Characteristics, uses and rationales of mark-based and grade-based assessment

Joanna Williamson Research Division

Introduction

Qualifications currently available in England include examples of both mark-based and grade-based approaches to assessment. Mark-based approaches require assessors to assign numerical marks to candidates' work, assisted by a mark scheme. Grade-based approaches avoid marks altogether, and require assessors to assign grades by evaluating candidates' work against grading criteria. These are statements that set out the knowledge or skills that must be demonstrated to gain a particular grade (e.g., "Deliver sports/activity sessions using effective communication" in a Sports qualification).

The choice between a mark-based or grade-based approach has important implications for a given qualification's overall assessment model. The choice determines what kind of judgement assessors are required to make, and the standards they must understand and apply. Mark-based and grade-based approaches also facilitate different ways of combining or aggregating judgements. For example, in a mark-based approach, the marks given for individual tasks can be added together. Despite these significant consequences, little has been written about the rationale for deciding upon a mark-based or grade-based approach, or the consequences of the decision.

This article outlines the characteristics, uses and rationales of mark-based and grade-based approaches to assessment. It focuses on the potential strengths and weaknesses of mark-based and grade-based approaches for internal assessment[1] in vocationally-related qualifications (VRQs). Qualifications of this type that are available in England, such as Applied Generals and Technical Awards, include examples of both mark-based and grade-based approaches, and, for this reason, are a particularly interesting context to consider.

Where are mark-based and grade-based approaches used?

Before reviewing the characteristics of mark-based and grade-based assessment approaches in depth, it is worth noting the contexts in which they are used. Table 1 shows where mark-based and grade-based approaches are used within common qualifications taken in England. General qualifications (GQs) in England such as the General Certificate of Education Advanced Level (GCE A Level) are overwhelmingly assessed using mark-based approaches. Applied and technical qualifications, by contrast, include examples of both mark-based and grade-based assessment.

Table 1: Common qualification types in England and their assessment approaches

Qualification type	*Qualification*	*Unit/component*[2]	*Approach*
Applied/ Technical	BTEC Nationals Level 3	Internally assessed units	Grade-based
		Externally assessed units	Mark-based
Applied/ Technical	Cambridge Nationals	Internally assessed units	Mark-based
		Externally assessed units	Mark-based
Applied/ Technical	Cambridge Technicals	Internally assessed units	Grade-based
		Externally assessed units	Mark-based
Applied/ Technical	General National Vocational Qualification (GNVQ)	All	Grade-based
Applied/ Technical	NCFE V Certs Level 2	Internally assessed units	Grade-based
		Externally assessed units	Grade-based
General	GCE A Level	All	Mark-based
General	GCSE	All	Mark-based
General	Cambridge Pre-U	All	Mark-based

Framework for considering strengths and weaknesses of mark-based and grade-based approaches

Neither mark-based nor grade-based assessment approaches are inherently 'better'. To judge their strengths and weaknesses, they need to be evaluated in terms of how well they support a particular assessment purpose. Broadly speaking, the assessment purpose of most qualifications is reliable and valid assessment of specified knowledge and skills, that is acceptable to those involved: teachers, students, regulators, awarding bodies and employers. More specific assessment purposes, however, depend on the particular aims and context of the qualification (e.g., to certify competence in a particular occupation).

Empirical evidence on the reliability of internal assessment, particularly in VRQs, is rare (Johnson, 2013). However, a broader body of research offers evidence about factors that affect reliability more

1. In *external assessment*, an awarding body sets the assessment task(s), specifies when and how candidates take the assessment, and carries out the marking/grading. In *internal assessment*, one or more of these activities is carried out by the candidate's school/college or workplace instead of an awarding body.

2. Many qualifications, including those listed in Table 1, are made up of multiple units or components. Candidates are assessed in these units/components separately, and the overall qualification result is then determined by combining the unit-level results. Different units/components may assess different areas of knowledge and skill. For example, a Sports qualification might contain three units/components assessing Sport Performance, Physiology and Coaching.

generally. Marker agreement can be increased either by improving the expertise of assessors, or by making the marking/grading task less demanding (Suto & Nádas, 2008). Since neither mark-based nor grade-based approaches can directly improve assessor expertise, the way in which they can best support reliable internal assessment is by reducing the demand of the marking/grading task.

