2016-09-21

Rigour in moderation processes is more important than the choice of method

Zahra, Daniel

10.1080/02602938.2016.1236183
Assessment & Evaluation in Higher Education
Routledge

Rigour in moderation processes is more important than the choice of method

Abstract

Processes for moderating assessments are much debated in higher education. The myriad approaches to the task vary in their demands on staff time and expertise, and also in how valid, reliable, and fair-to-students they appear. Medical education, with its diverse range of assessments and assessors across clinical and academic domains presents additional challenges to moderation. The current review focusses on medical education, considering double-marking and benchmarking as two broad classes of moderation procedure, and argues that it is the process more than the type of procedure which is crucial for successful moderation. The objective and subjective advantages and disadvantages of each class of procedure are discussed in light of our medical school's current practices, and with respect to the limited empirical evidence within medical education assessment.

Consideration of implementation is central to ensuring valid and reliable moderation. The reliability of assessor judgements depends more on the consistency of

assessment formats and the application of clear and agreed assessment criteria than on the moderation process itself. This article considers these factors in relation to their impact on the reliability of moderation, and aims to help assessors and students appreciate the diversity of these factors by facilitating their consideration in the assessment process.

Keywords: moderation; double-marking; benchmarking, medical education

Introduction

The central aim of moderation in assessment is to ensure uniformity of assessment standards – be this through shared understanding of criteria, expected knowledge, or consistent application of a marking scheme. Ideally moderation will achieve consensus between assessors in the marks awarded and standards that need to be achieved, reducing the impact not only of particularly stringent assessors but also overly lenient ones (Sadler, 2013). Although a range of different procedures appear in the literature for achieving this, comparisons between them can be misleading. Because of the variety of decisions which need to be made about how to implement moderation procedures, no two procedures are likely to share more than the broadest details.

Accepting the difficulty of trying to directly compare 'types' of moderation procedure, the many variations fall broadly into two common classes. Double (or multiple)-marking which typically involves every piece of student work being marked by two (or more) assessors, and benchmarking; typically involving all assessors marking a selection of student work or exemplar pieces. In both cases the marks awarded are compared in order to agree assessment criteria interpretation, weighting,

and final mark, and to determine whether any assessors are marking in a manner not consistent with the others.

However, as this review will emphasise, it is the specifics of implementation which should be the focus of moderation rather than the choice of procedure. With this in mind we take 'rigour' of implementation to mean the extent to which individuals engage in the process, as distinct from 'rigorous' in reference to the comprehensiveness of particular types of assessment. Similarly, the validity of an assessment should be evaluated independently of the moderation of that assessment. The structure, content, and delivery of an assessment should be appropriate to the skills and knowledge being assessed, as well as the context of the assessment (e.g. Brown, Bull, & Pendlebury, 1997). This validity is distinct from the issue of assessment reliability, which is discussed in relation to moderation below. Although the type, structure, and context of an assessment may necessitate a particular approach to moderation for logistic reasons, it does not detract from the central message of this article – that it is the specifics of how moderation is implemented that is of most importance, as opposed to whether the process might be considered double-marking or benchmarking.

Both classes of moderation procedure aim to reduce outlying marks and develop good inter-assessor reliability, yet how to define an 'outlier' and 'good' reliability are problems that need to be addressed by both. Similarly, the lack of research on the impact of assessors who give consistently mid-range 'defensive' marks (Hornby, 2003) suggests that in any moderation process, there is disproportionately more effort spent identifying outlying markers, and on achieving agreement between assessors irrespective of possible confounding factors. Furthermore, this review focusses on medical education, where the diverse range of assessment types highlight the need to

consider the specifics of each approach as well as the relative lack of empirical work on moderation in medical education.

Effect of assessment format and criteria on assessor reliability

It has been suggested that inter-assessor reliability (that is, the consistency in mark between assessors) for high-stakes medical education assessments should show correlations of r= 0.70-0.80 (Roberts, Shadbolt, Clark, & Simpson, 2014). Average inter-assessor reliability is in fact typically ~0.60, is affected by a wide range of factors, and will vary across submissions (e.g. Bloxham & Price, 2013, Elton & Johnston, 2002; Roberts et al, 2014).

