the criteria are couched in comparative terms, which rely upon intra-personal interpretation frameworks. McCurry (2003) demonstrated that a reliable composite image of a student was dependent upon aggregating data across teachers on the generic criteria and competencies and then across a range of subject areas to get a group or institutional reliability index which he called “soft” reliability. He found that individual teacher decisions were fraught with “noise” and uncertainty when unclear criteria were used. As such, when comparative statements are used to assess a generic aptitude on a single rating scale, the clarity and consistency of what is assessed is compromised in the pursuit of simplicity. When this clarity is lost, there is an increased need for expensive and time-consuming moderation procedures.

The validity of deriving generic criteria from generic competencies is also questionable, particularly given the uncertainty that generic competencies exist independent of context (Hager & Gillis, 1995). There has been a long history of debate as to whether competencies are domain specific and therefore ought to be assessed within traditional discipline boundaries or whether competencies can be context free. For example, Tognolini (2001), Griffin et al., (2001) and McCurry (2003) raised concerns about generic competencies and the relevant measurement qualities regarding transferability across contexts. Grummon (1997) reported that

> *The question of the transferability of skills and knowledge – which is the heart of the generic versus specific discussion – is one that has not been completely answered for either assessment or instruction. Some skills, like interpersonal skills, do seem to transfer. Others transfer only in part. For example, students may be able to read for meaning more easily in an occupational area of interest to them and be less able to read for meaning in a general subject area.*
>
> (Grummon, 1997. p. 1).

Similarly Tognolini (2001) argued that

> *There has been a hundred years of psychological research showing that students have great difficulty generalizing their skills across subject boundaries and this has been a source of contention for educators around the world.*
>
> (Tognolini, 2001, p. 7).

The importance of context and specialized knowledge has been well documented in the *novice to expert* literature (Chi, Feltovitch & Glaser, 1981; McCurry, 2003; McGaw, 1993). As expertise (which can be categorized as high levels of competence in an occupational area) is domain specific, individuals often demonstrate little capacity to transfer expertise from one context to another unless these two contexts are closely related (Hager & Gillis, 1995; McCurry, 2003). Yet one of the fundamental requirements of competence is the "ability to transfer and apply skills and knowledge to new situations and environments" (ANTA, 1997). Given the body of knowledge in the *novice to expert* literature, this requirement may be spurious, particularly if competence is dependent upon both specific knowledge acquisition and expertise (Stanley, 1993). As such, the notion of transferability of competence may need to be reconceptualised to reflect realistic human behaviour where transferability is limited to new, yet *related* contexts and environments. This is a common tenet of transfer of learning, but it is persistently ignored in the debate on competence and grading.

McCurry (2003) persuasively argued that when the context for application is either ignored or deemed irrelevant in the assessment, then the assessment is limited to measuring an individual's general aptitude or ability to learn (e.g., to learn the skills and knowledge for a new job). Alternatively, assessment of industry competency standards is expected to measure a specific set of skills and knowledge that people have learned and acquired within a given context. Given the importance of context in both the definition and the assessment of competencies, McCurry (2003) argued that there could only be specific competencies, such as industry and enterprise competency standards. He further argued that generic competencies should be reconceptualised as general abilities or aptitudes to avoid the misleading notion that they are forms of competencies.

McCurry's (2003) work has largely been influenced by the distinctions made in the field of psychometrics between tests of attainment and aptitude, with the former referring to measures of a person's potential to learn (i.e., general abilities), and the latter to what people have learned (e.g., industry specific competencies) (Groth-Marnat, 1990). McCurry (2003) argued that specific competencies can be used to measure student attainment, whilst generic skills can, and should, only be used to measure general aptitude.

Consistent with this distinction, Stanley (1993) outlined the difficulties associated with using general abilities (or generic competencies) as a measure of educational outcomes. He argued that they were more "dependent on the relative contributions of individual differences which people bring to the task of learning than on the direct outputs of instruction" (p. 147). Furthermore, he argued that any differences in demonstrated general abilities of candidates may reflect more inherent individual differences in ability patterns than any real differences in educational experience. He challenged the validity of measures of general abilities, particularly when used to determine educational outcomes.

