Research and Analysis

Overview - Grading Vocational & Technical Assessments

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

Contents

Overview

In 2017, Ofqual initiated a programme of research into grading within vocational and technical assessments. Taking a broad look at grading, we have explored policies, principles, and practices related to the grading of Vocational and Technical Qualifications (VTQs) in England, enabling us to deepen our engagement with such issues. We have supplemented this with a literature review focussing on grading within Technical and Vocational Education and Training (TVET) contexts in Australia – the only country with a significant body of relevant research and analysis to draw upon. We have released the products of these two pieces of work as:

1. *Grading Vocational & Technical Qualifications: Recent policies and current practices* (Newton, 2018a).
2. *Grading Competence-Based Assessments: Notes from a small literature* (Newton, 2018b).

The present report provides a general introduction to these two documents, explaining our rationale for producing them.

Grading Vocational & Technical Assessments

The focus of our programme is grading, by which we mean the award of higher grades beyond the passing grade. Grading is standard practice in relation to General Qualifications (GQs); such as the A level, which awards 5 higher grades (A* to D) beyond the passing grade (E). In TVET contexts, however, grading is not always standard practice, and its popularity has waxed and waned over time.

Grading became less common in TVET contexts, in England, with the rise of the Competence-Based Assessment (CBA) movement, which was associated with the introduction of National Vocational Qualifications (NVQs). CBA recommends a binary approach to recognising proficiency; meaning that candidates are assessed as either competent (Pass) or not-yet-competent (Fail). From this perspective, higher grades, such as Merit or Distinction, are simply irrelevant. Instead, candidates are differentiated purely in terms of the level at which they are entered for a qualification or apprenticeship (eg by seeking a Level 3 certificate rather than a Level 2 one). The design of many regulated VTQs in England has been heavily influenced by the CBA movement.

Over the past few years, however, grading has been promoted within a succession of high profile TVET reviews, from Wolf (2011), to Richard (2012), to Whitehead (2013). It is now Government policy to promote grading within VTQs and in new Apprenticeships; both to motivate learners to achieve a high level of proficiency, and to provide qualification users with high quality information on candidates’ proficiency levels.

This renewal of interest in grading raises fundamental questions of assessment design. In particular, it asks: what does good practice in grading in TVET contexts look like, and how does it differ, if at all, from good practice in grading in other contexts? It was in response to this question that we initiated our programme of research and analysis into grading within vocational and technical assessments. The purpose of the present document is to introduce the first two products from this programme, which are cited in full above, and which we shall abbreviate as:

1. *Grading VTQs*; and

2. *Grading CBAs*.

Because these reports include a number of relatively unfamiliar technical terms, and also introduce some entirely new technical terms, their publication is accompanied by a specially prepared *Glossary*.

The two reports explore policies, principles, and practices related to grading in TVET contexts in England and Australia, helping us to deepen our engagement with such issues. Each document contains an Executive Summary, providing a helpful introduction to the research.

The main body of the first report, *Grading VTQs*, presents results from a small-scale survey of grading practices across a sample of 18 regulated qualifications. As the first survey of its kind, it adopted a 'deep-dive' approach, exploring in detail how each of the sampled qualifications operates. This was based upon documentary analysis, supplemented by conversations with awarding organisation representatives.

The research identified a wide variety of grading practices. These were classified and discussed in terms of their underlying measurement models, and in terms of how they represented their measurement standards. Questions were identified relating to a variety of fundamental technical issues; including standardisation, grading and levelling, comparability, weighting, burden and backwash, and transparency.

The main conclusion from this first report is that VTQ grading, in England, is not underpinned by a straightforward, generally accepted, set of principles governing good practice. This raised the question of what such principles might look like.

The second report, *Grading CBAs*, investigated what the literature has to say about principles of good practice. Unfortunately, it became apparent that there is no authoritative literature on grading in TVET contexts. However, we identified a small body of work on CBA grading, from Australia, where a national debate on grading spans the best part of three decades. It appears that grading practices in Australia are at least as divergent as in England, if not more so. However, there has been far more discussion over why this is the case; and particular attention has been paid to critical lines of divergence, including the legitimacy of different kinds of grading criteria.

