Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria

Sue Bloxham, Birgit den Outer, Jane Hudson, Margaret Price

Sue Bloxham[a], S.Bloxham@cumbria.ac.uk

Birgit den Outer[b], B.den-outer@brookes.ac.uk

Jane Hudson[b], Jane.hudson@brookes.ac.uk

Margaret Price[b], meprice@brookes.ac.uk

[a] Faculty of Education, Arts & Business, University of Cumbria, Lancaster, UK. (corresponding author)

[b] Department of Business and Management Oxford Brookes University, Oxford, UK

Abstract

Unreliability in marking is well documented yet we lack studies that have investigated assessors' detailed use of assessment criteria. This project used a form of Kelly's Repertory Grid method to examine the characteristics that 24 experienced, UK assessors notice in distinguishing between students' performance in four contrasting subject disciplines: that is their implicit assessment criteria. Variation in the choice, ranking and scoring of criteria was evident. Inspection of the individual construct scores in a sub-sample of academic historians revealed five factors in the use of criteria that contribute to marking inconsistency. The results imply that whilst more effective and social marking processes that encourage sharing of standards in institutions and disciplinary communities may help align standards, assessment decisions at this level are so complex, intuitive and tacit that variability is inevitable. It concludes that universities should be more honest with themselves and with students and actively help students to understand that application of assessment criteria is a complex judgement and there is rarely an incontestable interpretation of their meaning.

Key words: assessment criteria; marking; grading; standards

MAIN TEXT

Variation in grading

Higher education assessors' inconsistency and unreliability are well documented and the causes have been investigated by a range of theoretical and empirical studies. A key source of variation is attributed to assessors' different professional knowledge, experience, values and institutions (Read, Francis and Robson 2005; Smith and Coombe 2006). Wolf (1995) contends that markers acquire fixed habits in their marking which can influence their grading in ways that they may not be aware of.

Use of assessment criteria is also considered a potential cause of variability. In particular, assessors may not understand the outcomes they are supposed to be judging (Baume, Yorke and Coffey 2004), may not agree with, ignore or choose not to adopt the criteria (Baume, Yorke and Coffey 2004; Ecclestone 2001; Orrell 2008; Smith and Coombe 2006) or interpret them differently (Webster, Pepper and Jenkins 2000). The language of criteria or standards can mask differences in interpretation (Moss and Schutz 2001). Studies have found that experienced markers are no better than novice markers at applying standards consistently. This is partly because new markers pay greater attention to marking and marking guidance (Ecclestone 2001; Price 2005). In addition, assessors use personal criteria beyond or different to those stated (Baume, Yorke and Coffey 2004; Broad 2003; Greatorex 2000; Price 2005; Read, Francis and Robson 2005; Webster, Pepper and Jenkins 2000) including the use of implicit assessment criteria not shared explicitly with students or other assessors (Hunter and Docherty 2011; Shay 2005).

A further source of variation related to this last point is that assessors attach importance to different qualities or aspects of student work (O'Hagan and Wigglesworth 2014; Read, Francis and Robson, 2005; Smith and Coombe 2006). A final reason emerging in the literature is that assessors have different expectations of standards at the different grade levels (Grainger, Purnell and Zipf 2008; Hand and Clewes 2000). Therefore, whilst markers may be working with shared criteria, they may not agree on 'how well the various criteria have been achieved' (Grainger, Purnell and Zipf 2008: 134). Overall, this body of research suggests that even where assessors agree marks (which, in the authors' experience, lecturers often claim to), this may not necessarily be for the same reasons.

The last few decades have witnessed considerable efforts to improve fairness and transparency in marking; securing appropriate standards and consistency in grading by using the tools of explicit criteria and rubrics (marking schemes) despite the evidence that use and interpretation of these tools may be an important source of grading inconsistency. However, we lack studies that have specifically investigated assessors' use of published and personal criteria. The project reported here sought to investigate the root of similarities and differences in the standards used by experienced assessors by exploring the nature of the criteria that assessors use as well as the commonality in both the selection of, and meaning given to, criteria by different assessors. It also sought to understand the relationship between inconsistencies at the level of individual constructs and assessors' overall judgement of quality. It asks whether an understanding of the roots of variation in judgement helps identify potential strategies for reducing variation and if this is possible.

Method

The research used a form of Kelly's Repertory Grid method (Fransella, Bannister and Bell, 2003) to collect data about the grading judgements of experienced assessors and to provide more robust data than was considered likely to arise solely from reported behaviour. This method facilitated the assessors in articulating the nuanced constructs they use in distinguishing between pieces of student work. Twenty-four experienced assessors in four contrasting disciplines, psychology, nursing, chemistry and history, were recruited from twenty diverse UK universities through open advertisement. They had sufficient experience within their discipline to be appointed as a reviewer of assessment standards at a minimum of one other university (known as an external examiner in the UK), although many participants had been involved in several external examining appointments that took place over a number of years. Assessors were provided with five examples of student work and relevant assessment criteria where available. Borderline work was selected to help tease out the nuanced deciding factors in judgement although the assessors were not given this information. Contextual information, such as year and place of study, previous marks for the work and credit weighting, was also not provided.