To understand marking/grading task demand in internal VRQ assessment, it is important to acknowledge two important aspects of the marking/grading context. The first is that internal VRQ assessments frequently assess portfolios of evidence from tasks such as projects and practical activities, characterised by relatively low levels of task constraint (Ahmed & Pollitt, 2011). Assessing portfolios is arguably fairly demanding, due to "the complex nature of the assessment task (multiple, often non-standard and probably complex assignments)" (Johnston, 2004, p.395) and the fact that assessors must deal with a large amount of information (Johnson, 2008a, p.28), which may not necessarily be well-organised. The second aspect to acknowledge is that, internal assessors typically form a much larger group than examiners, but may only have experience of assessing their own students, and may have varying levels of skill.

Characteristics of mark-based and grade-based approaches

Considered in abstract, the only inherent difference between mark-based and grade-based approaches is the difference outlined at the start of this article. In a mark-based approach, assessors assign numerical marks to candidate work, assisted by a mark scheme, whereas in a grade-based approach, assessors evaluate candidates' work against grading criteria to decide upon a grade, avoiding marks altogether. Mark-based and grade-based assessments found in practice, however, differ across a range of characteristics. Some differences result directly from the choice of a mark-based or grade-based approach, whilst others instead reflect the rationales and characteristics of assessment traditions that are strongly associated with mark-based or grade-based approaches. The differences can be classified into those to do with *making judgements*, and those to do with *aggregating judgements*. The following sections explore these differences, and the evidence for how they may form advantages or disadvantages.

Making judgements

Rating scale

A mark scheme designed for tasks with low levels of constraint, such as those found in internal VRQ assessments, typically assigns marks by describing different levels of candidate performance. Figure 1 shows an extract from such a levels-based mark scheme. Each row describes a different level of performance, and the mark bands show what ranges of marks should be awarded to work at these levels. Using this mark scheme, assessors compare candidates' work to 'the standard' by deciding which level best describes the candidate's work, and to what extent. Levels of candidate performance are discriminated through comparison to marking criteria that describe the quality of candidate work expected at different points on the mark scale.

Level	*Mark range*	*Marking criteria*
Band 1	1–4 marks	Selects formulae and functions to produce a solution which has **limited** capacity to meet user requirements.
Band 2	5–7 marks	Selects formulae and functions to produce a solution that includes elements of efficiency and satisfies **some** of the user requirements.
Band 3	8–10 marks	Selects formulae and functions to produce a solution that is **effective** and efficient and in the main **accurately** meets user requirements.

Figure 1: Extract from a levels-based Cambridge Nationals mark scheme (OCR, 2017)

In grade-based approaches to internal VRQ assessment, candidate work is typically compared to the standard through the assessor deciding whether individual grading criteria have been met or not met. Individual grading criteria specify separately identifiable skills or competencies, and candidates are assessed against each. This is not a necessary feature of grade-based approaches, but is the dominant practice found in grade-based VRQ assessments, including BTECs and Cambridge Technicals.

The key advantage of a mark-based approach is that it can capture more information about candidate performance, recording the extent to which a candidate achieved something, not just that it was (or was not) achieved. A commonly found disadvantage of mark-based approaches is that it can be difficult to differentiate between multiple points on a rating scale, even for experienced assessors (Johnson, 2011). This increases task demand, and may also lead to underuse of certain marks, with a negative effect on reliability.

Wolf (1993) discusses rating scale choice in the specific context of criterion-referenced assessment, emphasising that the choice depends on "the inherent logic of the subject (whether there are key, recognisable, thresholds) and also on the context (the use to which it is being put)" (p.23). In particular, Wolf stresses:

> *There is nothing about criterion-referenced testing which ties it to a pass-fail, on-off approach. Criterion-referenced assessments produce a 'distribution' of performance (or, in a formal test, marks) in exactly the same way as any other assessment does. A single pass-fail is ONE way to partition that distribution but only one. (p.23)*

Wolf argues that a binary (met/not met) judgement may not be the most reliable way to assess performance against a standard, on the basis that qualifications assess human behaviour, and "… human behaviour does not usually fall into 'on/off' categories but instead along a continuum" (Wolf, 1993, p.25). For this reason, "The decision about whether someone should be placed on one side or the other of a particular criterion line can consequently be difficult to make – and the assessor's judgement fallible" (p.25). The consequences for reliability are that aggregation of binary decisions may lead to a result that does not reflect candidate achievement: the result "may be quite untrue to the underlying performance of the candidate", whereas the aggregation of more finely-graded judgements can be fairer because of the additional information that is captured at the first assessment stage (p.25).

