RESEARCH ARTICLE

The impact of peer assessment on mathematics students' understanding of marking criteria and their ability to self-regulate learning

Chris Brignell, School of Mathematical Sciences, University of Nottingham, Nottingham, UK. Email: chris.brignell@nottingham.ac.uk.
Tom Wicks, School of Mathematical Sciences, University of Nottingham, Nottingham, UK. Email: tom.wicks@nottingham.ac.uk.
Carmen Tomas, Educational Excellence Team, University of Nottingham, Nottingham, UK. Email: carmen.tomas@nottingham.ac.uk.
Jonathan Halls, Educational Excellence Team, University of Nottingham, Nottingham, UK. Email: jonathan.halls2@nottingham.ac.uk.

Abstract

At the University of Nottingham peer-assessment was piloted with the objective of assisting students to gain greater understanding of marking criteria so that students may improve their comprehension of, and solutions to, future mathematical tasks. The study resulted in improvement in all four factors of observation, emulation, self-control and self-regulation thus providing evidence of a positive impact on student learning.

The pilot involved a large first-year mathematics class who completed a formative piece of coursework prior to a problem class. At the problem class students were trained in the use of marking criteria before anonymously marking peer work. The pilot was evaluated using questionnaires (97 responses) at the beginning and end of the problem class. The questionnaires elicited students' understanding of criteria before and after the task and students' self-efficacy in relation to assessment self-control and self-regulation.

The analysis of students' descriptions of the criteria of assessment show that their understanding of the requirements for the task were expanded. After the class, explanation of the method and notation (consistent and correct) were much more present in students' descriptions. Furthermore, 67 per cent of students stated they had specific ideas on how to improve their solutions to problems in the future. Students' self-perceived abilities to self-assess and improve were positively impacted. The pilot gives strong evidence for the use of peer-assessment to develop students' competencies as assessors, both in terms of their understanding of marking criteria and more broadly their ability to self-assess and regulate their learning.

Keywords: peer-assessment, assessment criteria, formative assessment, rubric-based scoring, analytic rubrics.

1. Introduction

1.1. Assessment context – NSS and marking criteria

In the UK it is established that assessment related National Student Survey (NSS) questions perform consistently lower than the other areas of satisfaction or even the overall satisfaction. This study pays attention to a particular element of the satisfaction with assessment: 'assessment criteria have been made clear to me in advance'. The significance of this question is primarily about validity of assessments. For an assessment to be valid, the expected or required performance should be understood by all stakeholders, of which students are a primary one (Messick, 1994). Whilst the NSS

questions have adopted a political significance, particularly with the introduction of regulatory subject level Teaching Excellence Framework (TEF) ratings, this paper explores how marking criteria can be communicated clearly to students in advance of mathematics assessments.

The very nature of marking criteria is contested in the literature, and no less in practice, but can criteria be accurately communicated to ensure validity of the assessment? Mathematics is often marked in a holistic manner, where an overall judgement is made, but analytic rubric marking, where several judgements are made on identified individual criteria, is sometimes proposed with the intention of increasing openness and objectivity (Swan and Burkhardt, 2012).

Sadler (2009) expresses the view that criteria are intrinsically vague and cannot be defined clearly. As a consequence, he argues in favour of holistic marking and urges practitioners to engage students in the practice of 'evaluative experiences'. This particular line of argument has seen the development of a novel form of peer-assessment within mathematics known as 'comparative judgement' (Jones and Alcock, 2014; Jones and Sirl, 2017).

In contrast, several reviews have indicated rubrics are beneficial instruments for instruction by clarifying goals for all users, both markers and students (Jönsson and Svingby, 2007; Reddy and Andrade, 2010; Brookhart, 2018). Also, in agreement with Sadler's early discussion of evaluative experiences, student engagement in evaluative judgement has grown conceptually (Boud et al., 2018). There is also empirical evidence that engaging students in peer, self and co-assessment can have a positive impact on students' self-regulation, motivation and self-efficacy, their own confidence in self-perceived abilities (Winstone et al., 2017; Evans, 2013; Boud et al., 2018).