A week before the Kelly's Repertory Grid exercise, researchers asked the assessors to read five assignments addressing the same task, which was typical for their discipline. During the activity, researchers presented assessors with different combinations of three out of the five assignments and asked them to identify how two were the same but differed from the third. They were asked to name these differences, for example two might have 'clear structure' whilst the other was 'confused'. The first of these qualities, the similarity, was placed on the left-hand side of a grid and the other, the difference, was placed on the right hand side. In this way, the 'constructs' by which the assessors discriminated between examples of

student work were elicited. This process was repeated until all possible combinations of three were exhausted, or until time ran out, with each trio of assignments creating another line on the grid. These constructs are the characteristics that assessors noticed in distinguishing between student work and therefore presumed by the researchers to be a verbal representation of their implicit assessment criteria. Assessors then ranked each assignment against these self-generated constructs and, finally, provided an overall grade for each piece. As the grading was not an exacting exercise and assessors did not have access to wider contextual information, no validity was accorded to the overall marks given except in observing the spread of marks and the assessors' perceptions of the relative worth of the five assignments. In other words, overall grades were only used to see the extent to which the assessors judged work to be of the same or different standards.

Data analysis attended to the range of constructs, ranking of constructs by importance, shared constructs across a discipline, consistency of scoring within each construct, and consistency of overall judgement for each piece. A classification of 'surface' and 'global' was used, with global constructs referring to disciplinary knowledge and academic qualities, such as depth of knowledge, analysis, and argumentation, and surface constructs referring to more generic and technical qualities, such as grammar, citation, presentation and register.

Findings

Consistency in overall judgement.

Consistency between the assessors' overall judgement, as evidenced by how they graded the assignments, reflects other studies of reliability in marking in revealing little inter-assessor agreement. Only one of the twenty pieces was assigned the same rank by all six

assessors in any of the disciplines. All other assignments were given grades that 'ranked' them against the other assignments in at least three different positions (i.e. best, second best, and so on). Nine of the twenty assignments were ranked both best and worst by different assessors. See Table 1, which presents the variable ranking of the different assignments. Differences in ranking might be expected for work selected for its borderline nature where judgement of difference is likely to be nuanced. However, the assessors did not see the work as borderline in the main, with only 7 of the 21 assessors placing them all within two adjacent grade bands.

(Insert Table 1 about here)

Assessors' use of constructs

The Kelly's Repertory Grid exercise elicited thirty seven constructs with the method artificially limiting the maximum per examiner to ten. The mean number of constructs per assessor was 7.4 although this masks a range of 3 to 10 between the different assessors. The mean and median for each subject discipline were roughly similar. Within the 37 constructs, 4 were surface constructs. The remainder were global constructs. There was evidence of some sharing of constructs within disciplines although this is was not at all universal.

The psychologists generated 18 different constructs, 7 of which were in the criteria provided. In relation to global constructs, u*se of evidence* and *argument* appeared in 4/6 assessors' grids as did *referencing* and *academic style* in the surface constructs. The nurse tutors elicited 15 constructs of which only 5 were included in the criteria. It should be noted that these criteria were framed more in terms of learning outcomes to be demonstrated, for example the student should have 'Discussed the centrality of holism in the care of patients'.

In practice, the constructs elicited tended not to focus on these outcomes but on the way in which they had been met, for example the depth of knowledge displayed or the quality of analysis. This may explain why there was a lack of consistency between the nurses on the constructs elicited with many being generated by only one or two assessors.

Only two constructs were generated by at least four of the chemistry assessors. It should be noted that they had not been provided with assessment criteria or a model answer for their exam scripts and this may have reduced the likelihood of consistency. They generated 16 constructs, one of which (*Quality of explanation)* was elicited from all six assessors. All but two constructs were generated by fewer than four assessors with nine elicited from only one assessor. They were not particularly concerned with surface constructs.

Assessors working in History generated 18 constructs and 7 of these related to the assessment criteria. There was some consistency with 6 constructs used by at least 4 assessors. All assessors used *historiography and* 5 out of 6 used *structure* and *academic style*. Ten constructs were elicited from only one assessor. Whereas the constructs shared by the psychologists tended to be surface characteristics, the historians shared more global constructs. A lack of emphasis on surface characteristics may be a feature of the good quality of these characteristics in the essays provided as this was remarked upon by three assessors. See table 2 for a summary of assessors' use of constructs.