Hence, any assessment based on generic competencies is limited to the assessment of differences in general aptitude (McCurry, 2003), which may not reflect any differences in the vocational experiences of students (Stanley, 1993), thus questioning the validity of such measures. Consequently, assessments that use generic scoring criteria, which have been derived from generic competencies, to differentiate among performance would be difficult to defend. This is accentuated when comparative language is used to differentiate the performance levels (e.g., refer to Figure 1) thus impacting on the inter-rater reliability of such measures. This review leads us to conclude that the use of generic criteria cannot yield meaningful interpretations of competencies and hence cannot be applied consistently with the foregoing principle.

**Specific criteria**

Specific criteria establish the rules for judging the quality of evidence of learning or competency. They are content specific and are assessed within traditional discipline or industry boundaries and are context dependent (Pascoe, 2001). A standards referenced approach, a subset of criterion referencing, uses specific criteria to define levels of performance along a developmental continuum. The continuum is used for interpretive purposes to define and report a range of achievement levels. As Wolf (1993) argued, one of these levels defines the performance expected in the workplace and, therefore, reflects the cut–point for competence.

The notion of a developmental continuum of learning in workplace competence assessment was field tested in the Australian VET System recently when the Australian National Training Authority (ANTA) (through the NSW Board of Vocational Education and Training) commissioned the development of standards referenced interpretation models for assessing competencies. ANTA commissioned a national study in increasing the recognition by both industry and higher education of VET in School programs within senior secondary certificates of education (Griffin, Gillis & Calvitto, 2004). The Australian Research Council (ARC) commissioned a study into the public safety and public services industries (Connally, Jorgensen, Gillis & Griffin, 2003; Griffin, Gillis, Connally, Jorgensen & McArdle, 2003). Both studies required specific criteria to be established according to strict principles and guidelines. In both studies, a standards referenced interpretation framework required the development and use of scoring rubrics that were expressed in the form of ordered, ***transparent*** descriptions of quality performance that were ***specific*** to the unit(s) of competency; underpinned by a ***theory*** of learning; and were ***hierarchical*** and ***sequential.*** Both studies demonstrated that the interpretative model:

- allowed for ***multiple levels*** of performance quality to be identified along developmental continua;
- could be used for ***differentiation*** as well as ***recognition*** purposes;
- ***minimized implementation*** **costs** by using the same assessment evidence to report a range of assessment outcomes (e.g., competent/not yet competent decisions, performance levels of grades and marks) without any need to gather additional evidence or extend the assessment process (as is the case for some assessment models that use generic criteria);
- provided ***flexibility*** and ***autonomy*** for the assessors as it decentralized the assessment task development and merely standardized the interpretation of the evidence; and
- gave a ***substantive meaning*** to the grades, scores or marks.

The standards referenced approach to competence assessment, in which varying levels of quality performance were defined along a developmental continuum, was also consistent with the outcomes of the review of Training Packages for ANTA (Schofield & McDonald, 2004). In particular, it was consistent with the recommendations associated with the "expansion of the notion of competency to include a combination of higher level skills, where appropriate" (p. 17).

In their investigation of assessment of higher order competencies, Griffin et al. (2003) demonstrated that subject matter experts could develop the frameworks through a process of "unpacking" units of competency. This required an analysis of the elements, performance criteria, range of variables and evidence guides to develop a set of quality criteria that could differentiate levels of performance of individuals being assessed against the particular unit. They showed that the use of specific criteria allowed reliable decisions to be made about individual performances. Other studies developed progressive sets of performance levels and demonstrated that quality of performance mattered both in VET in School programs (Griffin et al., 2004) and in industry based assessments (e.g., Bateman, 2003; Connally, 2004; Connally et al., 2003; Nicholson, 2004). Each of these studies showed that the specialists were able to identify the levels of performance and define an acceptable level of performance required in the workplace without detriment to workplace competence assessment.