Reliability is improved when there are explicit outcomes against which to judge assessments (Baume & Yorke, 2002), and there is also some evidence that reliability can be improved by marking work on an element-by-element (serial) basis (Nystrand, Cohen, & Dowling, 1993); this is in contrast to assessments for which each assessor provides a single, overall (holistic) grade(see Mitchell & Anderson, 1986, for discussion of this approach in the essay component of the Medical College Admission Test). As an example of this, Baume, Yorke, and Coffey (2004) report a study of portfolios in which 60% of comparisons between element scores, but only 39% of comparisons between overall outcomes, showed exact agreement. Furthermore, when outcomes within one grade of each other were considered a match, agreement rose to 80%. This not only highlights the impact of a change in process, but also the importance of definitions of agreement. These results are similar to those found in other evaluations and across clinical competency assessments (Pitts, Coles, & Thomas, 1999). Portfolio assessments demonstrate the range of challenges to reliability as elements may also take different formats. In this case reliability needs to be maintained within and across format types as well as assessors. Reliability may be improved in such assessments by altering the

assessments themselves (e.g. uniform rating scales across components, clearly defined criteria) rather than the moderation processes, yet it also highlights the need to consider what would constitute consistent assessment between markers and across formats. Students within one grade of each other may appear similar, but if one grade is a passing one and the other a failing one, the difference to the student is likely to seem much larger than assessors who consider one grade difference 'agreement'. Furthermore, whether and to what extent greater or lesser variability would be accepted across different element formats will have similar implications for both student perceptions and whether the assessors consider themselves to be in agreement. The level of variability acceptable in the moderation process should take into account the type and context of the assessment.

The work of Baume and Yorke (2002), Nystrand et al (1993) and Pitts et al (1999) make it clear that the structure of an assessment affects reliability, and this in turn emphasises the need to consider how they are moderated – for example, whether individual element scores are considered, or overall scores are compared. Knowing that factors such as these impact marks, it should be clear that similar factors will also affect the moderation process. Even when detailed marking criteria are employed, they are often open to interpretation and applied differently depending on whether the assessment is summative (e.g. leads to graduation or professional certification) or formative (e.g. used to evaluate learning; Gibbs & Simpson, 2004). Use of criteria will also vary depending on whether an assessor holds a positivist (objective standard) versus interpretationist (relative) view of the skills to be assessed (Elton & Johnston, 2002). This will subsequently have implications for assessor reliability in that it impacts whether marks are awarded by norm or criterion referencing, and highlights again the need for moderation processes to give due consideration to shared understanding

between markers of not only the definitions of the criteria, but also how judgements relate to both the rating scale and group (criterion or cohort norm) against which the assessment of performance is made.

This discussion of criterion and norm referencing of assessment standards may seem unnecessary within some disciplines where there has been a trend towards criterion referencing. However, given our focus on medical education, within which there remains widely used methods of norm-referenced standard setting (within both double-marking and benchmarking) the distinction is important and one which should be considered in the moderation of student work.

With respect to individual assessment criteria, although they should be applied to each piece of work independently of other submissions, there is always likely to be an element of cross-student comparison (Bloxham, Boyd, & Orr, 2011). The only way of overcoming this form of norm-referencing would be for each assessor to only assess one submission, thus making moderation between assessors impossible. This norm-referencing may be more of an issue when benchmarking is employed, where a smaller number of scripts are seen by all assessors with potentially varying levels of expertise, thus giving a less detailed, less representative, and potentially biased picture of the overall distribution. Even with full cohort double-marking, norm-referencing may lead only to agreement in rank order, but not necessarily specific marks. In both cases, and even if each assessor were to only mark one script, informal discussions between assessors may lead to cross-student comparisons whether intentional or subconscious. On the one hand, this would seem to undermine the moderation processes. On the other, however, it highlights the need to take into account the wider context within which moderation takes place, and how these factors might be incorporated or at least acknowledged within the process. One thorough approach to this is a social

constructionist perspective outlined by Rust, O'Donovan, and Price (2005). The proposal in their model is to create the constructs, knowledge, and assessment criteria through shared discourse. Specifically, learning and increased understanding of the assessment criteria develops as an emergent property of engagement with the moderation process and shared ownership of the criteria (see Elwood & Klenowski, 2002, for an interesting discussion). This is likely to increase engagement and parity in the application of these criteria when marking, and can be seen in anecdotal accounts of assessor training and moderation in our own school (work currently in progress). Whatever philosophical approach is adopted however, features such as constructive alignment and explicit shared assessment criteria are clearly essential (e.g. Fry, Ketteridge, & Marshall, 2003).