Insert table 2 about here

Ranking of importance of constructs within disciplines

On completion of the exercise in generating constructs, each assessor was asked to rank their constructs in order of importance. Individual assessors generated different numbers of constructs and therefore a construct ranked at 5 might be relatively unimportant to a respondent who only generated 5 constructs. On the other hand, ranking a construct as 5th out of a list of 10 probably indicates a greater significance to the assessor. The data is based on the raw rank numbers rather than figures adjusted for the individual's number of constructs. Overall, in all subjects, there was some consistency in ranking constructs but only at the very broad level that 'surface' constructs were consistently ranked lower than global characteristics. See Table 3 for the relative ranking of global and surface constructs by discipline. There were no other clear agreements over the ranking of individual constructs within any subject area.

(insert Table 3 about here)

There were specific examples of assessors ranking apparently similar constructs very differently and this may point up the potential weakness in reported rankings and their interpretation. For example, one assessor generated two similar constructs, '*overall adherence to assignment brief' and* '*following of specific assignment requirements*' which appear to be describing the same criteria and were coded similarly. Yet the assessor ranked them as first and fourth in order of importance. Another assessor gave a rank of 1 for *quality of writing* and a rank of 8 for *reasonably strong writing skills* although these constructs appear very similar. These unexpectedly different rankings for similar constructs may be

further evidence of the difficulty in verbally representing these constructs to adequately reflect the nuanced differences in meanings.

Consistency of scoring individual constructs

As part of the Kelly's Repertory Grid exercise, for each construct generated, assessors were asked to score all five assignments on a count from 1 to 5 depending on how well it matched the construct identified. For example, if an assignment was close to the construct identified on the left hand side of the grid (similarity) it would be given a 1 and if it was close to the opposite construct on the right hand side of the grid, it would be awarded a 5. The assessor could use other numbers on the scale 1-5 for pieces which were more or less similar to either pole of the construct and, indeed, some used numbers beyond that range to stress work that was stronger or weaker on a specific construct than the scripts in the trio from which that construct was generated.

This part of the exercise enables us to see the extent to which assessors judge work to be of a similar standard in relation to a 'named' quality. To a certain extent, this comparative process was made difficult by the lack of shared constructs between assessors. However, there were 17 constructs which were used by at least four assessors within a subject discipline and these have been used for analysis.

A review of the scores for these 17 constructs shows that, out of a potential 85 opportunities, there are only 2 examples where all the assessors within a discipline awarded the same score for the same construct. There are 9 incidences where all assessors came within two scores and 42 instances (approximately half) where assessors rated the 5 different essays from 1 to 5 for the same construct. Consequently, this data suggests that

examining differences at the construct/ criteria level, might help us to understand inconsistencies in assessors' overall grading judgements. The following close analysis of the historians is designed to investigate these inconsistences.

History

It is reasonable to argue that constructs can only be analysed at a disciplinary level because although terms such as structure and argument may be used across subjects, they are likely to hold different meanings depending on the discipline. We selected history because the assessors demonstrated the greatest commonality in the constructs used (6 used by at least four assessors) as well as displaying the general features found in all disciplines such as variability across overall judgement, the constructs used, ranking of constructs and construct scores. The greater commonality may reflect the range of epistemological and scientific differences in history compared with other subjects or the assignments used.

Overall consistency in judgement

The researchers were hesitant to use the overall marks assigned for the reason discussed above so they are only used here to illustrate the range of grades and how the assignments were compared with one another by the assessors. The specific grades awarded should not be interpreted any more significantly than that. Tables 4 and 5 present the historians' overall assessment of the students' work and reveal considerable variation in judgement about the relative worth of the assignments although they were supplied with a set of standard history essay marking criteria. There is a broad span of grades from a 1st (A – top pass) to a 3rd (D - low pass) despite the assignments all being graded as borderline 2.1.- 2.2 (B - C) when they were originally marked. This replicates Read, Francis and Robson (2005)

who found that two undergraduate history essays received 6 different degree classifications when marked by 50 assessors. Individual assignments vary in the amount of difference between assessors' judgement. There is fairly strong agreement over the best assignment but much more mixed reviews of the other four although C and D are generally judged to be weaker.

Variation in overall judgement does not appear to be a feature of 'harder' or 'softer' markers but may be a feature of markers who were more or less prepared to use the extremities of the scale; that is firsts, thirds and fails (see Table 4, final column). This reinforces the argument that staff vary considerably both in the marks they give (Yorke, 2008) and in the shape of their mark distributions (Heywood, 2000) and provides some support for moderation of these distributions.

(Insert tables 4 & 5 about here)

Interestingly, overall agreement on a mark by assessors appears to mask variability in individual criteria. Therefore whilst essay E has many high appraisals, it also comes in for some poor evaluation in specific constructs by different assessors: for example, *loose structure* (1), *unawareness of historiography* (3), *fairly poor academic style and general presentation* (4), *not enough use of historiography* (5), *and does not use primary sources* (6). However, in the case of this assignment, (and unusually in our findings) assessors award strong scores for most constructs and overall ranking is both high and consistent. This tends to reinforce arguments about the holistic nature of judgement (Sadler 2009) where assessors will balance different aspects of a piece of work.