## Rubrics

A rubric is defined as "any rule, explanatory comment" (Geddes & Grosset, 1999, p.509) used in making a judgment of quality. In an assessment context, a rubric refers to the "scoring rules" and, in this case, statements that describe levels of quality in performances of workplace tasks. Rubrics define the rules for judging the performance. There are several elements to a useful rubric. For example:

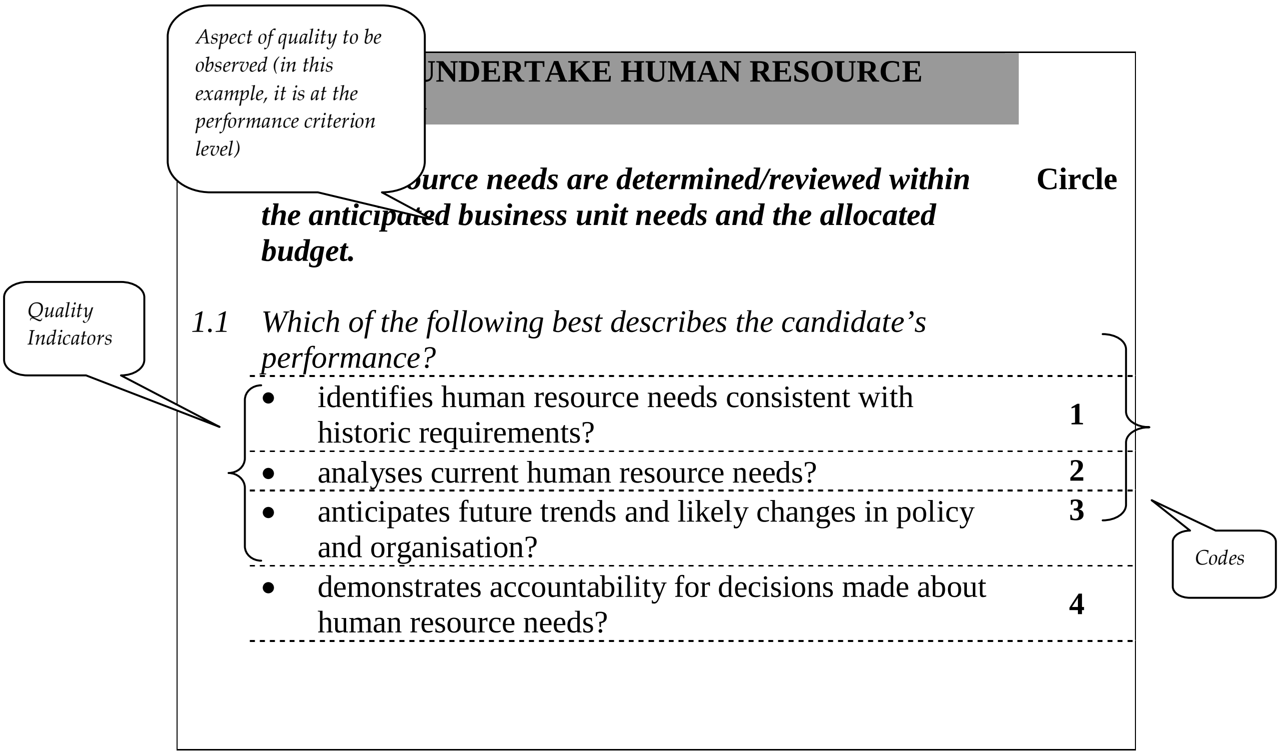
- the quality of performance is described in levels from low to high;

- each aspect of quality to be judged separately is important for the purpose of the assessment; and
- for each aspect of quality, rubrics provide a commentary describing the defining features of work at each level of performance.

Huba & Freed (2000)

The use of rubrics that defined quality of performances was central to both a criterion referenced and a standards referenced assessment interpretation approach. In the greater recognition of VET in Schools study (Griffin et al., 2004), rubrics were developed for performance criteria. An example is provided in Figure 2.