The use of criteria raises the issue of how to determine whether a particular skill or area of knowledge has been addressed within each submission, and what to do if there is insufficient evidence for a given skill – whether that be in a given assessment, over the course of a year, through issues with the assessment, or circumstances beyond the control of assessors and students. Direct knowledge of the students being assessed has been shown to reduce disagreement over such elements, yet assumptions about other student characteristics can bias marking (Baume & Yorke, 2002). This again suggests that reliabilities can be altered by changing how the moderation process is conducted. For example, increasing the uniformity of knowledge about students is likely to increase reliability but may reduce the validity as irrelevant factors (e.g. knowledge of personal circumstances, perceptions of motivation in specific sessions) inform decisions. This could be achieved by ensuring truly anonymous marking in both benchmarking and double-marking. However, given the different ways educators engage with students, and the different extents to which they will know each of their

students – even down to details such as writing style, this may not always be practical and an appreciation for and awareness of these extraneous factors needs to be kept in mind during the moderation process.

Variation in standards between samples and resolution of discrepancies between assessors

Another issue to consider when evaluating moderation procedures is the fact that different subsamples of work evaluated by different assessors are likely to be of differing standards (Yorke, Bridges, & Woolf, 2000). Unless all assessors mark all submissions and have the same knowledge and expertise, some assessors may seem to be outliers when in fact the difference is a result of their sample containing genuinely low or high achieving students. This is not an issue for double-marking if the entire cohort is marked by the same assessors, but in reality it is often the case that different pairs (groups) of assessors will mark different subsets. Equally, if different pairs of assessors mark different subsets of students, the power-relation between the assessors introduces another form of potential bias. Less senior members of staff are more likely to adjust their marks toward those of more senior members (Orr, 2007), and this will vary across assessor pairs.

This may be overcome, particularly in double-marking, by keeping all marks anonymous or not sharing them between assessors before the moderation phase (Partington, 1994); though such an approach has its own issues. It has been suggested that blind double-marking may lead to a 'defensive' approach, with assessors opting for 'average' marks in order to minimise the chance of a discrepancy (Hornby, 2003). This is likely to unduly advantage low achievers and disadvantage high achievers. The extent

to which this affects benchmarking has not been studied empirically, but it is likely to depend on the group dynamics and the consequences of low inter-assessor agreement.

If discussion is used to reach a consensus rather than adopting mathematical approaches to resolution, this can help develop understanding of the criteria and their application (Price, 2005), leading to improved reliability in future assessments. Discussion of the reasoning behind each mark may be preferable to mathematical approaches as it can highlight material worthy of credit that had been missed by one or more assessors. It has been noted that some assessors like this opportunity to compare their marking and justifications with other assessors. Some, however, report feeling uneasy discussing justifications during moderation and subsequent exam boards (Hand & Clewes, 2000). This discussion element is not unique to double-marking, and these points also apply to benchmarking, yet the largest benefits to group understanding are likely to come from benchmarking as the criteria, marking and justifications are discussed in larger groups, thus helping reduce bias introduced by assessor options and interpretations. Any discussion between assessors will allow the development of the criteria, but the larger the group, the more diverse the views, and the wider the range of experience available to draw on in refining the criteria. In addition, comparison of marks and discussion of criteria in larger groups may go some way to avoiding the effects of power-relationships between pairs of assessors by providing a more varied group of individuals. This approach is not infallible, and the presence of particularly senior staff, or those with particular responsibilities for the assessment, may still have disproportionate influence; though this in turn may be minimised by the use of facilitators, anonymising submissions and comments, or discussion protocols.