The main conclusion from this second report is that, despite repeated attempts to identify principles of good practice for grading in TVET contexts, the Australians have had only limited success.

Why is grading even an issue?

Both reports raise fundamental questions concerning grading in TVET contexts. Yet, on reflection, it is not unreasonable to ask why this issue should even be up for debate. After all, the idea of grading is not new. Learners have been graded in all sorts of educational contexts, all over the world, for decades and decades. Should we not already know what good practice in grading vocational and technical assessments looks like?

In fact, there are all sorts of reasons that make it hard for us to know exactly what good practice looks like, including these:

1. it can be tricky to pin grading down;
2. grading theory is still maturing;
3. grading in TVET contexts is potentially enigmatic;
4. grading purposes can be tricky to disentangle; and
5. there has been relatively little research and analysis.

Providing an introduction to why we need to engage more deeply with the issue of grading in TVET contexts, each of these reasons is briefly explored, below. The second, third and fourth reasons are explored in more depth in the two main reports.

It can be tricky to pin grading down

Grading is often discussed as though it were a discrete process that can be separated from the wider assessment procedure within which it is located, and studied in isolation. Indeed, at one end of the continuum of grading definitions, the narrow end, this is more-or-less true. However, at the other end of the continuum of grading definitions, the broad end, this could not be further from the truth.

At the narrow end, grading is no more nor less than the approach that is adopted to classifying learners, who have already been rank ordered (via an assessment process) into meaningful groups. Indeed, the sole purpose of this grouping process is to add meaning to the rank ordering, so that assessment results can be interpreted accurately and usefully. The smallest number of groups into which any cohort of candidates might be divided is two, ie those who pass versus those who fail. The cut-off mark that separates those who pass from those who fail might be decided with reference to a certain proficiency; for example, the minimum level of proficiency required to practise safely and competently within an occupational field. Alternatively, it might be decided with reference to a certain group; for example, the mark that separates the top 70% of the cohort from the bottom 30%.

At the narrow end of the continuum, grading is theorised and practised somewhat differently, depending on the purpose(s), context(s), and population(s) targeted by the assessment procedure. Approaches can be classified in various ways, for example:

1. norming (eg Kolen, 2006);[1]
2. linking and equating (eg Holland and Dorans, 2006);[2] and
3. setting performance standards (eg Hambleton and Pitoniak, 2006).[3]

Grading in GQ contexts, in England, would tend to be located towards this narrow end. GQ grading tends to refer to the process by which grade boundaries are located along component-level mark scales, from which qualification-level grade boundaries are derived. This involves establishing a link between grade boundary standards on one examination (eg this year's physics A level) and grade boundary standards on another (eg last year's physics A level). It is therefore an example of category 2, above.

Conversely, at the broad end of the continuum, grading cannot be separated from the procedure within which it is located. At this end, grading tends to refer to the assessment procedure itself, including many if not all of the features and processes that comprise it (including the process for eliciting evidence, the process for evaluating performances, the process for aggregating information, the process for reporting results, and so on). At this end of the continuum, an assessor will be directly responsible for grading learners – that is, they will make an overall grading judgement, or they will aggregate a series of lower-level grading judgements – and it might be entirely up to them how they choose to do so; indeed, they may even choose to do so differently for different learners.

Grading in many local, school-based contexts – particularly in countries like the USA with a tradition of relatively high stakes school-based assessment – might be located towards the broad end of the continuum of grading definitions. To the extent that grading, defined like this, incorporates a plethora of features and processes, it cannot be neatly classified procedurally. This also makes it harder to identify principles of good practice for grading at this end of the definitional continuum.