**Figure 2. An example of rubrics designed at the performance criterion/element level of competency.**



**UNDERTAKE HUMAN RESOURCE**

***urce needs are determined/reviewed within the anticipated business unit needs and the allocated budget.*** **Circle**

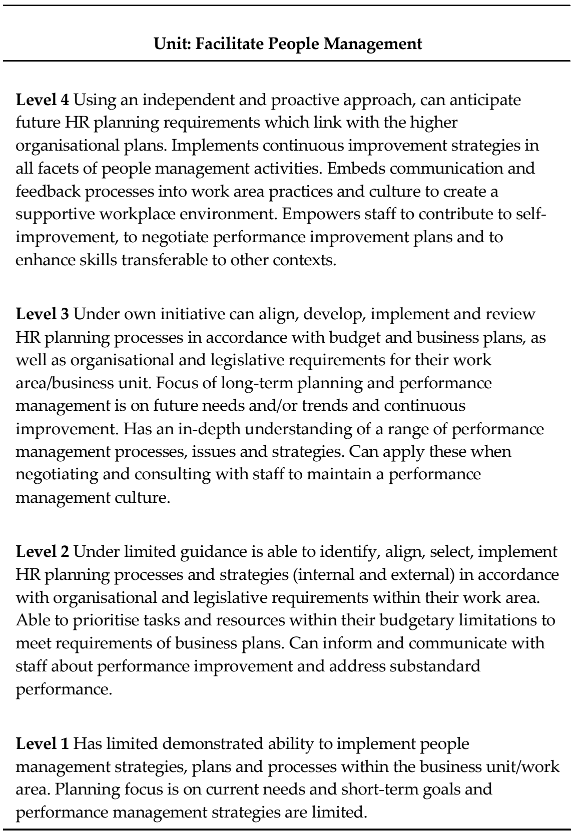
*1.1 Which of the following best describes the candidate's performance?*

- identifies human resource needs consistent with historic requirements? **1**
- analyses current human resource needs? **2**
- anticipates future trends and likely changes in policy and organisation? **3**
- demonstrates accountability for decisions made about human resource needs? **4**

Aspect of quality to be observed (in this example, it is at the performance criterion level)
Quality Indicators
Codes

It is also possible to define rubrics at a unit level and an example of such rubrics is provided in Figure 3.

**Figure 3: Using rubrics at the unit level of competency.**



**Unit: Facilitate People Management**

**Level 4** Using an independent and proactive approach, can anticipate future HR planning requirements which link with the higher organisational plans. Implements continuous improvement strategies in all facets of people management activities. Embeds communication and feedback processes into work area practices and culture to create a supportive workplace environment. Empowers staff to contribute to self-improvement, to negotiate performance improvement plans and to enhance skills transferable to other contexts.

**Level 3** Under own initiative can align, develop, implement and review HR planning processes in accordance with budget and business plans, as well as organisational and legislative requirements for their work area/business unit. Focus of long-term planning and performance management is on future needs and/or trends and continuous improvement. Has an in-depth understanding of a range of performance management processes, issues and strategies. Can apply these when negotiating and consulting with staff to maintain a performance management culture.

**Level 2** Under limited guidance is able to identify, align, select, implement HR planning processes and strategies (internal and external) in accordance with organisational and legislative requirements within their work area. Able to prioritise tasks and resources within their budgetary limitations to meet requirements of business plans. Can inform and communicate with staff about performance improvement and address substandard performance.

**Level 1** Has limited demonstrated ability to implement people management strategies, plans and processes within the business unit/work area. Planning focus is on current needs and short-term goals and performance management strategies are limited.