With respect to students, a greater understanding of the moderation process can also be beneficial to learning. In the same way that discussions amongst assessors

clarify the assessment criteria by creating shared understanding, student knowledge of the moderation process improves the value of each assessment and the associated feedback. This is particularly apparent in peer-assessment that includes an element of moderation (Bloxham, Hughes, & Adie, 2016; Elwood & Klenowski, 2002).

Generalisability

One assumption in benchmarking is that because the sample marks have been agreed, then all other marks are likely to be comparable as well. Work by Harlen (e.g. 2014) in primary education supports this assumption to some extent, but the degree to which it holds in medical education has not been investigated. It seems likely, given the variability caused by criteria specificity and element-versus-holistic judgement methods that the extent to which the assumption of generalisability holds will be variable – even where combination methods are used to standard set. One such issue with generalisability has been raised with respect to the complexity of the ideas being assessed. When they are abstract concepts (e.g. professionalism or bedside-manner), subjective interpretation might be expected to lead to greater variability within assessors (Partington, 1994). This highlights the importance of discussing the justification of each mark, as the same marks may have been awarded but for different reasons (Rust, 2007). Double-marking of the entire group of students does not suffer from the same assumption, as every submission is awarded two independent marks and every discrepancy discussed. However, it only addresses the issue between particular pairs of assessors, and as with the issue of power-relations, many of the strengths of double-marking are undermined when different pairs of assessors mark different subsets of students; the procedure becomes comparable to benchmarking in this embodiment.

Training, alternatives, and burden

Cannings, Hawthorne, Hood, and Houston (2005) report moderate reliability from double-marking, but reiterate the claim made by others that most variability is due to factors other than assessor variability. They present a range of methods for evaluating marking, along with a flow-chart indicating when work should be double-marked. This is particularly useful when it is noted that training in the use of criteria provides very little improvement in reliability (Newble, Hoare, & Sheldrake, 1980). Selective application of double-marking based on the criteria of Cannings *et al* may help reduce the time and resource burden required to double-mark, but again makes the procedure comparable to benchmarking.

Bloxham (2009) draws a number of these points together, arguing that although moderation procedures improve perceptions of an assessment, objectively they add little as they still need to be applied within a particular setting and in relation to the socially constructed concept of what knowledge is being assessed. This view that moderation procedures increase burden but add little reliability or accuracy to the marks is a view held by many other authors (e.g. Cannings et al., 2005; Gibbs, 2006), and applies whether the procedure employed is double-marking or benchmarking (Bloxham, 2009). The key benefits of moderation are most likely the result of developing a consensus and understanding amongst assessors, and a pre-assessment discussion of an example or small sample of work is potentially more beneficial than multiple iterations of marking and re-marking during any moderation procedure (Smith, 2012). Such a discussion of criteria or exemplars before marking commences is part of benchmarking but may also be built in to double-marking approaches.

Bloxham (2009) also suggests that moderation across all assessments taking place within a module, more akin to benchmarking over a large sample, may provide a

better understanding of assessor reliability than individual level moderation, similar to moderation between pairs of assessors in double-marking. Such module-level moderation might take the form of statistical analysis of assessors, but this in itself may take many forms. Furthermore, in a benchmarking procedure, comparisons can more easily be made between assessors within each sample as subgroup variability is reduced by having all assessors mark the same sample of work. Criteria for suitable sample sizes, appropriate analysis, and methods of identifying low reliability or inconsistent assessors need careful consideration, and any evaluation across sample groups may be affected by genuine variability in the quality of submissions in each sample. Double-marking, unless implemented with the same assessor pair marking all submissions, makes overall analysis of assessor reliability difficult, and is potentially unnecessary if a single mark is awarded by agreement between the pair. This raises the issue of logistical challenges in relation to ensuring scripts are transferred between first and second assessors, and the added complication if marks are to be kept hidden during the process. If marks awarded by the first assessor are visible to the second, then the order in which the assessors mark the scripts should be balanced to avoid systematic bias.