Mark scheme structure

Levels-based mark schemes vary in the extent to which they are *holistic* or *analytic* in structure. In holistic levels-based mark schemes (see, for

example, Figure 1 in the Appendix), each level of response described may encompass multiple component skills or features, without guidance on how to weigh their relative importance. Assessors are required to decide which band provides the best fit to the candidate's response, although a candidate response may exhibit varying levels of performance against the component skills described. The disadvantage of such 'best fit' levels-based mark schemes is that marker agreement may be undermined by leaving aspects of the assessment judgement without explicit guidance. Centres and assessors may vary in how they weigh component skills, which introduces a threat to validity.

An analytic levels-based mark scheme (see, for example, Figure A2 in the Appendix), describes levels of candidate performance separately for multiple strands or component skills. Ahmed and Pollitt (2011) argue that explicitly analytic mark schemes are the most reliable way to mark unconstrained tasks, because they minimise the chance that an assessor has to make an assessment judgement without guidance from the mark scheme. Explicitly analytic mark schemes give "more help … by making it clear what distinguishes better from poorer responses" (Tisi, Whitehouse, Maughan, & Burdett, 2013, p.24). This is particularly valuable for tasks that are unconstrained, with hence less predictable candidate responses. A disadvantage of analytic mark schemes, however, is that, the more closely component skills are analytically decomposed and specified, the more information assessors are required to assimilate, and the more the mark scheme resembles a lengthy points-based mark scheme instead of a levels-based mark scheme (Pinot de Moira, 2013). Points-based mark schemes allocate marks to objectively identifiable words or ideas that are listed in the mark scheme, and are considered inappropriate for tasks with low levels of constraint (such as those found in internal VRQ assessments), not least because they require the pre-specification of all possible credit-worthy points.

Assessment criteria

The clarity of assessment criteria is a direct way for a marking or grading approach to affect the demand of a marking/grading task and thereby marker agreement. Ahmed and Pollitt argue that assessors are most effectively guided when a mark scheme "interprets the important aspects of the trait in the specific ways in which they should appear in responses" (Ahmed & Pollitt, 2011, p.275). Even so, the interpretation of descriptive assessment criteria can present difficulties. These arise from the need to interpret the vocabulary used, which will be influenced by subject-specific construct perceptions, assessors' professional experiences, and also their personal experiences and values (Johnson, 2013; Nádas, Suto, & Grayson, 2012). Assessors may encounter difficulties in separating quality from quantity when making judgements, and may perceive some criteria to overlap. The findings of Johnson (2008a, 2008b) illustrated these difficulties in the context of mark-based VRQs, and showed that marker agreement was negatively affected where these difficulties occurred.

Grade-based approaches found in internal VRQ assessments, such as those of Cambridge Technicals and BTEC qualifications, are underpinned by the logic of an 'objective' judgement of performance against individual grading criteria. Such grade-based approaches typically feature grading criteria that are far more concise than descriptive marking criteria. Taken in isolation, this seems to be an advantage since it imposes lower cognitive demands than multiple pages of high-density text (Pinot de Moira, 2013). On the other hand, the much shorter criteria necessarily provide assessors with less information. Furthermore, whilst descriptive criteria in a mark-based approach are notoriously vulnerable to variable interpretation, the concise criteria in grade-based approaches may still suffer from this problem.

Grade-based assessment has in the past been assumed to be more reliable than other forms, due to the 'transparency' of criteria based on competences. In practice, research and experience have conclusively demonstrated that "even tightly written specifications of criteria are capable of multiple interpretations" (Baird, Beguin, Black, Pollitt, & Stanley, 2012, p.55; Eraut, Steadman, Trill, & Parkes, 1996; Greatorex & Shannon, 2003; Wyatt-Smith & Klenowski, 2013). The extremely detailed criteria written for GNVQ assessment, for example, could not prevent "very low agreement on key indicators" and a "'lack of consensus' relating to grading standards" (Wolf, 1998, p.438). Carter and Bathmaker (2017), more recently, illustrated how supposedly unambiguous assessment standards in grade-based VRQ assessment can be interpreted by assessors as uncomfortably flexible.

Wolf (1993) argues that it is misleading to conceive of competency and criterion-based assessment approaches as qualitatively different from other assessment approaches, since no criterion can ever be entirely transparent, and human performances are intrinsically variable. It follows that candidate performance "cannot be fitted mechanistically to either a written list of criteria or an exemplar" and will require at least some assessor judgement (Wolf, 1993, pp.16–17). Wolf emphasises that "while assessment systems may vary in the *degree* to which these complex judgements come into play, such judgements are universal to all assessments" (Wolf, 1993, p.17).