Regardless of the level of specificity, the aspect of quality to be observed (often referred to as the “criteria”) had to be directly related to the industry competency standards if a standards referenced interpretative framework was to have been used. The rules surrounding the development of rubrics remained the same, regardless of the level of specificity of the assessment criteria. Griffin (1997) developed a set of rules for defining rubrics. Rubrics must:

1. reflect levels of **quality of performance**. Each recognisable, different level of quality needs to be defined within each criterion to be observed. They should reflect the quality of cognitive, affective or psychomotor learning that is demonstrated in the candidates’ performances;

2. enable an **inference** to be made about developmental learning. They should not be just counts of things right and wrong or sequential steps in a process;

3. **discriminate** between levels of learning and performance quality;

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 language should be descriptive, enable inference and avoid the use of comparative terms;

6. be **transparent** in that they are written such that candidates can verify their own performance against the rubrics;

7. be **developmental** so that each successive level code implies a higher level of performance quality;

8. be **internally coherent** such that they should consistently describe performances in the same domain of learning;

9. reflect the level of performance quality (or difficulty) **relative** to all other rubrics and codes as stipulated in a quality matrix; and

10. lead to **reliable** and consistent judgments across judges. To this effect, no aspect of performance should have more than four or five levels. If more levels are required the task or sub-task should be split for coding purposes and two sets of rubrics developed.

The competency assessment used in both studies was a derivation of approaches used elsewhere. The studies drew on the lessons learned in the shift from norm-referenced scaling procedures to those in which developmental levels of performances were preferred (Masters, 1998; McGaw, 1997). The approach had been used in the Program for International Student Assessment (PISA), the Third International Mathematics and Science Study (TIMSS), the state-wide basic skills tests used in Australian schools from grades 3, 5, 7 and 9 (e.g., Victorian AIM tests, Northern Territory Multilevel Assessment Program (MAP), the NSW Basic Skills Test) as well as the NSW High School Certificate (McGaw, 1997). It became clear that the standards referenced approach was compatible with competence assessment and that it also yielded differentiating data about candidates that could be used in other contexts.

### Applied once competence has been achieved

Williams and Bateman (2003) proposed that grading should only be "applied once competence has been determined" in what appeared to be an attempt to minimize the use of marks and percentages. Percentages were apparently regarded as the common basis for determining normed grades. This resulted in a misconception that assessment has to be a two-tiered approach, requiring first the decision about competence and then the use of supplementary criteria to make other judgments for assigning grades. However, given that assessments can be interpreted in terms of progress along a developmental continuum (Glaser, 1981) the two-tiered approach becomes redundant (e.g., Griffin et al., 2001; Masters, 1998; McGaw, 1997; Stanley, 1993; Wolf, 1993). The continuum can be partitioned into levels, and one level can be used for competent/not yet competent decisions (Wolf, 1993). In keeping with this approach, an assessment only requires a single decision regarding the level on the developmental continuum that best describes the student's performance. As the developmental continuum is typically hierarchical, any demonstration of performance above the cut-point for competence would mean that competence has been achieved, and that the quality of performance was beyond the minimal level of performance required for competency. This approach eliminates the need to make more than one decision in the deliberation process.

Any model that uses supplementary generic criteria (e.g., the Western Australian graded performance model) requires two distinct decisions to be made and recorded. The first is related to declaring competence and the second requires a judgment in a different aptitude domain. According to McCurry (2003), this should be avoided because of issues associated with lack of validity and reliability. Whilst the additional information may be useful for purposes other than competence assessment, assessing generic abilities adds to the workload and provides information that can be tangential to the primary purpose of the assessment.

Despite Williams and Bateman's (2003) concern with the use of marks, and especially percentages, grading can be based on scores in a criterion referenced framework, but the scores and grades must have meaning. Meaning is defined as the capacity to describe a performance level on a developmental continuum. For example, in a review of the NSW High School Certificate, McGaw (1997) proposed that

> *… with careful analysis of the characteristics of the performances that yield different marks on the scales for each question it would be possible to move towards descriptors that permit a substantive interpretation of the [Geography] achievement scale.*
>
> (p.18).

He further argued that the

> *… solution to that problem is not to take the marks examiners assign to answers to the different questions and add them up but to use a statistical model that takes account of the differences in difficulty of questions in estimating the achievement level of students .*
>
> (McGaw, 1997 p.18).