Similar points have been made specifically in relation to Objective Structured Clinical Examinations (Newble *et al*., 1980). Although these typically show acceptable inter-assessor reliability, substantial improvements can be achieved by identifying and removing the most variable assessors. In addition, training on the assessment criteria was deemed unnecessary for consistent assessors, and found to have no effect on inconsistent assessors. Notable as the only empirical study, although investigating assessment of applied accounting skills, O'Connel *et al* (2015) compared the effects of pre-assessment marking workshops on assessor marks, and showed workshops significantly reduced the standard deviation of marks awarded to a given set of

submissions. This emphasises the need to consider how consistency and reliability should be assessed (Cronbach, Linn, Brennan, & Haertel, 1997), and how to account for differences in underlying ability between groups of students. It further highlights the potential benefits of changes elsewhere in the process of assessment rather than focussing solely on moderation of the marking process at the time of assessment (e.g. repeated sampling and assessment to provide more reliable assessment of student performance).

Alternative moderation procedures to consider are best thought of as different implementations of benchmarking and double-marking. Benchmarking procedures may vary the size of the sample, with larger samples giving a better estimate of overall reliability. They may also vary in the sample used, whether it consists of student submissions or exemplars created by staff. Double-marking may vary the number of pairs of assessors used, the size of each subset, and the methods used to identify and resolve discrepancies. These may range from taking the mean of the two marks to requiring a third, fourth or subsequent assessors until there is a majority consensus. An alternative approach is to single mark work following benchmarking, but subject failing (and excellent) submissions to double-marking (Cannings et al, 2005). There is, unfortunately, no empirical, work to the authors knowledge, on the effects of any of these implementations on reliability or perceived validity, but it seems reasonable to argue against 'mathematical' approaches on the grounds that the resultant 'average' grades may bear no resemblance to grade descriptors or assessment criteria associated with the average grades, and thus be of little help to the student in terms of feedback. This issue is related in many ways to the development and shared understanding of assessment criteria – they must be clear, applied, and understood in the same way by each assessor; as well as the issues related to group discussion of awarded marks.

Different assessors may give different grades for different reasons, and may overlook elements that others have based their judgements on. Discussion of the submissions, as opposed to a mere averaging of discrepant grades or scores, thus seems to have more pedagogic value to both assessors and students (Fry et al, 2003).

Student perceptions

From a student perspective, detailed qualitative feedback may improve the perceived validity of the assessment, but this may be difficult to implement in practice (van der Vleuten et al., 2012). Transparent communication of any feedback and moderation processes also improves student perceptions and staff confidence in dealing with complaints as the results are based on a wider consensus (Partington, 1994). There is often an element of psychometric analysis involved in judgements of reliability and validity, both in relation to assessment data and moderation data. These elements have often been criticised, but Schoenherr and Hamstra (2016) provide an insightful discussion of how these ostensibly statistical methods of evaluating validity and reliability have themselves developed out of holistic, qualitative discourse surrounding these important components of assessment rigour.

Peer benchmarking – allowing students to assess and discuss each other's work - by students could help align student expectations with their performance and increase satisfaction by allowing them to develop a better understanding of the moderation process, and discuss how and why a school employs the methods it does. The disconnect between perceived and actual quality of moderation has been cited as a possible cause for the persistence of poor ratings of assessment and feedback in UK national student surveys (http://www.hefce.ac.uk/lt/nss/) even after measures to improve moderation procedures (Gibbs, 2006). What has not been considered is that students are

not always made aware of changes to moderation processes, and rarely have the choices justified to them.

This also highlights the need to balance the cost of moderation procedures against their benefits. For example, lengthier moderation processes may make decisions more defensible, and marginally more reliable, but they are likely to delay feedback to students and reduce its usefulness (Gibbs & Simpson, 2004). In this respect benchmarking would seem to provide the most balanced approach, but only if a suitable sample is chosen for the purpose and the efficiency is valued more than the thoroughness of a complete double-marking procedure.

Moderation outside medical and higher education

Considering assessment outside medical and higher education, a review of assessment practices focussing on GCSEs and A-levels in England found that no exam boards use double-marking in its strict sense, favouring instead benchmarking approaches (Office of Qualifications and Examinations Regulations, 2014). The logistical and financial difficulties involved in double-marking prohibited its implementation, particularly when double-marking only offers marginal improvements on single-marking. No comparison was made to other moderation procedures, and the number of students involved in GCSE and A-level examinations is much larger than the numbers typically enrolled on individual modules or programmes in higher education.