In both mark-based and grade-based approaches, efforts can be made to increase the clarity of assessment criteria. However, there are risks attached to doing so. One is that criteria containing more clarifying information for assessors (with the aim of lowering marking task demand) almost invariably add to the total amount of information that must be read and understood. Pinot de Moira (2013) concluded that criteria giving assessors *less* information could result in more reliable marking due to the lower cognitive demand when mark schemes were uncluttered and required less reading. A second risk is that by pursuing clarity of criteria in order to increase the reliability of assessment, criteria can be altered in a way that threatens validity. Eraut et al. (1996, p.5) summarised the tensions between clarity of standards and validity in vocational assessment as follows: "The search for perfect reliability leads toward tests whose completion and marking allow for no possible margin of error and end up measuring nothing worth knowing. Pursuing perfect reliability leads to meaningless assessment."

Aggregating judgements

Mark-based and grade-based approaches may differ in both the aggregation of assessment judgements within units/components, and in the aggregation of unit-level results to form the overall qualification result. Since all aggregation involves loss of information, it can be argued that the optimal aggregation of assessment judgements would be no aggregation at all. Reporting a lengthy profile of marks or grades, however, tends not to be acceptable to users of results (such as employers and universities).

Within mark-based assessment, two main aggregation methods are possible. Either marks can be added together, or marks can be mapped onto a different numerical scale and then added. An example of the

latter approach is the *Uniform Mark Scale* (UMS) used until recently in modular GCE A Levels (Gray & Shaw, 2009). Within grade-based assessment, many methods of aggregation are possible. One approach is to convert grades into points, and then calculate a total score or average, for example, *Grade Point Average* (GPA). Another approach is to apply an algorithm or rule based on the profile of unit-level grades. For example, the rule could specify that the qualification grade is determined by the lowest grade achieved on any single unit.

The methods used to aggregate judgements can lead to very different outcomes from the same initial set of results. Aggregation affects assessment reliability, and also affects the validity and acceptability of a qualification result, due to decisions about what the aggregation preserves and what is lost. Thomson (1992), discussing possible aggregation methods for modular qualifications like the General Certificate of Education (GCSE), emphasises that different aggregation methods must be evaluated in light of the purpose of the qualification: "… the philosophy of the awarders is an essential element in trying to decide which system is best, or most appropriate. Whether the system is quantitative or qualitative, the decision to choose between them will always be qualitative" (Thomson, 1992, p.7).

The characteristics of aggregation methods that differ include the degree of compensation allowed between units, the reliability of overall results, the likelihood of anomalous results, and the loss of information that occurs between assessment and overall grades. The following sections compare mark-based and grade-based aggregation methods in these terms, and consider their advantages and disadvantages for supporting common assessment purposes.

Compensation

Mark-based approaches to assessment almost invariably involve the addition of marks achieved for different questions or tasks, and hence an element of compensation. If a unit-level result is obtained by adding the marks achieved for each learning outcome, for example, low performance against one learning outcome can be offset by high performance against another. By contrast, many grade-based approaches to assessment, including those used in Cambridge Technicals and BTECs, use *conjunctive* or *hurdle rules* to aggregate within units, whereby a candidate must achieve a given grade across *all* assessment criteria before they are awarded that grade for the unit as a whole.

Compensation is "a guiding principle of most examinations" (Cresswell, 1988, p.370) and in the context of GQs is widely agreed to be fairer and more reliable than non-compensatory approaches (Cadwallader, 2014, p.14; Cresswell, 1987). Arguments for fairness include offering students flexibility, and avoiding results being determined by "the simplest task in which you fail" (Cresswell, 1987, p.251; citing Forrest & Shoesmith, 1985). The latter is considered not just unfair in a general sense, but specifically at odds with the goal of rewarding achievement. Where assessment criteria are differentiated into levels of performance and aggregated using a conjunctive or hurdle approach, there is "failure to reward appropriately the achievement of candidates whose demonstrated attainments do not fit the hierarchical pattern anticipated" (Cresswell, 1994, p.50).

Compensatory aggregation may also be considered an advantage due to matching the practice of GQs. Where students (and their parents and teachers) are accustomed to compensatory aggregation, and qualification results are likely to be compared with results from assessments with compensatory aggregation, assessment hurdles may be considered particularly unfair. In 2018, the Office of Qualifications and Examinations Regulation (Ofqual) advised awarding bodies that the 'must pass' examination hurdles present in Applied General and Tech Level qualifications should be reconsidered, since A Level qualifications did not include similar hurdles, and Applied General and Tech Level candidates were therefore being unfairly disadvantaged (Beach, 2018, March 15).
It can be argued that it is problematic to evaluate VRQ assessment practice in terms of GQ practice, since VRQs have different aims and purposes. However, the differences can be overstated. Ofqual's argument recognises that there exists substantial overlap in the uses of VRQ and GQ results. Not least, many Applied General and Tech Level candidates will use their results to compete with GQ candidates for entrance to Higher Education and employment.