The difference in the historians' appraisal of individual constructs was further investigated and five potential reasons were identified that link judgement about specific qualities in assignments to potential variation in grading.

Reason 1: Using different criteria to those published

It is important to note that the Kelly's Repertory Grid method invited the assessors to articulate what they noticed in the student work rather than apply a set of criteria to the essays. They generated 18 constructs and 7 of these related to the assessment criteria. The overall list of constructs generated by the historians is:

Difficult title/ question attempted

Good attempts to define constructs

Attempts to set up essay with introductory paragraph

Understanding of wider context

Quality of explanation

English/ grammar/ proof reading

Referencing/ citation

Analysis/ critical analysis

Addresses the question

Structure/ organisation

Good conclusion

Style/ Academic style/ register

Presentation/ legibility

Historiography

Wide reading,

Depth/ quality of Knowledge

Developing argument

Use of theory

A glance at this list suggests that none of these constructs would be misplaced in the assessment criteria of an undergraduate history assignment. In practice, lists of assessment criteria are normally shorter; this does not mean, however, that assessors do not draw on the wider list, whether consciously or unconsciously (Sadler 2009). These results support arguments suggesting that additional criteria are used above those that are explicitly stated. A key difficulty with this apparent practice is the potential for assessors to vary in the additional criteria they use and the likelihood that they may not be conscious of all the criteria they use in making judgements.

Reason 2: Assessors have different understanding of shared criteria

A second reason for the link between variation in assessors' grading and their judgement about specific qualities in assignments is that assessors hold a different concept of what a criterion means. Table 6 illustrates this variability by presenting the scores for *engagement with historiography,* a construct which was not in the stated criteria but which all historians articulated.

(*Insert Table 6 about here)*

The wording of the actual constructs elicited from the 6 historians is:

1. *Engages well with historiography > Hardly engages with historiography (reversed)*

2. *Historiographically determined > Less determined by historiography*

3. *Engagement with historiography > Unawareness of historiography*

4. *Awareness of historical debate, historiography > Absence of debate*

5. *Clear investigation of previous arguments in the area > Not enough use of historiography*

6. *Engages with historiography > Doesn't explicitly discuss the historiography.*

It was judged that all these constructs referred to the quality of historiography yet the judgements vary considerably. Assessors 1 and 3 use very similar language but have variable scores particularly with essays B, D, and E. For example, assessor 1 considers essay E to engage well with historiography whereas assessor 3 considers that it shows unawareness of historiography. Likewise, assessor 3 reports that essay D engages with historiography but assessors 1, 2, 4 and 6 report limited or no engagement. Can these assessors be sharing a view of what engagement with historiography means when made concrete in actual student assignments?

Similar findings are presented in Table 7 where the assessors' scores for the construct of *developing argument* are displayed. The actual language of the constructs by the four historians were:

1. *Argument excellent > argument adequate*

3. *Argument focus > narrative focus (reversed)*

4. *Reasonable argument > superficial argument*

5. *Clear exposition of argument > contradiction of argument.*

The first three assessors agree in their scores although the actual constructs used vary somewhat. This is discussed below. The fourth assessor's judgements of the texts suggest that she conceives of argument quite differently; although s/he uses the language *clear exposition of argument* as the positive end of the construct, the scores are so out of alignment with the other assessors that it is hard to conclude that they accord the same meaning to the term.

(insert Table 7 about here)

A clear conclusion, if we accept the assumption that the constructs represent assessors implicit criteria, is that although they appear to use similar criteria, in practice they interpret such criteria differently and this has the potential to contribute to differences in standards.

Reason 3. Assessors have a different sense of appropriate standards

A further reason for variation indicated by the data is that although assessors may agree on what a construct, for example *developing argument,* means, they may have a different sense of what constitutes *excellent, adequate,* and *weak* in relation to *argument*. This indicates an issue of standards rather than interpretation.

This point is illustrated by the assessors' judgement of the quality of *argument* (see Table 7 and accompanying list of constructs). With the exception of assessor 5, there seems to be a

strong agreement between the assessors on the stronger and weaker assignments in relation to *argument*. But whereas assessor 1 sees the better essays as demonstrating *excellent argument* assessor 4 describes it merely as *reasonable argument* and assessor 3 as having an *argument focus*. Likewise, with the weaker appraisals, the language of the constructs suggests that standards for the same assignment differ from *superficial* to *adequate argument.* Might it be that these three assessors do hold a shared meaning of *argument* but diverse understandings of what constitutes *excellent, reasonable, adequate* and *superficial?* If so, this variation in standards might explain some variety in their overall judgement. This finding replicates Grainger, Purnell and Zipf (2008) and Hand and Clewes (2000) who found that assessors have different expectations of the standards required at various levels. This variation can lead to disagreement over grades.