Recommendations

The perceived advantages and disadvantages of the two types of moderation process are summarised in Table 1.

[Insert Table 1 about here]

For the gains made in reliability relative to the resources involved, benchmarking appears to be most efficient but this may rely on clear justification and communication to students who may otherwise feel that double-marking is the most appropriate means of moderation. To maximise the effectiveness of either method careful consideration should be given to how reliability, consistency, and outliers are defined and measured. In conclusion, after consideration of the literature and reflection on opinions and practices within our medical school, there are a number of recommendations which may be of use to other schools: 1. Ensure availability of sufficient staff with appropriate subject expertise in order to make discussions beneficial; 2. Provide appropriate staff training and facilitate agreement on criteria before marking, with example scripts covering the full range of expected marking (i.e. including top, middle and bottom); 3. Manage differences in staff seniority and expertise (in subject knowledge, assessment responsibility, and experience of being an assessor) by ensuring assessors are blinded to the identity of any other assessor who is allocated the same work to mark; and, 4. Be transparent with students about the moderation rationale and process to ensure perception of fairness. In summary, regardless of the approach taken to moderation, each stage of the entire assessment process should be carefully thought through, and this is more important than the type of procedure adopted.

References

Baume, D., and 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.

Baume, D., Yorke, M., and Coffey, M. 2004. What is happening when we assess, and how can we use our understanding of this to improve assessment? *Assessment and Evaluation in Higher Education*, 29 (4): 451-477.

Bloxham, S. 2009. Marking and moderation in the UK: false assumptions and wasted resources. *Assessment and Evaluation in Higher Education,* 34 (2): 209-220.

Bloxham, S., Boyd, P., and Orr, S. 2011. Mark my words: the role of assessment criteria in UK higher education grading practices. *Studies in Higher Education,* 36 (6): 655-670.

Bloxham, S., Hughes, C., and Adie, L. 2016. What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices, *Assessment and Evaluation in Higher Education,* 41(4): 638-653

Bloxham, S., & Price, M. 2013. External examining: fit for purpose? *Studies in Higher Education, 40*(2): 195-211.

Brown, G., Bull, J., and Pendlebury, M. 1997. *Assessing Student Learning in Higher Education*. London: Routledge

Cannings, R., Hawthorne, K., Hood, K., and Houston, H. 2005. Putting double marking to the test: a framework to assess if it is worth the trouble. *Medical Education,* 39 (3): 299-308.

Cronbach, L. J., Linn, R. L., Brennan, R. L., and Haertel, E. H. 1997. Generalisability analysis for performance assessments of student achievement or school effectiveness *Educational and Psychological Measurement,* 57 (3): 373-399.

Elton, L., and Johnston, B. 2002. *Assessment in universities: a critical review of research*. York: Higher Education Academy.

Elwood, J. and Klenowski, V. 2002. Creating communities of shared practice: the challenges of assessment use in learning and teaching. *Assessment and Evaluation in higher Education*. 27(3): 243-256.

Fry, H., Ketteridge, S., & Marshall, S. 2003. *A handbook for teaching and learning in higher education: Enhancing academic practice* (2nd ed.). London: Kogan Page.

Gibbs, G. 2006. Why assessment is changing. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education*. London: Routledge.

Gibbs, G., and Simpson, C. 2004. Conditions under which assessment supports student learning. *Learning and Teaching in Higher Education,* 1 (1): 3-31.

Hand, L., and Clewes, D. 2000. Marking the difference: an investigation of the criteria used for assessing undergraduate dissertations in a business school. *Assessment and Evaluation in Higher Education,* 25 (1): 5-21.

Harlen, W. (2014). *Assessments, Standards, and Quality of Learning in Primary Education*. York: Cambridge Primary Review Trust.

Hornby, W. 2003. Assessing using grade-related criteria: a single currency for universities? *Assessment and Evaluation in Higher Education,* 28 (4): 435-454.