The logic of hurdle-based aggregation reflects the logic of criterion-referencing and competence-based assessment[3]. Compensation between skills is intentionally avoided, so that a given grade 'guarantees' that a particular set of criteria has been met (Greatorex, 2001, p.7), a position that depends on assuming that individual criteria are assessed without error. The guarantee is the principal advantage of hurdle-based aggregation: It becomes "possible to ensure that all candidates have successfully completed tasks deemed appropriate to the award of their grades" (Good & Cresswell, 1988, p.49). The lack of this guarantee is, correspondingly, the chief disadvantage of compensatory approaches. The compensation "obscures the *meaning* of a grade" (Cadwallader, 2014, p.4), since a student with a given grade may have performed uniformly well across assessment criteria, but may also have excelled in some (unknown) areas whilst having weaknesses in other (unknown) areas.

The disadvantages of hurdle-based aggregation correspond, as expected, to the advantages of compensatory approaches. When used within units, hurdle-based aggregation means that unit-level results are determined by the lowest-graded criterion that a candidate has failed to achieve. This means that students' grades "no longer reflect their average achievement" (Good & Cresswell, 1988, p.49), and is arguably incongruent with the aim of rewarding achievement. The introduction of hurdles can also lower achievement at qualification level, sometimes dramatically (Baume & Yorke, 2002; Taylor, Pritchard, & Gray, 2006). Acquah and Malpass (2015), discussing assessment design decisions for a Technical Baccalaureate, suggest two further disadvantages of hurdle-based aggregation. Firstly, hurdles may demotivate learners who feel they are unlikely to pass one element. Secondly, a washback effect could direct centres and students to focus disproportionately on elements that may not be passed, neglecting other areas of learning.

Information loss

Mark-based aggregation methods preserve a high level of information about students' performances at unit level. Retaining the fine-grained information about unit-level performance "until the last moment" means that overall grades "can be based on maximum evidence" (Thomson, 1992, p.42).

3. Wolf (1995) discusses how competence-based assessment can be seen as a specialised, vocationally-focused form of criterion-referenced assessment, and offers the following definition:

> *Competence-based assessment is a form of assessment that is derived from the specification of a set of outcomes; that so clearly states both the outcomes – general and specific – that assessors, students and interested third parties can all make reasonably objective judgements with respect to student achievement or non-achievement of these outcomes; and that certifies student progress on the basis of demonstrated achievement of these outcomes. Assessments are not tied to time served in formal educational settings* (Wolf, 1995, p.1)

Grade-based assessment collects less fine-grained information about candidate performance than mark-based assessment at the point of initial assessment judgements. Hurdle-based aggregation methods within units then preserve only the grade achieved across all learning outcomes. It can be seen as a disadvantage of such aggregation methods, both for perceived fairness and for the usefulness of qualification grades, that a candidate who has performed minimally at a particular grade (i.e., is judged to have 'just' met the necessary criteria and no more) is awarded the same grade as a candidate who has met all necessary criteria easily, and has perhaps met a high proportion of higher grade criteria in addition. On the other hand, it could be argued that the level of distinction between candidates is sufficient for the purposes of the qualification: The relevance of grading (at all) in VQs and VRQs can be challenged (e.g., Johnson, 2008c), and although regulated qualifications including Applied Generals and Tech Levels are required to grade and reward achievement (DfE, 2016), this does not necessarily require the fine-grained distinctions enabled by a mark-based approach. It could alternatively be argued that the information loss is significant, but acceptable in light of the overall weight of advantages conferred by a grade-based approach.

As in within-unit aggregation, different methods of aggregating unit-level results preserve different levels of information. A conjunctive approach in which the qualification grade must be met or exceeded in every unit (meaning that the qualification grade is equal to the candidate's lowest unit-level grade) will discard any information about higher achievements by the candidate in selected units. A points-based system in which unit-level grades are converted to points and then added, on the other hand, will capture these achievements, as well as introducing compensation between units.

Reliability

The compensation in mark-based aggregation confers an advantage in terms of reliability. Adding multiple marks allows compensation between positive and negative errors of measurement (not just between student strengths and weaknesses), so that "aggregate marks can reasonably be taken to be a reliable measure of candidates' general performances" (Cresswell, 1988, p.363). The absence of this 'cancelling out' of measurement error is, correspondingly, a disadvantage of aggregation approaches that do not permit compensation. Chester (2003) noted that where results or measures are combined using a conjunctive aggregation rule, "the reliability of the decision (in the example, whether a student should receive a diploma) is that of the least reliable measure" (p.33).