Reason 4. Criteria contain diverse sub-criteria

The construct *structure and organisation* illustrates the potential for complexity within any assessment criterion even though they may appear to be simply stated in a list of assessment criteria. This construct shows the same pattern of agreement and disagreement, some of which might be explained by the reasons already discussed. Another reason may be that the erratic similarities and differences between assessors across the 5 assignments is a result of the complexity of criteria. The constructs used were:

1. *Keeps a logical and analytical structure all the way through > loose structure*
2. (a) *Thematically and analytically structured > Narrative dominated by chronological approach*

(b) Balanced in level of attention to all structural components > imbalanced in level of attention to all structural components

3. *Effective structure > weak structure*

4. *Extremely well-structured > not so well structured*

6. Clear structure and signposting > jumps in with no signposting

Assessors 4 and 6 tend to agree on essays C, D and E but have much less agreement on assignments A and B (see Table 8). Is that because *signposting* (from 6's construct, but which does not feature in 4's construct) is more of an issue in essays A and B?

(*Insert Table 8 about here)*

Perhaps different ways of talking about structure give us some insight into the variation in lecturers' understanding of the concept. For our assessors it seems to include, for example, *logical structure, analytical structure and thematic structure.* This complexity is hardly signalled in the published criterion for this essay, which was 'the quality of structure and focus'. In these assessors' constructs, there is a suggestion of a sub-set of criteria under the umbrella of *structure* and therefore it is hardly a surprise that we find different appraisals for each essay. This complexity is not so evident in all of the constructs, some of which tend to be more qualitative judgements of the same term; for example, *excellent argument, adequate argument, superficial argument,* and *reasonable argument.*

Reason 5. Assessors value and weigh criteria differently in their judgements

Each assessor was asked to rank their constructs in order of importance for judging student work. In common with other disciplines, the historians tended to give lower rankings for surface characteristics compared with global constructs. Overall, there was considerable difference in the ranking of shared constructs by the different assessors.

Therefore a final reason why assessors' use of criteria to grade work may lead to variation is in the value and weighting they assign to these. For example, if *structure* is of paramount importance to one assessor (2) and second in importance to another (4) but only 5th out of 7 (1) and 5th out of 10 (3) for others, it is possible to see how the balancing or weighting of different components which consciously or unconsciously forms part of the judgement process is likely to produce variable results.

It is important before moving to a discussion of the findings to note that the choice of method may have influenced the results. Kelly's Repertory Grid required assessors to generate their own descriptions of constructs and we had to make a judgement about the commonality between constructs in order to compare scores. Amongst this coding, there will be cases where apparently shared descriptions did not reflect shared meaning. In such cases it is not surprising that there was variation in scores. Nevertheless, even if this was a shortcoming of our method, it also signals the limitations of simply worded assessment criteria in capturing all the nuances of assessor meaning. Further studies of assessors' understanding of frequently used terms might help clarify the extent of difference.

The interviews provided an opportunity for assessors to give feedback on the Kelly's Repertory Grid method and five described it as difficult to do. In particular, assessors commented on issues of language (its nature, use and interpretation) and the lack of contextual information (level, nature of teaching, knowledge of students). In relation to ranking the importance of constructs, assessors suggested that this might change depending on aspects of context such as the students' level of study. Others rejected a criterion-by-criterion approach to marking. They talked about having to weigh up different factors and the fact that formulaic combinations of criteria often led to the 'wrong' mark. Others expressed no difficulty or felt the resulting list was appropriate. Consequently, the ranking data set out above can probably only be considered to provide a broad picture of assessors' views on the importance of different constructs and further research is needed.

Discussion

The findings set out above replicate the message of other studies in demonstrating considerable variation in assessors' grading where complex higher education tasks are involved. It is important to stress that we perceive this inconsistency to be a reflection of the complex and intuitive nature of judgement at the higher education level, and should not be interpreted as criticism of the assessors. Nevertheless, it is not surprising that some researchers and students claim that grades are more dependent on who marks an essay than its content (O'Hagan and Wigglesworth 2014) with the potential for a sense of unfairness and dissatisfaction in students.

It seems that some assignments increase the diversity of judgements whereas others elicit greater agreement. Efforts to improve consistency have frequently focused on detailing

assessment criteria or rubrics and therefore this study specifically examined experienced assessors' use of criteria in differentiating student performance in order to investigate sources of variable judgement. Variation in the choice, ranking and scoring of criteria was evident across the sample and inspection of the individual construct scores in a sub-sample of academic historians suggested five possible factors that may cause inconsistency. It is probable that these factors combine in some way in the grading behaviour of any individual. These findings indicate that criteria are likely to have limited power in achieving consistent judgement. Shared language is insufficient to ensure shared interpretation of common criteria. In addition, we cannot be confident that only published criteria will be drawn upon for judgement or that they will be weighted similarly by markers.