In a recent article Dawson (2017) identified 14 elements that practitioners need to define when designing rubrics and marking criteria. These elements span the objective of the assessment (what knowledge or skill is being tested) and the scoring strategy (how are marks arrived at) but also how the criteria are articulated to students and other markers required to implement the criteria. In practice, many of these optional decisions are left to the discretion of the practitioners.

1.2. Making criteria clear and self-regulation

The importance of the ways in which marking criteria are used with students has been stressed with the development of the concept of evaluative judgement (Boud et al., 2018). Evaluative judgement provides a conceptual framework for practitioners that brings together multiple known formative practices (e.g. rubrics, peer and self-assessment, use of exemplars). In the absence of the 'evaluative judgement' umbrella these practices are understood as separate methods. Evaluative judgement provides a coherent framework for practitioners to actively and explicitly promote students' ability to judge their own work and that of others. Similar concepts exist and predate evaluative judgement in the literature (e.g. evaluative competence by Sadler, 1989; assessment literacy by Price et al., 2012).

Within this evaluative judgement framework, Panadero and Broadbent (2018) make connections with self-regulated learning. Four levels of self-regulation exist (observation, emulation, self-control and self-regulation). Each level is incremental, although not in a linear fashion. In the observation level students observe an expert performing the task. In HE mathematics this might be during a lecture or problem class. Emulation is students performing the task themselves in the presence of the example. For mathematics students this might be a formative homework task attempted using the lecture notes as guidance. However, our aim is for students to reach self-control and self-regulation where students can attempt similar, and then unseen, problems in the absence of experts or model answers. Panadero and Broadbent (2018) propose that rubrics and peer-assessment tasks can assist students in achieving this aim. This framework for instruction including the use of rubrics

derives from a pre-existing evidence base of the positive impact of the formative use of rubrics for learning (Panadero and Jönsson, 2013).

1.3. Scoring systems in mathematics

Swan and Burkhardt (2012) note that while all assessment involves judgement, scoring systems in mathematics tend to reward answers, which is quick and objective, rather than mathematical reasoning, which is arguably more important but harder to judge. The main scoring systems currently in use are summarised below.

- Point-based scoring. A common and traditional scoring system where numerical marks are awarded for method, accuracy or explanation at each step of the solution. While easy to implement, marks are task-specific rather than an absolute measure of mathematical ability.
- Criterion-based scoring. The whole response is assigned a level based on pre-defined descriptors. The descriptors enable the student to be measured against absolute standards but converting levels to numerical scores is somewhat subjective.
- Rubric-based scoring. This retains the holistic element of criterion-based scoring but levels are awarded for different elements of performance, e.g. formulating a model or interpreting an answer, to pinpoint areas of strength and weakness.
- Comparative judgement. Responses are ranked by making relative judgements rather than judgements against criteria, which may be easier for inexperienced markers. However, scores are norm-referenced and the basis for judgements can be unclear.

Newton (1996) shows that point-based scoring has high reliability. However, our experience is that students don't benefit from points-based scoring. In the 2018 NSS, only 69% of mathematics students at our institution agreed that marking criteria were made clear, and only 64% agreed that they had received helpful comments on their work. Similarly, in a focus group in 2017/18, four out of seven of our mathematics students reported the current feedback did not help them understand where marks had been discounted.

Swan and Burkhardt (2012) suggest that criterion-based scoring is more useful for formative work because of its ability to feedforward to unseen tasks. Rubric-based scoring gives a more detailed judgement that communicates which facets of an answer are valued. See Mertler (2001) for examples of criterion-based and rubric-based scoring methods.

1.4. Objective of the present study

This study, in seeking to enhance transparency of assessment criteria to mathematics students, trialled the use of analytic rubric-based scoring and peer assessment. In the context of mathematics, we wanted to evaluate the impact of these alternative instructional approaches on helping mathematics students to become self-regulated learners. The study aimed to provide insights into students' understanding of expectations of the present task (self-control) and their ability to plan future actions (self-regulation). Both aspects are crucial to student autonomy and evaluative judgement.