Amongst the 18 sampled qualifications discussed within *Grading VTQs*, some operated grading in the narrow sense. For these qualifications, grading was simply a matter of determining grade boundaries, on a mark scale derived for a unit test, to classify candidates into one grade or another. Where the qualification comprised two or more unit tests, this introduced an additional element of aggregation, albeit sometimes fairly trivially so (eg unit Distinction + unit Distinction = Distinction overall). Other qualifications within this report, however, operated grading in a broader sense, which could not be reduced to grade boundary determination. For many of these qualifications, dozens if not hundreds of criterion-level performance-grading judgements ultimately contribute to the overall qualification grade; typically rendering grading and aggregation intrinsically and non-trivially intertwined.

Within *Grading CBAs*, approaches even further towards the broad end of the continuum were identified; for instance, where assessors made a single grading

[1] Expressing the results (ie grades) of candidates from a particular cohort relative to the performance of candidates from a known population (the norm-group).

[2] Applying the same standards across two or more assessments; often by establishing a link between the minimum mark that is worthy of each grade on each assessment.

[3] Determining new standards for a particular assessment; often by determining the minimum mark that is worthy of each grade.

judgement for each learner, based upon the entire body of evidence collated during their course of learning.

Grading theory is still maturing

Although research into grading – defined both narrowly and broadly – can be traced back well over a century (eg Latham, 1886), a highly influential conceptual distinction was drawn just half a century ago; and its implications, both technical and educational, are still being worked through. This distinction, between norm-referencing and criterion-referencing, was drawn in the early 1960s by Robert Glaser, who was working in the USA (eg Glaser, 1963). It concerned the meaning that is attached to assessment results, and, more specifically, whether this is defined relative to:

- a specified group of learners (norm-referencing), eg able to perform better than 90% of the year group; or to
- specified proficiency profiles (criterion-referencing), eg able to perform X, Y, and Z.

This distinction was technically significant, because it encouraged assessment designers and developers to make assessment results easier to interpret. It was also educationally significant, because it set learners the (self-directed) goal of attaining specific learning outcomes, rather than setting them the (others-directed) goal of attaining learning outcomes better than their peers. Both of these features – technical and educational – made criterion-referencing politically attractive, on an international scale. In England, Government promoted criterion-referencing heavily during the early- to mid-1980s. Consequently, England's public examinations took on certain of the trappings of criterion-referencing from the late-1980s onwards.

Unfortunately, the theory of criterion-referencing has often been misunderstood (eg Glaser, 1994; Linn, 1994), and frequently misapplied (eg Popham, 1994). In England, although it is widely assumed that public examinations transitioned from being norm-referenced (from the 1950s to mid-1980s) to being criterion-referenced (from the late-1980s onwards), the truth is that they were never strictly norm-referenced, and they did not become strictly criterion-referenced (Newton, 2011).[4] It is perhaps better to say that they have always been attainment-referenced; and that they have changed over time more in terms of practices than in terms of principle.

Grading in TVET contexts is potentially enigmatic

Assessment in TVET contexts did, however, begin to change radically towards the end of the 1980s, both in terms of practices and in terms of principle. This change was orchestrated through the introduction of National Vocational Qualifications (NVQs), which were strongly criterion-referenced. The proficiency profile to which each NVQ was referenced was an occupational or professional standard of competence. This version of criterion-referencing, which emerged during the 1990s, exemplified a form of CBA that was characterised by:

[4] See also https://ofqual.blog.gov.uk/2017/03/17/mythbusting-3-common-misconceptions/

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

The original idea of the NVQ was that it should be defined purely in terms of outputs from learning (ie having attained specified learning outcomes), and not at all in terms of inputs to learning (ie having followed a course of a certain duration). Its corollary was that a learner could be assessed as competent, and so certificated, at their own pace; whether that required less time than the conventional course of learning, or more. The only important issue, from this perspective, was whether or not the learner had achieved the requisite competence to practise. If they had, then they should receive their certificate; if not, then they should continue learning. As noted above, the idea of grading beyond the competence threshold is irrelevant, here, and NVQs were not graded.