This study further clarifies the notion of personalised 'standards frameworks' (Bloxham, Boyd and Orr 2011) which are conceived of as the lens through which individuals read student work. 'Standards' frameworks' are dynamic; they are constructed and reconstructed through involvement in communities and practices including engagement with student work, moderation and external examiners' feedback (Sadler 1989; Crisp 2008; Bloxham, Boyd and Orr 2011). Similar notions have been theorised such as 'teachers conceptions of quality' (Sadler 1989, 127), assessors' 'evaluative frameworks' (Broad 2003), 'assessors' interpretive frameworks' (Shay2005, 665), 'pre-formed knowledge structures' (Crisp 2008, 250) and 'ways of understanding the world' (Read, Francis and Robson 2005, 242). The research reported here provides a greater insight into the contributing elements of 'standards frameworks' in terms of choice of criteria (whether intuitive or deliberate), understanding of them, variation in standards for individual criteria and value and weighting accorded to them.

There is a view that individuals' 'standards frameworks' are heavily influenced, but not determined, by subject discipline norms (Shay 2005) which ensures a level of agreement over standards. Indeed, the interview material generated as a second part of this study (Bloxham, den Outer, Hudson and Price, forthcoming) indicated that these experienced assessors have a personal commitment to their discipline and, because they have studied at a high level in their fields, consider that they all think the same way about standards. However, these results contradict that view, suggesting that interpretation and understanding of these norms varies significantly in practice.

Recommendations

What are the implications of this work for better securing consistent standards amongst markers? A natural response is that we need to detail assessment criteria more fully to improve shared understanding, creating rubrics which stipulate the required achievement for each grade in each criterion. However, it is unlikely that we can capture the wide range of criteria used, the nuance in interpretation, the complexity of individual criteria and the determination of standards in a usable form. Such detail is likely to make marking an overly onerous process, limit independent thought and originality in students and encourage middling grades if individual criteria are scored. It may prevent markers awarding good marks to unexpected responses (Herbert, Joyce and Hassall 2014). The general use of holistic marking practices in higher education (Bloxham, Boyd and Orr 2011; Sadler 2009) suggests that such an approach will be undermined by staff working backwards from a holistic judgement to determine commensurate marks for individual criteria as found in Grainger, Purnell and Zipf (2008). More importantly, this type of technical enhancement has been dismissed by other researchers (Price 2005; Swann and Ecclestone 1999) who suggest

that marking grids and assessment criteria are insufficient on their own because application of a marking scheme to a specific assignment is a 'social construct' negotiated between the members of that assessment community and influenced by their tacit knowledge (Baird *et al.* 2004).

The latter points reinforce the view that we need fresh thinking about reliability, fairness and standards in higher education assessment and that our current reliance on criteria, rubrics, moderation and standardising grade distributions is unlikely to tackle the proven lack of grading consensus. One way forward worth considerably more investigation is the use of community processes aimed at developing shared understanding of assessment standards. We need to understand what these community processes might look like, how they can be evidenced, how much is needed to gain adequate consistency and how they can be made sustainable. Experiments are taking place to engage university teachers in the type of activity and dialogue that can bring about consensus and alignment of standards with learning outcomes (Watty et al. 2014) but we need further evidence on the impact and sustainability of such practices.

The real challenge emerging from this paper is that, even with more effective community processes, assessment decisions are so complex, intuitive and tacit that variability is inevitable. Short of turning our assessment methods into standardised tests, we have to live with a large element of unreliability and a recognition that grading is judgement and not measurement (Yorke 2011). Such a future is likely to continue the frustration and dissatisfaction for students which is reflected in satisfaction surveys. Universities need to be more honest with themselves and with students and help them to understand that application of assessment criteria is a complex judgement and there is rarely an

incontestable interpretation of their meaning. Universities need to be more honest in helping students to understand that application of assessment criteria is a complex judgement and there is rarely an incontestable interpretation of their meaning. Indeed, there is some evidence that students who have developed a more complex view of knowledge see criteria as guidance rather than prescription and are less dissatisfied (Bell, Mladenovic and Price 2013).

Accepting the inevitability of grading variation means that we should review whether current efforts to moderate are addressing the sources of variation. This study does add some support to the comparison of grade distributions across markers to tackle differences in the range of marks awarded. However, the real issue is not about artificial manipulation of marks without reference to evidence. It is more that we should recognise the impossibility of a 'right' mark in the case of complex assignments and avoid over-extensive, detailed, internal or external moderation. Perhaps a better approach is to recognise that a profile made up of multiple assessors' judgements is a more accurate, and therefore fairer, way to determine the final degree outcome for an individual. Such a profile can identify the consistent patterns in students' work and provide a fair representation of their performance without disingenuously claiming that every single mark is 'right'. It would significantly reduce the staff resource devoted to internal and external moderation reserving detailed, dialogic moderation for the borderline cases where it has the power to make a difference. This is not to gainsay the importance of moderation which is aimed at developing shared disciplinary norms as opposed to superficial procedures or the mechanical resolution of marks.