As part of the national VET in Schools differentiating scored assessment project, Griffin and colleagues (2001) adopted McGaw's recommendation that a standards referenced framework should be used. They developed an efficient and inexpensive process of scoring competence assessments. Subject matter experts, appointed by Industry Training Advisory Boards (ITABS), developed scoring rubrics for performance criteria and weighted them according to their capacity to discriminate performance amongst students. This eliminated the need to conduct sophisticated statistical analysis requiring specialist skills in psychometrics. The model was field tested in VET in Schools programs as part of a national ANTA funded project on greater recognition (Griffin et al., 2004), and in assessment of higher order competencies in industry (Connally et al., 2003). The model was also validated in a number of postgraduate studies (e.g., Bateman, 2003; Connally, 2004; Nicholson, 2004) focusing on training package units addressing human resource management and manual handling competencies.

Scores are not necessarily bad things in competence assessment. If the scores have meaning and can be translated into performance descriptions, then the utility of the competence assessment is increased as it can incorporate quality criteria and differentiating scores that can be used for other subsidiary purposes. Supplementary criteria in other domains of learning may be useful for a range of additional reasons but should not be confused with the central purpose of the assessment: to identify and measure competence. Wolf's (1993) advice is important. Competence is just one level on a continuum of proficiency and the capacity to extend beyond two levels adds value to the assessment and to the use to which the assessment can be put.

## Conclusion

This article explored a number of principles that have been proposed in the VET sector for graded assessment. The principles have been examined from both a theoretical and practical assessment perspective. Grading can be done in a criterion referenced framework. It was argued that current competence assessment practices in VET have not reflected a criterion referenced interpretive framework. This was largely due to a failure to recognize an underlying developmental continuum. The continuum is needed so that a candidate's progress can be mapped when developing grading criteria and associated performance level descriptors (Griffin et al., 2003; Masters, 1993; McGaw, 1997). A clear and theoretically sound understanding of the continuum is needed to inform the development of meaningful and valid criteria and descriptors of quality performance. Generic criteria could not be defended in assessments of achievement levels as they limit the assessment to measurement of general aptitudes (McCurry, 2003), which may be related to neither the competencies of interest nor the training experience (Stanley, 1993). Hence this raises concern about the validity of generic criteria for competence assessment, whether differentiated or not. The consistency of interpretation of levels using vague and comparative descriptive statements to measure such generic abilities (McCurry, 2003) destroys the classical reliability and removes validity when they are applied to competence assessment. A standards referenced interpretive approach, on the other hand, satisfies the requirements of criterion referencing (and hence competence assessment) and enhances the content, construct and criterion validity as well as the inter-rater reliability of the assessment. It also minimizes the implementation costs associated with the assessment because it enables the same evidence to be used to report a range of assessment outcomes (e.g., competent/not yet competent and the performance level achieved in terms of a grade), using a single judgment of the candidate's performance level on the developmental criterion referenced continuum.

The following principles underpinned the approach to competence assessment trialed in both the VET in School Programs (Griffin et al., 2004) and in industry (Connally et al., 2003). As they reflect broad concepts of assessment and reporting that have acceptance across a range of educational contexts, the problems associated with using flawed rules or technically incorrect principles have been avoided. The principles listed below were derived to accommodate idiosyncratic differences in practices across education systems and were applicable when reporting both dichotomous (competent/not yet competent) and polychotomous decisions (levels on a continuum), thus increasing flexibility and applicability. The principles are that

11. the system of assessment and reporting must be situated in a **theory** of learning and assessment;

12. the procedures and assessment must satisfy both a normed and **criterion referenced** interpretation;

13. the model, approach used, assessment method, materials and decisions must be **transparent** and externally **verifiable** through a formal audit process;

14. the assessment procedure and model must be **resource – sensitive** in both development and application;

15. 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;

16. 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;

17. the procedure must have both face and construct **validity**;

18. the procedure must be demonstrably fair, equitable and **unbiased**;

19. the model must be **communicative** and satisfy the information needs of stakeholders in a quality assurance context that must be accommodated; and

20. the scores and assessments are amenable to statistical and/or consensus moderation to ensure **consistency** of decisions and accuracy of score.