Grading CBAs charts the 'Grade Debate' – the national debate on grading in TVET contexts – as it unfolded in Australia from the mid-1990s onwards. This debate focused upon whether it is desirable and feasible to grade CBAs. Some key Australian stakeholders disagreed with both propositions. Other stakeholders, assuming that it was desirable, explored ways in which it might be made feasible.

Results presented in *Grading VTQs* illustrate how the core characteristics of CBA – including atomistic specification, mastery measurement, and exhaustive sampling – have strongly influenced the design of many current VTQs in England. They also illustrate a multiplicity of ways in which grading can be operationalised within CBA-influenced qualifications.

Both *Grading VTQs* and *Grading CBAs* emphasise that grading in TVET contexts does not need to be operationalised in terms of the traditional CBA model. Indeed, both raise the question of whether there are circumstances in which alternative models might be more suitable.

Grading purposes can be tricky to disentangle

Assessment purposes can be viewed from a number of quite different perspectives; but the two most important are the information perspective and the engagement perspective. From the information perspective, the purpose of an educational assessment is to provide a certain kind of information about a learner. Educational assessment results are generally designed to provide information concerning a learner's level of proficiency[5] in a domain of learning.

Conversely, from the engagement perspective, the purpose of an educational assessment is to secure a certain kind of engagement between the learner and their

[5] Also known as attainment, or competence.

course of learning. Educational assessment procedures are often designed to help motivate learners; and one way of doing so is via grading, that is, by recognising the attainment of higher levels of proficiency.

Generally speaking, assessment design is driven primarily by the information perspective (to provide accurate and useful information), with the engagement perspective as a secondary consideration (to motivate candidates). Assessment, and grading in particular, can go wrong when assessment design decisions are driven too heavily by the engagement perspective, with too little consideration given to the information perspective.

Problems can also arise when those who use assessment results interpret them differently from how they have been designed to be interpreted. As noted above, educational assessment results are generally designed to provide information concerning a learner's current level of proficiency in a domain of learning. Often, though, assessment results are interpreted as though they somehow indicated a learner's aptitude, ie their potential for achieving success in the future. The legitimacy of this over-interpretation may, to some extent, depend on the conditions of learning associated with the assessment result. For instance, when it can be assumed that learners have all followed a course of fixed duration, the over-interpretation may be more legitimate than when this cannot be assumed. *Grading CBAs* discusses a number of 'thought experiments' along these lines.

There has been relatively little research and analysis

Compared with the situation for conventional school-based assessments, there has been relatively little research and analysis into the technical functioning of vocational and technical assessments. This is true not just for grading, but for all aspects of technical functioning.

In England, a strong tradition of GQ research and analysis has existed for the best part of a century (see Crofts and Caradog Jones, 1928; Petch, 1953). It has focused particularly upon issues of standards and comparability (eg Bardell, Forrest and Shoesmith, 1978; Forrest and Shoesmith, 1985; Newton, et al, 2007; Baird, et al, 2018), including approaches to grading (eg Whittaker and Forrest, 1983). This tradition has been driven largely, and often collaboratively, by the awarding organisations; although regulatory bodies have also been actively involved (eg Secondary Schools Examinations Council, 1932; Schools Council, 1979; Qualifications and Curriculum Authority, 1999). In recent years, Ofqual has played a key role (eg Ofqual, 2016; Holmes and Rhead, 2018).

Yet, a similar tradition has not emerged for VTQs. Having said that, it is certainly possible to point to many examples of research and analysis related to VTQs in England (eg Black, He and Holmes, 2017; Boyle and Rahman, 2013; Curcin et al, 2014; Ecclestone, 2002; Greatorex, 2005; Isaacs, 2013; Jessup, 1991; Johnson, 2008; Murphy, et al, 1995; Oates, 2004; Qualifications and Curriculum Authority, 2006; Smith, 1996; Wolf, 1995). However, there has never been an identifiable VTQ research and analysis community; the outputs have tended to be sporadic and isolated rather than cumulative; and the amount of work produced over the years has remained quite small, certainly in comparison with the situation for GQs. As noted earlier, *Grading VTQs* documents the first survey of its kind into VTQ grading practices.