Douglas and Mislevy (2010) used a simulation method to estimate the classification accuracy of component-based assessments or measures formed by different aggregation rules. Considering multiple measures of agreement, Douglas and Mislevy (2010) found that "adding up scores provides the most consistently accurate decision for all students" whilst "the conjunctive rule [hurdle aggregation] does a better job of correctly identifying students who have not acquired the necessary skills at the expense of misclassifying students who in fact have acquired such mastery" (p.302).

Cresswell (1988) suggested that non-compensatory aggregation is appropriate only when knowledge about performance in individual components is the priority. In such cases, the overall grade is required to carry a different meaning to the meaning it holds in GQs. The argument for the greater reliability of mark-based aggregation no longer applies, since component-level results "do not purport to be and should not be judged as measures of general performance" (p.364), and their aggregated marks do not capture what is supposed to be assessed. In the case of a VQ used to certify competence, the disadvantages introduced by hurdle-based aggregation are accepted, since the risk of a 'false positive' (incorrectly certifying competence) is a more serious problem than failing to reward instances of high-achievement. However, VRQs such as Applied Generals, Tech Levels and Technical Awards do not certify competence in an occupation. In fact, it is required that they are not occupation specific and that they reward student achievement (DfE, 2016). A high level of compensation (which risks awarding too generously to students with gaps in their understanding, and obscures the meaning of the grade awarded) therefore seems to represent a lower threat to the assessment purpose than aggregation with little or no compensation (which risks failing to reward achievement in the areas where students have excelled).

Anomalies

The aggregation of unit-level results can result in several kinds of anomaly. Firstly, the aggregation method can cause one given set of unit grades to result in more than one overall grade – as would happen when, for example, units are graded, but unit-level marks are aggregated and final grades are based on the overall aggregated mark scale. Secondly, an aggregation method can result in a 'mismatch' between unit grades and overall grade, for example where a profile of unit-level Merit grades results in an overall Distinction grade. Thirdly, an aggregation method can result in multiple sets of unit grades resulting in the same overall grade. Of these anomalies, the third type is not considered problematic (unless prevalent to a very high degree, for example, almost all unit grade profiles resulting in an overall Merit grade). The first two types, on the other hand, are considered undesirable (Cresswell, 1988; Thomson, 1992).

Mark-based aggregation via UMS avoids the second type of anomaly, since the qualification-level grade boundaries are obtained by adding the UMS grade boundaries at unit level. The first type of anomaly does occur, so that "the same grade profile does not lead to a well-defined overall grade" (Taylor et al., 2006, p.20). However, the variation among overall grades is caused by retaining information about the quality of performances at unit level, which, in Thomson's view, makes such anomalies "justifiable" (Thomson, 1992, p.137), particularly when transparency is ensured by reporting unit-level UMS results.

A chief advantage of grade-based methods is that anomalous qualification-level grades do not occur. Since unit-level grades are themselves the information used to calculate the overall grade, the same set of unit-level grades always results in the same qualification-level grade. A secondary advantage of this is that students are very easily able to assess their progress towards an overall qualification result.

Conclusions

Table 2 summarises the potential advantages and disadvantages of mark-based and grade-based assessment approaches, as used in VRQs.

The overall implications of choosing between mark-based and grade-approaches are not straightforward. Although research has identified numerous assessment features which support marker agreement, these

Table 2: Summary of marking/grading approach advantages and disadvantages

Mark-based approaches	*Grade-based approaches*
▲Detailed criteria in level-based grid provide guidance on distinguishing better responses from poorer responses, supporting validity and marker agreement	▼Grading criteria provide assessors with less information on how to distinguish between responses
▲Assessors able to recognise different levels of achievement at the same criteria	▼Forces assessors to make binary decisions about performance for each criterion
▼Detailed descriptive criteria vulnerable to variable interpretation	▼Concisely written criteria still open to variable interpretation
▼High level of marking task demand due to amount of information given to assessors, likely to weaken marker agreement and also acceptability	▲Brevity of assessment criteria is likely to result in lower marking task demand, which may result in higher marker agreement and help acceptability
▼Ambiguity introduced if a levels-based mark scheme describes multiple skills/ features within a single level of response without guidance on their relative weighting	▲Separate (binary) decision for each criterion is unambiguous
▲Compensation within units supports reliability of unit-level results	▼Hurdle aggregation within units is a risk to the reliability of unit-level grades
▲Aggregation via addition of marks supports reliability of overall results	▲Points-based aggregation of units uses all information captured at unit level
▲High levels of compensation may be perceived as more fair to students	▼Hurdle aggregation within units is hard to justify if certification of competence is not the aim
▲High levels of compensation perhaps better aligned with aim to recognise student achievement	▼Hurdle aggregation within units can be seen as against aim of rewarding success
▼Performance against criteria cannot be deduced from unit or overall grade	▲Unit grade is able to 'guarantee' that candidate has met minimum level of performance against the criteria for that unit