## References

ANTA 1997,  Guide to development of training packages. Australian Training Products, Melbourne.

ANTA 2001, Australian Quality Training Framework: Standards for Registered Training Organisations. Australian Training Products, Melbourne.

ANTA 2004, Training package development handbook. Part 2 endorsed components: Chapter 2 Developing units of competency. Australian National Training Authority, Melbourne.

Bateman, A 2003, The appropriateness of professional judgment to determine rubrics in competency based assessments, Unpublished Master of Assessment and Evaluation thesis, The University of Melbourne.

Bowden, J, & Masters, G 1993, Implications for higher education of a competency based approach to education and training. Australian Government Publishing Service, Canberra.

Connally, J 2004, A multi source measurement approach to the assessment of higher order competencies, Unpublished Doctoral thesis, The University of Melbourne.

Connally, J, Jorgensen, K, Gillis, S, & Griffin, P 2003, 'An integrated approach to the assessment of higher order competencies.' Refereed paper presented at the Sixth Australian VET Research Association Conference, *The changing face of VET*, Sydney, 9-11th April.

Chi, MTH, Feltovitch, P & Glaser, R 1981, 'Categorisation and representation of physics problems by experts and novices', Cognitive Sciences, vol *5*, pp121-151.

CSB-Assessors and Workplace Trainers 1993, Competency standards for assessors, PSI Consultants, Melbourne.

Foreman, D, Davis, P, & Bone, J 2003, Assessment practices at diploma and advanced diploma levels within training packages, National Centre for Vocational Education Research, Adelaide.

Foyster, J 1990, Getting to grips with competency-based training and assessment, TAFE National Centre for Research and Development, Leabrook.

Geddes & Grosset 1999, New English dictionary and thesaurus, Children's Leisure Products Limited, Scotland.

Gillis, S 2003, Domains of vocational assessment decision making. Unpublished Doctoral thesis, The University of Melbourne.

Glaser, R 1981, 'The future of testing: A research agenda for cognitive psychology and psychometrics', American Psychologist, vol 36(9), pp923-936.

Griffin, P 1995, 'Competency assessment: Avoiding the pitfalls of the past', Australian and New Zealand Journal of Vocational Education, vol 3(2), pp33 - 59.

Griffin, P 1997, Developing assessments in schools and workplaces, The University of Melbourne, Faculty of Education, Deans Series on Education.

Griffin, P, Gillis, S, & Calvitto, L 2004, Connecting competence and quality: Scored assessment in Year 12 VET, Submitted to the NSW Board of Vocational Education and Training.

Griffin, P, Gillis, S, Connally, J, Jorgensen, K, & McArdle, D 2003, A multi-source approach to assessing higher order competencies, A project report to the Australian Research Council, Canberra.

Griffin, P, Gillis, S, Keating, J, & Fennessy, D 2001, Assessment reporting of VET courses within senior secondary certificates in Expanding Opportunities for Youth: Greater Industry Recognition of Achievement in VET in School Courses, NSW Board of Vocational Education and Training, Sydney.

Griffin, P & Nix, P, 1991, *Educational assessment and reporting: A new approach.* Harcourt Brace Jovanovich, Sydney.

Groth-Marnat, G 1990, Handbook of psychological assessment (2nd Edition), Wiley-Interscience Publication, New York.

Grummon, PTH 1997, Assessing students for workplace readiness, Centrefocus Number 15, http://ncreve.berkeley.edu/CentreFocus/CF15.html.

Hager, P, Athanasou, J, & Gonczi, A 1994, Assessment technical manual, Australian Government Publishing Services, Canberra.

Hager, P & Gillis, S 1995, Assessment at higher levels of competence in W.C. Hall (Ed) Key aspects of competency based assessments, National Centre for Vocational Education Research Ltd, Adelaide., pp.59-72.

Hall, W, & Saunders, J 1993, Getting to grips with assessment. National Centre for Vocational Education Research, Adelaide.

Huba, ME, & Freed, JE 2000, Learner-centered assessment on college campuses: Shifting the focus from teaching to learning, Allyn & Bacon, Boston, MA.