Where does this leave us?

For each of these reasons, and others too, it is hard for the assessment profession to pinpoint exactly what good practice in grading in TVET contexts looks like. This has motivated Ofqual to become more deeply engaged with this issue. It is not simply that grading raises highly technical challenges, which defy straightforward resolution, eg standardisation, the relationship between grading and levelling, comparability, weighting, and transparency. It is also that the nature of these technical challenges will differ markedly according to how it is operationalised; and, more fundamentally, how it is conceptualised, eg narrowly or more broadly. Much work, both empirical and analytical, remains to be done to explore these issues and their consequences in depth.

Grading is not just a technical matter, but an educational one, too. In addition to the potential of grading to engage learners with their course of learning, its potential to disengage both learners and their teachers/trainers needs also to be recognised; for instance, when grading practices are poorly designed, or simply take up too much time. All of these factors, technical and educational, need to be weighed against each other when designing optimal grading models and practices.

Where does this leave us? Unfortunately, it leaves us closer to the beginning of a dialogue than to its resolution. Building upon the Grade Debate in Australia, the accompanying reports help to provide us with a solid foundation upon which to deepen our engagement with grading issues in TVET contexts in England.

Key issues from *Grading VTQs* and *Grading CBAs* will be presented and discussed at a conference entitled *Driving Good Practice in Grading Vocational and Technical Assessments*, scheduled for 11 December 2018. We hope that this will mark the beginning of a broader conversation on grading vocational and technical assessments amongst scholars, policy makers, and practitioners in England.

References

Baird, J., Isaacs, T., Opposs, D. and Gray, L. (2018). *Examination Standards: How measures and meanings differ around the world*. London: University College London Institute of Education Press.

Bardell, G.S., Forrest, G.M. and Shoesmith, D.J. (1978). *Comparability in GCE*. Manchester: Joint Matriculation Board.

Black, B., He, Q. and Holmes, S.D. (2017). *Vocational and Technical Qualifications: Assessment functioning of external assessments. An overview of the functioning of assessments in 27 qualifications and 49 units.* Ofqual/17/6319. Coventry: Office of Qualifications and Examinations Regulation.

Boyle, A. and Rahman, Z. (2013). *The Internal Reliability of Some City & Guilds Tests*. Ofqual/13/5257. Coventry: Office of Qualifications and Examinations Regulation.

Crofts, J.M. and Caradog Jones, D. (1928). *Secondary School Examination Statistics*. London: Longmans, Green and Co.

Curcin, M., Boyle, A., May, T. and Rahman, Z. (2014). *A Validation Framework for Work-Based Observational Assessment in Vocational Qualifications*. Ofqual/14/5374. Coventry: Office of Qualifications and Examinations Regulation.

Ecclestone, K. (2002). *Learning Autonomy in Post-16 Education: The politics and practice of formative assessment*. Oxford: RoutledgeFalmer.

Forrest, G.M. and Shoesmith, D.J. (1985). *A Second Review of GCE Comparability Studies*. Manchester: Joint Matriculation Board (on behalf of the GCE Examining Boards).

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

Glaser, R. (1994). Criterion-referenced tests: Part II. Unfinished Business. *Educational Measurement: Issues and Practice*, 13 (4), 27–30.

Greatorex, J. (2005). Assessing the evidence: different types of NVQ evidence and their impact on reliability and fairness. *Journal of Vocational Education & Training*, 57 (2), 149-164.

Hambleton, R.K. and Pitoniak, M.J. (2006). Setting performance standards. In R.L. Brennan (Ed.). *Educational Measurement* (4th ed., pp.433–470). Westport, CT: American Council on Education. Praeger Press.

Holland, P.W. and Dorans, N.J. (2006). Linking and equating. In R.L. Brennan (Ed.). *Educational Measurement* (4th ed., pp.187–220). Westport, CT: American Council on Education. Praeger Press.