Legend: ▲ = Advantage ▼ = Disadvantage

can be in tension with each other (e.g., providing clarification versus conciseness), and it is not clear what the net effects of particular marking and grading approaches are. In the context of VRQs, there are well-documented challenges that apply to both mark-based and grade-based approaches. Studies in VRQ assessment have not explicitly compared the extent to which mark-based and grade-based approaches succeed in addressing these challenges.

Within VRQs, both mark-based and grade-based approaches assess candidates' performances against assessment criteria. However, there are differences in how they do this. This article has highlighted ways in which these differences reflect different underlying assessment rationales. The features of mark-based approaches to VRQ assessment typically reflect dominant practices from GQs, and an underlying concern with recognising the extent of candidate achievement. By contrast, grade-based assessment in VRQs tends to reflect the strong association of grade-based assessment with criterion-referenced and competency-based assessment.

It is important to reiterate that the advantages and disadvantages of each approach can only be weighed in light of assessment purpose. In the case of VRQs, this may not always be transparent to students and other stakeholders, especially in comparison with better known examples such as GQs and wholly VQs.

References

Acquah, D. K., & Malpass, D. (2015). The technical baccalaureate: providing excellence in vocational education? *Assessment in Education: Principles, Policy & Practice, 24*(1), 96–117.

Ahmed, A., & Pollitt, A. (2011). Improving marking quality through a taxonomy of mark schemes. *Assessment in Education: Principles, Policy & Practice, 18*(3), 259–278.

Baird, J.-A., Beguin, A., Black, P., Pollitt, A., & Stanley, G. (2012). The Reliability Programme: Final report of the Technical Advisory Group. In D. Opposs & Q. He (Eds.), *Reliability Compendium* (pp.771–838). Coventry: Office of Examinations and Qualifications Regulation.

Baume, D., & Yorke, M. (2002). The Reliability of Assessment by Portfolio on a Course to Develop and Accredit Teachers in Higher Education. *Studies in Higher Education, 27*(1), 7–25.

Beach, P. (2018, March 15). *Applied Generals and Tech level qualifications: Additional guidance on the interpretation of Department for Education technical guidance for performance table qualifications* [Open letter]. Coventry: Ofqual. Retrieved from https://www.gov.uk/government/publications/applied-generals-and-tech-levels-additional-guidance.

Cadwallader, S. (2014). *Developing grade descriptions for the new GCSEs: Considerations and challenges*. Manchester: AQA Centre for Education Research and Practice.

Carter, A., & Bathmaker, A.-M. (2017). Prioritising progression over proficiency: Limitations of teacher-based assessment within technician-level vocational education. *Journal of Further and Higher Education, 41*(4), 460–474.

Chester, M. D. (2003). Multiple Measures and High-Stakes Decisions: A Framework for Combining Measures. *Educational Measurement: Issues and Practice, 22*(2), 32–41.

Cresswell, M. J. (1987). Describing examination performance: Grade criteria in public examinations. *Educational Studies, 13*(3), 247–265.

Cresswell, M. J. (1988). Combining grades from different assessments: How reliable is the result? *Educational Review, 40*(3), 361–382.

Cresswell, M. J. (1994). Aggregation and awarding methods for National Curriculum assessments in England and Wales: A comparison of approaches proposed for Key Stages 3 and 4. *Assessment in Education: Principles, Policy & Practice, 1*(1), 45–62.

DfE. (2016). *Technical and applied qualifications for 14 to 19 year olds. Key stage 4 and 16 to 19 performance tables from 2019: technical guidance for awarding organisations*. London: Department for Education.

Douglas, K. M., & Mislevy, R. J. (2010). Estimating Classification Accuracy for Complex Decision Rules Based on Multiple Scores. *Journal of Educational and Behavioral Statistics, 35*(3), 280–306.

Eraut, M., Steadman, S., Trill, J., & Parkes, J. (1996). *The assessment of NVQs*. Research Report No. 4: University of Sussex Institute of Education.

Forrest, G. M., & Shoesmith, D. J. (1985). *Monitoring standards in the General Certificate of Secondary Education*. Manchester: Joint Matriculation Board.