Johnstone, I & Evans, G 2001, ‘Assessing competencies in higher qualifications’ in Training package assessment materials, Department of Education and Training & Youth Affairs, Canberra.

Masters, G 1993, Certainty and probability in assessment of competence, Paper presented at the VEETAC National Assessment Research Forum on Competency-based Assessment Issues, Sydney.

Masters, G 1998, ‘Standards and assessment for students and teachers: A developmental paradigm’, Incorporated Associated of Registered Teachers of Victoria Seminar Series, May 1998, No. 74.

Mayer Committee 1992, Putting general education to work: The key competencies report, AEC/MOVEET, Melbourne.

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 a work related skills and the whole school assessment of generic skills, Australian Council for Educational Research, Melbourne.

MCEETYA 1995, Australian Qualifications Framework – Implementation handbook, Ministerial Council on Education, Employment, Training and Youth Affairs, Canberra.

McGaw, B 1993, Competency based assessment: Measurement issues, Paper presented at the National Assessment Research Forum, NSW TAFE Commission, Sydney.

McGaw, B 1997, Shaping their future: Recommendations for reform of the Higher School Certificate, NSW Board of Vocational Education and Training, Sydney.

Murphy, KR, & Davidshofer, CO 1998, Psychological testing: Principles and applications (4th ed.), Prentice Hall, London.

Nicholson, K 2004, Trial of a standard referenced framework for the defining and measuring of the manutention competency, Unpublished Master of Assessment and Evaluation thesis, The University of Melbourne.

NAWTB 1998, Training package for assessment and workplace training: BSZ98, Australian Training Products, Melbourne.

NTB 1992, Policy and guidelines (2nd ed.), National Training Board, Canberra.

Pascoe, S 2001, Generic versus content-driven assessment, Paper presented at the Australian Curriculum, Assessment and Certification Authorities (ACACA) Annual Conference 2001, Sydney.

Peddie, R 1997, ‘Difficulty, excellence and levels: Implications for a qualifications framework’ Australian and New Zealand Journal of Vocational Education Research, vol 5(2), pp56-76.

Rumsey, D 1997, Reporting of assessment outcomes within competency based training and assessment programs under New Apprenticeship, Australian National Training Authority, Melbourne.

Rumsey, D 2003, Think piece on the training package model, Australian National Training Authority, Brisbane.

Schofield, K & McDonald, R 2004, Moving on: report of high level review of training packages. Australian National Training Authority, Melbourne.

Smith, E 1996, ‘Study of the penetration of competency based training in Australia’, Journal of Research in Post-Compulsory Education, vol 1(2), pp169-185.

Smith, L 2000, Issues impacting on the quality of assessment in vocational education and training in Queensland, Department of Employment, Training and Industrial Relations, Brisbane.

Stanley, G 1993, ‘The psychology of competency-based education’, in C. Collins (Ed.), The debate on competencies in Australian education and training, The Australian College of Education, Canberra, pp.145-153.

Thomson, P, Mathers, R, & Quirk, R 1996, Grade debate: Should we grade competency based assessment? National Centre for Vocational Education Research, Adelaide.

Tognolini, J 2001, Generic versus content-driven assessment. A paper presented at the Australasian Curriculum Assessment and Certification Authorities (ACACA) Conference, Sydney, July 26.

Western Australian Department of Training and Employment 1999, Performance with merit in competency based training, WADTE, Perth.

Western Australian Department of Training, 2001, Graded Performance Assessment: Professional Development Materials, WADT, Perth.

William, D 2000, August, Integrating formative and summative functions of assessment, Paper presented to Working Group 10 of the International Congress on Mathematics Education, Makuhari, Tokyo.

Williams, M, & Bateman, A 2003, Graded assessment in vocational education and training: An analysis of national practice, drivers and areas for policy development, National Centre for Vocational Education Research, Adelaide.

Wolf, A 1993, Assessment issues and problems in a criterion based system, Further Education Unit, Stamford, UK.