Sponsorship We are grateful to the UK Quality Assurance Agency and the UK Higher Education Academy for their joint sponsorship of this research.

References

Baird, J., J. Greatorex and J.F. Bell. 2004. "What makes marking reliable? Experiments with UK examinations." *Assessment in Education: Principles, Policy & Practice* 11 (3): 331-348. Doi:10.1080/0969594042000304627

Baume, D., M. Yorke, and M. Coffey. 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. Doi: 10.1080/02602930310001689037

Bell, A., R. Mladenovic, and M. Price. 2013. "Students' perceptions of the usefulness of marking guides, grade descriptors and annotated exemplars." *Assessment & Evaluation in Higher Education* 38 (7): 769-788. Doi:10.1080/02602938.2012.714738

Bloxham, S., P. Boyd, and S. Orr. 2011. "Mark my words: the role of assessment criteria in UK higher education grading practices." *Studies in Higher Education* 36 (6): 655-670. DOI:10.1080/03075071003777716

Bloxham, S., B. den Outer, J. Hudson, and M. Price. Forthcoming. "External peer review of assessment: an effective approach to verifying standards?" *Higher Education Research and Development.*

Broad, B., 2003. *What We Really Value: Beyond rubrics in teaching and assessing writing.* Logan, Utah, Utah State University Press.

Crisp, V., 2008. “Exploring the nature of assessor thinking during the process of examination marking.” *Cambridge Journal of Education* 38(2): 247–64. Doi 10.1080/03057640802063486.

Ecclestone, K. 2001. “I know a 2:1 when I see it: understanding criteria for degree classifications in franchised university programmes.” *Journal of Further and Higher Education* 25 (3): 301-313. 10.1080/03098770126527

Fransella, F., D. Bannister and R. Bell (2003) *A Manual for Repertory Grid Technique* (2nd Edition). Chichester: Wiley

Grainger, P., K. Purnell, and R. Zipf. 2008. “Judging quality through substantive conversations between markers.” *Assessment and Evaluation in Higher Education* 33(2): 133–42. Doi: 10.1080/02602930601125681

Greatorex, J., 2000, ‘Is the glass half full or half empty? What examiners really think of candidates' achievement’, paper presented at the *British Educational Research Association Conference,* 7–10 September 2000, Cardiff.

Hand, L. and D. Clewes. 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. DOI:10.1080/713611416

Herbert, I.P., J. Joyce, and T. Hassall. 2014. “Assessment in Higher Education: The Potential for a Community of Practice to Improve Inter-marker Reliability.” *Accounting Education*, DOI: 10.1080/09639284.2014.974195

Heywood, J. 2000. *Assessment in Higher Education.* London: Jessica Kingsley.

Hunter, K. and P. Docherty. 2011, “Reducing variation in the assessment of student writing.” *Assessment and Evaluation in Higher Education* 36(1): 109–24. **DOI:** 10.1080/02602930903215842

Moss, P.A. and A. Schutz. 2001. “Educational Standards, Assessment and the search for consensus. *American Educational Research Journal.”* 38 (1): 37-70. doi:10.3102/00028312038001037

O’Hagan, S.R and G. Wigglesworth. 2014. “Who's marking my essay? The assessment of non-native-speaker and native-speaker undergraduate essays in an Australian higher education context.” *Studies in Higher Education*, DOI: 10.1080/03075079.2014.896890

Orrell, J. 2008, “Assessment beyond belief: The cognitive process of grading” in *Balancing dilemmas in assessment and learning in contemporary education*, edited by A. Havnes and L. McDowell, 251-263. London: Routledge.

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. **DOI:**10.1080/02602930500063793

Read, B., B. Francis, and J. Robson. 2005 “Gender, bias, assessment and feedback: analyzing the written assessment of undergraduate history essays.” *Assessment and Evaluation in Higher Education* 30 (3):241-260. **DOI:**10.1080/02602930500063827

Sadler, D. R., 1989 “Formative assessment and the design of instructional systems.” *Instructional Science* 18(2): 119–44.

Sadler, D.R. 2009. “Indeterminancy in the use of preset criteria for assessment and grading.” *Assessment & Evaluation in Higher Education* 34 (2): 159-179. **DOI:**10.1080/02602930801956059

Shay, S. 2005. “The assessment of complex tasks: a double reading.” *Studies in Higher Education* 30(6): 663–79. DOI:10.1080/03075070500339988

Smith, E. and K. Coombe. 2006. “Quality and qualms in the marking of university assignments by sessional staff: an exploratory study.” *Higher Education* 51 (1):45-69. Doi:10.1007/s10734-004-6376-7

Swann, J. and K Ecclestone. 1999. “Litigation and learning: tensions in improving university lecturers’ assessment practice.” *Assessment in Education,* 6 (3): 357-375. Doi: 10.1080/09695949992801