Good, F. J., & Cresswell, M. (1988). *Grading the GCSE*. London: The Secondary Examinations Council.

Gray, E., & Shaw, S. (2009). De-mystifying the role of the uniform mark in assessment practice: concepts, confusions and challenges. *Research Matters: A Cambridge Assessment publication, 7*, 32–37.

Greatorex, J. (2001, September). *Can vocational A levels be meaningfully compared with other qualifications?* Paper presented at the British Educational Research Association Conference, University of Leeds,UK.

Greatorex, J., & Shannon. (2003, September). *How can NVQ Assessor's judgements be standardised?* Paper presented at the British Educational Research Association Conference, Heriot-Watt University, Edinburgh, UK.

Johnson, M. (2008a). Assessing at the borderline: Judging a vocationally related portfolio holistically. *Issues in Educational Research, 18*(1), 26–43.

Johnson, M. (2008b). Exploring assessor consistency in a Health and Social Care qualification using a sociocultural perspective. *Journal of Vocational Education & Training, 60*(2), 173–187.

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

Johnson, S. (2011). *A focus on teacher assessment reliability in GCSE and GCE*. Coventry: Office of Qualifications and Examinations Regulation.

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education, 28*(1), 91–105.

Johnston, B. (2004). Summative assessment of portfolios: An examination of different approaches to agreement over outcomes. *Studies in Higher Education, 29*(3), 395–412.

Nádas, R., Suto, I., & Grayson, R. (2012, September). *Analyse, evaluate, review….How do teachers with differing subject specialisms interpret common assessment vocabulary?* Paper presented at the European Conference for Educational Research, University of Cádiz, Spain.

OCR. (2017). *Cambridge Nationals in ICT – Specification* (Version 4 – July 2017). Retrieved from http://www.ocr.org.uk/Images/82572-specification.pdf

Pinot de Moira, A. (2013). *Features of a levels-based mark scheme and their effect on marking reliability*. Centre for Education Research and Policy paper. Manchester: AQA.

Suto, I., & Nádas, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education, 23*(4), 477–497.

Taylor, M., Pritchard, J., & Gray, E. (2006). *Grading the specialised diploma*. Manchester: AQA.

Thomson, D. G. (1992). *Grading modular curricula. Final report of the GCSE modular aggregation research and comparability study*. Cambridge: Midland Examining Group.

Tisi, J., Whitehouse, G., Maughan, S., & Burdett, N. (2013). *A review of literature on marking reliability research (Report for Ofqual)*. Slough: National Foundation for Educational Research.

Wolf, A. (1993). *Assessment issues and problems in a criterion-based system*. London: Further Education Unit.

Wolf, A. (1995). *Competence-based assessment*. Buckingham: Open University Press.

Wolf, A. (1998). Portfolio Assessment as National Policy: the National Council for Vocational Qualifications and its quest for a pedagogical revolution. *Assessment in Education: Principles, Policy & Practice, 5*(3), 413–445.

Wyatt-Smith, C., & Klenowski, V. (2013). Explicit, latent and meta-criteria: types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy & Practice, 20*(1), 35–52.

Appendix

Level	*Marks*	*Description*
3	15–20	Accurate and detailed knowledge of … is demonstrated. The response is highly focused on the question. Evaluation/discussion is effective and logically developed.
2	8–14	Generally accurate knowledge of … is demonstrated, with some omissions. The response is generally focused on the question. Some effective evaluation/discussion is present.
1	1–7	Limited knowledge of … is demonstrated. The response relates to the topic but addresses the question only to a limited extent. Response is mainly descriptive. Evaluation/discussion is limited or absent.
	0	No relevant content.

Figure A1: Holistic levels-based mark scheme – hypothetical example

Level	*Strand 1*		*Strand 2*		*Strand 3*	
	Marks	*Description*	*Marks*	*Description*	*Marks*	*Description*
3	6–8	Accurate and detailed knowledge of … is demonstrated.	5–6	The response is highly focused on the question	5–6	Evaluation/discussion is effective and logically developed.
2	3–5	Generally accurate knowledge of … is demonstrated, with some omissions.	3–4	The response is generally focused on the question.	3–4	Some effective evaluation/ discussion is present.
1	1–2	Limited knowledge of … is demonstrated.	1–2	The response relates to the topic but addresses the question only to a limited extent.	1–2	Response is mainly descriptive. Evaluation/discussion is limited or absent.
	0	No relevant content.	0	No relevant content.	0	No relevant content.

Figure A2: Analytic levels-based mark scheme – adapted from the holistic example in Figure A1