Holmes, S.D and Rhead, S. (2018). *A level and AS mathematics: An evaluation of the expected item difficulty*. Ofqual/18/6344. Coventry: Office of Qualifications and Examinations Regulation.

Isaacs, T. (2013). The diploma qualification in England: an avoidable failure? *Journal of Vocational Education & Training*, 65 (2), 277-290.

Jessup, G. (1991). *Outcomes: NVQs and the emerging model of education and training*. London: The Falmer Press.

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

Kolen, M.J. (2006). Scaling and norming. In R.L. Brennan (Ed.). *Educational Measurement* (4th ed., pp.155–186). Westport, CT: American Council on Education. Praeger Press.

Latham, H. (1886). *On the Action of Examinations Considered as a Means of Selection*. Boston: Willard Small.

Linn, R.L. (1994). Criterion-referenced measurement: A valuable perspective clouded by surplus meaning. *Educational Measurement: Issues and Practice*, 13 (4), 12–14.

Murphy, R., et al (1995). *The Reliability of Assessment of NVQs*: Report presented to the National Council for Vocational Qualifications. Nottingham: School of Education, University of Nottingham.

Newton, P.E. (2011). A level pass rates and the enduring myth of norm-referencing. *Research Matters*, Special Issue, 2, 20–26.

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

Newton, P.E. (2018b). *Grading Competence-Based Assessments: Notes from a small literature*. Coventry: Office of Qualifications and Examinations Regulation.

Newton, P.E., Baird, J., Goldstein, H., Patrick, H., & Tymms, P. (2007). *Techniques for Monitoring the Comparability of Examination Standards*. London: Qualifications and Curriculum Authority.

Oates, T. (2004). The role of outcomes-based national qualifications in the development of an effective vocational education and training system: the case of England and Wales. *Policy Futures in Education*, 2 (1), 53-71.

Ofqual (2016). *An Investigation into the 'Sawtooth Effect' in GCSE and AS / A level Assessments.* Ofqual/16/6098. Coventry: Office of Qualifications and Examinations Regulation.

Petch, J.A. (1953). *Fifty Years of Examining*. London: Harrap & Co.

Popham, J.W. (1994). The instructional consequences of criterion referenced clarity. *Educational Measurement: Issues and Practice*, 13 (4), 15–18.

Qualifications and Curriculum Authority (2006). *Comparability Study of Assessment Practice*: Personal licence holders. London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (1999). *Five Yearly Review of Standards in 16+ Examinations: 1976-1996*. London: Qualifications and Curriculum Authority.

Richard, D. (2012). *Richard Review of Apprenticeships*. Available online: https://www.gov.uk/government/publications/the-richard-review-of-apprenticeships

Schools Council (1979). *Standards in Public Examinations: Problems and possibilities. Occasional Paper 1*. London: Schools Council.

Secondary School Examinations Council. (1932). *The School Certificate Examination. Being the report of the panel of investigators appointed by the Secondary School Examinations Council to enquire into the eight approved School Certificate examinations held in the Summer of 1931*. London: HMSO.

Smith, V. (1996). The General National Vocational Qualification experience: an education or just a qualification? *Research in Post-Compulsory Education*, 1 (3), 373-384.

Whitehead, N. (2013). *Review of Adult Vocational Qualifications in England*. London: UK Commission for Employment and Skills. Available online: https://www.gov.uk/government/publications/review-of-adult-vocational-qualifications-in-england--2

Whittaker, R.J. and Forrest, G.M. (1983). *Problems of the GCE Advanced Level Grading Scheme*. Manchester: Joint Matriculation Board.

Wolf, A. (1995). *Competence-Based Assessment*. Berkshire: Open University Press.

Wolf, A. (2011). *Review of Vocational Education: The Wolf Report*. Available online: https://www.gov.uk/government/publications/review-of-vocational-education-the-wolf-report

Published by:

ofqual

Earlsdon Park
53-55 Butts Road
Coventry
CV1 3BH

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

November 2018

Ofqual/18/6441/1