Watty, K., M. Freeman, B. Howieson, P. Hancock, B. O’Connell, P. de Lange, and A. Abraham. 2013. “Social moderation, assessment and assuring standards for accounting graduates.” *Assessment & Evaluation in Higher Education,* DOI: 10.1080/02602938.2013.848336

Webster, F., D. Pepper, and A. Jenkins. 2000. “Assessing the Undergraduate Dissertation.” Assessment and Evaluation in Higher Education 25 (1): 71–80. DOI:10.1080/02602930050025042

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

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

Yorke, M., 2011. “Summative assessment: dealing with the ‘measurement fallacy.” *Studies in Higher Education* 36(3): 251–73. DOI:10.1080/03075070903545082

Table 1 Range of assessors' ranking of assignments

	Assignments				
	A	**B**	**C**	**D**	**E**
Psychology	3rd-5th	1st – joint 2nd/3rd	1st-5th	1st-5th	1st – joint 4th/5th
Nursing	1st- joint 3rd/4th	1st-5th	Joint 1st/2nd – 5th	Joint 1st/2nd – 4th	1st – joint 3rd/4th
chemistry	1st-5th	Joint 1st/ 2nd – Joint 4th/5th	Joint 1st / 2nd – 5th	1st – 3rd	1st-5th
History	Joint 1st / 2nd – 3rd	Joint 1st / 2nd – 4th	Joint 2nd/ 3rd – 5th	Joint 2nd/ 3rd – 5th	1st – Joint 1st/2nd

Table 2: Number and consistency of constructs generated

	Psychology	Nursing	Chemistry	History
Number of constructs generated	18	15	16	18
Number of global constructs shared by at least 4 assessors	2	2	1	5
Number of surface constructs shared by at least 4 assessors	2	2	1	1
Number of constructs aligned with assessment criteria provided	7	5	Not applicable	7
Number of constructs used by only 1 assessor	9	5	9	10

Table 3: Ranking of global and surface constructs by subject (1 = high rank)

	Psychology		Nursing		Chemistry		History	
	Median	**Mean**	**Median**	**Mean**	**Median**	**Mean**	**Median**	**Mean**
Global constructs	3	2.9	3	3.09	3	3.9	3.5	3.7
Surface constructs	5	5.1	6	5.6	7	6.5	7	7.25

Table 4: Overall grade by history assessors

Assessor	Essay A	Essay B	Essay C	Essay D	Essay E	Range of grades for each assessor
1	1st	1st	Low 2.1	2.2/ 2.1	1st	1st-2.2/2.1 (A-B/C)
2	2.1	2.2	3	2.1	1st	1st-3rd (A-D)
3	Low 2.i	Mid 2:2	Low 2.2	Low 2.2	Mid 2.1	Mid 2.1-Low2.2. (B-C)
4	Mid 2:2	Mid2:1	2:2-3rd	3rd	1st	1st-3rd (A-D)
5	2.2	2.1.	2.1	2.1	2.1	2.1.-2.2 (B-C)
6	2.1	2.1	2.2	3rd	1st	1st-3rd (A-D)
Range of marks for each essay	1st – 2.2 (A-C)	1st – 2.2 (A-C)	2.1. – 3rd (B-D)	2.1.–3rd (B-D)	1st – 2.1 (A-B)	

Table 5: Range of assignment rankings by history assessors. (J1/2 = joint 1st/2nd)

	Assignments				
Assessor	A	B	C	D	E
1	J1/2	J1/2	4	5	J1/2
2	J2/3	4	5	J2/3	1
3	2	3	J4/5	J4/5	1
4	3	2	4	5	1
5	4	J2/3	J2/3	5	1
6	J2/3	J2/3	4	5	1
Range of rank	J1/2-3rd	J1/2-4th	J2/3-5th	J2/3 - 5th	J1-J1/2

Table 6: Assessors scores by assignment for *Engagement with historiography*

	Construct: engagement with historiography				
Assessors	Essay A	Essay B	Essay C	Essay D	Essay E
1	2	1	3	5	1
2	4	4	2	5	1
3	3	4	1	1	5
4	4	2	5	5	1
5	3	3	1	3	5
6	5	5	6	5	1

Table 7: Assessors scores by assignment for *Developing argument, argumentation*

	Construct: Developing argument; argumentation				
Assessors	Essay A	Essay B	Essay C	Essay D	Essay E
1	1	2	5	4	3
2	Did not use construct				
3	1	2	5	4	-1
4	1	2	5	4	-1
5	5	3	1	2	3
6	Did not use construct				

Table 8: Assessors scores by assignment for *structure*

	Construct: structure				
Assessors	Essay A	Essay B	Essay C	Essay D	Essay E
1	1	2	2	3	5
2	1 2	2 5	5 5	1.5 2	0 1
3	2	3	8	5	1
4,	4	2	5	5	1
5	Did not use construct				
6	1	4	5	5	1