Mitchell, K. and Anderson, J. 1986. Reliability of holistic scoring for the MCAT essay. Educational and Psychological Measurement. 46(3):771-775.

Newble, D. I., Hoare, J., and Sheldrake, P. F. 1980. The selection and training of examiners for clinical examinations. *Medical Education,* 15 (5): 345-349.

Nystrand, M., Cohen, A. S., and Dowling, N. M. 1993. Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment,* 1: 53-70.

O'Connell, B., De Lange, P., Freeman, M., Hancock, P., Abraham, A., Howieson, B., and Watty, K. 2015. Does calibration reduce variability in the assessment of accounting learning outcomes? *Assessment and Evaluation in Higher Education,* 41(3):311-349.

Office of Qualifications and Examinations Regulations. 2014. *Review of Double Marking Research*.

Orr, S. 2007. Assessment moderation: constructing the marks and constructing the students. *Assessment and Evaluation in Higher Education,* 32 (6): 645-656.
Partington, J. 1994. Double-marking students' work. *Assessment and Evaluation in Higher Education,* 19 (1): 57-60.
Pitts, J., Coles, C., and Thomas, P. 1999. Educational portfolios in the assessment of general practice trainers: reliability of assessors. *Medical Education,* 33 (7): 515-520.
Price, M. 2005. Assessment standards: the role of communities of practice and the scholarship of assessment. *Assessment and Evaluation in Higher Education,* 30 (3): 215-230.
Roberts, C., Shadbolt, N., Clark, T., and Simpson, P. 2014. The reliability and validity of a portfolio designed as a programmatic assessment of performance in an integrated clinical placement. *Medical Education,* 14 (197): http://dx.doi.org/10.1186/1472-6920-14-197.
Rust, C. 2007. Towards a scholarship of assessment. *Assessment and Evaluation in Higher Education,* 3 (2): 229-237.
Rust, C., O'Donovan, B., & Price, M. 2005. A social constructivist assessment process model: how the research literature shows us this could be best practice. *Assessment and Evaluation in Higher Education, 30*(3), 231-240.
Sadler, D. R. 2013. Assuring academic achievement standards: from moderation to calibration. *Assessment in education: Principles, Policy, and Practice,* 20 (1): 5-19.
Schoenherr, J.R. and Hamstra, S.J. 2016. Psychometrics and its discontents: an historical perspective on the discourse of the measurment tradition. Advances in Health Science Education. 21(3):719-729.
Smith, C. 2012. Why should we bother with assessment moderation? *Nurse Education Today,* 32 (6): e45-e48.
van der Vleuten, C., Schuwirth, L. W., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K., and Tartwijk, J. 2012. A model for programmatic assessment fit for purpose. *Medical Teacher,* 34 (3): 205-214.
Yorke, M., Bridges, P., and Woolf, H. 2000. Mark distributions and marking practices in UK higher education. *Active Learning in Higher Education,* 1 (1): 7-27.

Table 1. Compares a range of properties for benchmarking and double-marking. The extent to which each property is seen as advantageous or disadvantageous will vary depending on implementation, and how much utility or value is attributed to each property.

Benchmarking	
Advantages	Disadvantages
• Provides opportunity for detailed discussion of the sample and criteria among all assessors. • Allows assessors to compare their marks to a range of other assessors. • Allows assessors to learn from each other about marking criteria and justification of the marks awarded. • Enables effective assessment even if there is only one individual with expertise in a given area.	• Sample may not be representative of the cohort and assumes that agreement on the sample generalises to the cohort. • May decrease student perceptions of fairness, particularly if submissions cannot be effectively anonymised.

Double-Marking	
Advantages	Disadvantages
• Doubles the amount of feedback given to students. • Potentially reduces impact of assessor bias and increases perceptions of fairness if scripts cannot be effectively anonymised. • Assessors can be exposed to a wider range of student abilities.	• Doubles the number of submissions to be assessed and hence increases the burden on faculty time. • Potentially delays release of results. • Resolution methods may require additional assessors. • Limits discussion of criteria to the assessor pair. • Potential power bias within pairs of assessors. • Logistical considerations keeping marks hidden between assessors, or alternating the role of first assessor.