2. Developing and evaluating a peer-assessment activity

2.1. Preparation and class format

Approximately 250 first year mathematics undergraduates studying a compulsory first year module were asked to complete three questions for homework. A copy of the questions is in appendix 1. The questions related to linear systems of equations and the use of row and column operations to invert

matrices and find determinants. Solutions were submitted in advance of the class and anonymised via the use of student identification numbers.

At the beginning of the class a module lecturer explained the aims of the activity, to help students understand the marking criteria so they can self-evaluate the work. Students were provided with a copy of the model solutions and the marking criteria, as shown in table 1. An example mock script was projected on the screen and the lecturer demonstrated the application of the criteria and anticipated common errors were discussed.

Students marked an anonymous piece of coursework, rating the script for accuracy and clarity using the level descriptors in table 1. They also provided written feedback to justify their decision, and discussed in small groups of two to four students their perceptions and decisions on scoring.

The class was concluded with a class discussion where around five students were invited to present their scoring and feedback. Marked scripts were then returned to staff for checking prior to being returned to students via personal tutors.

Table 1: Original marking criteria rubric

Mark	Accuracy	Clarity
0	No genuine attempt made at answering the question.	No genuine attempt is made to explain the method or use correct notation.
1	The solution contains multiple errors.	There is little explanation or correct use of notation.
2	The correct method is applied but with one or two minor errors (e.g., incorrect addition at an intermediate step).	Most steps are explained and notation is mostly correct.
3	The method is correctly applied with no minor errors.	Clear explanation of method. Consistent and correct use of notation throughout.

2.2. Evaluation

Students completed a questionnaire at the beginning and end of the class in order to capture students' understanding of criteria in the assessment before the class and after. A copy of the questionnaire is in appendix 2. In particular, students were asked, 'In your own words, what makes a problem solution excellent?' both pre-class and post-class. In order to capture the impact on planning actions, in the post-class questionnaire they were also asked, 'Have you had any specific ideas on how to improve your solutions to problems in the future? If YES, please describe these briefly.'

One part of evaluative judgement and students' autonomy is linked to confidence in their own abilities. A Likert scale question posed before and after the class aimed to capture their confidence in performing well in types of assessment and their ability to self-assess.

Table 2: Subdivision of marking criteria rubric

	Accuracy		Clarity	
Mark	Method	Errors	Explanation	Notation
0	No genuine attempt made at answering the question	No genuine attempt made at answering the question	No genuine attempt is made to explain the method	No genuine attempt is made to use correct notation
1	An incorrect method is applied	The solution contains multiple errors	There is little explanation	There is little correct use of notation
2	The correct method is applied	One or two minor errors	Most steps are explained	Notation is mostly correct
3	The method is correctly applied	No minor errors	Clear explanation of method	Consistent and correct use of notation throughout

2.3. Data analysis

Students' responses to the questionnaire were originally coded according to the marking criteria rubric, shown in table 1. However, it was decided that the areas of Accuracy and Clarity could be subdivided into Method and Errors, and Explanations and Notation, respectively. This subdivision is shown in table 2. It is maintained that this refined rubric is more suited to the current analysis, as the amalgamation of areas could mask differences in student responses between pre-class and post-class. For example, students discussing Method pre-class and Method and Errors post-class would be viewed as scoring the same if their responses are coded only by the area of Accuracy.

Investigation of students' responses indicated that some discussed *legibility*. For example, they suggested that an excellent answer should be 'legible', 'neatly presented' or 'clearly written'. Therefore, students' responses were also analysed for the area of legibility. In the initial whole class discussions, legibility of answers (in terms of neatness of handwriting) had not been discussed.

Consequently, students' responses were coded for five categories: Method, Errors, Explanation, Notation and Legibility. Students' responses were marked for presence (1) or absence (0) of each category. To establish the extent of inter-rater reliability, 16% of the data were coded by a second researcher. Cohen Kappa indicated a very high level of agreement ($K = .953$, $p < .001$). McNemar's test was used to test for significant differences from pre- to post-class.

Students' self-reported ratings in their confidence at assessing their own work and writing good solutions, both pre-class and post-class, on a Likert scale from 1 (not confident at all) to 5 (very confident) were analysed using a related samples Wilcoxon signed ranks test.

2.4. Sample and Ethics

In total 97 students responded to the pre- and post-class questionnaire, but not all questions were answered by all students. Where appropriate, the number of responses to specific questions is provided in the analysis below. Prior to data collection, students were informed the questionnaire was optional, that data would be anonymised and separate to any other assessment activities, and asked to provide informed consent. The ethical procedures applied in this study were approved by a University of Nottingham ethics committee and we report no conflicts of interest.

3. Results

3.1. Students' awareness of task requirements: observation level

To evaluate students' awareness of marking criteria, data is taken from students' pre- and post-class explanations of 'what makes a problem solution excellent?'. Table 3 shows the percentage of responses which related to the five coding categories (see table 2). It shows improvement in all five categories from pre- to post-class. That is, all areas of the subdivided rubric (Method, Errors, Explanation and Notation) were discussed more post-class than pre-class. A similar increase was also observed for responses discussing legibility. Analysis using McNemar's test indicates that there was a significant difference from pre- to post-class in the areas of Method, Explanation and Notation.

Table 3: Differences in rubric areas mentioned in students' responses (n=80)

Area	Pre-class percentage of participants	Post-class percentage of participants	Percentage point change	*p.*
Method	32.5%	52.5%	20.0	.001
Errors	42.5%	55.5%	10.0	.096
Explanation	75.0%	87.5%	12.5	.041
Notation	2.5%	22.5%	20.0	<.001
Legibility	17.5%	26.3%	8.8	.118

3.2. Student awareness of their own abilities: self-control

Students were asked to rate their confidence at assessing their own work and writing good solutions, both pre-class and post-class, on a Likert scale from 1 (not confident at all) to 5 (very confident). In both categories the median rating rose from 3 pre-class to 4 post-class, with 47% of respondents reporting an increased confidence in assessing their own work and 61% reporting an increased confidence in writing good solutions. A Wilcoxon signed-rank test showed both these improvements to be significant ($p < .001$).

3.3. What to include in future assessments: self-regulation

When asked 'have you had any specific ideas on how to improve your solutions to problems in the future?', 67.3% of students stated that they did have ($n = 95$). Table 4 shows the rubric areas that related to students' explanation about what to include in future work. The table shows no student discussed the rubric areas of Method or Errors. The most popular areas discussed were Explanation and Notation. These results do not map to the significant improvements found in the previous section (see table 3). Refinement of the questionnaire to tailor it to the types of question being discussed may improve the reliability of the responses.

Table 4: What to include in future work, coded by rubric area (n=63)

Area	Percentage of participants
Method	0.0%
Errors	0.0%
Explanation	84.1%
Notation	31.7%
Legibility	17.5%

4. Discussion

The ubiquity of point-based scoring in mathematics for summative assessment is likely to prevail for some time to come. Its convenience, speed and high reliability are all good reasons for its dominance in a landscape dominated by traditional closed book exams (Iannone and Simpson, 2011). However, for formative assessment, the peer assessment activity described in this article has been found to be highly effective in enhancing students' understanding of both the current task's requirements and their ability to plan next steps to improve solutions to future unseen problems.

This contrasts to standard feedback methods, such as annotations on student scripts or general feedback to the class, which may highlight errors with the solution to the present task but might not enhance the feed-forward to future solutions. Indeed, this cohort has been exposed to these standard feedback methods previously, so the fact this task gave students new insights demonstrates that peer-assessment using absolute performance descriptors provided students with fresh understanding of marking criteria that was not gained from standard feedback and point-based scoring implemented previously. The dissatisfaction in the quality and quantity of feedback to mathematics students to support student learning is well-documented (Bidgood and Cox, 2002), but this study shows further research into the use of rubric-based scoring for formative assessment is worth pursuing.

Other forms of peer-assessment, such as comparative judgement (Jones & Alcock, 2014), are available. However, rubric based scoring with explicitly stated criteria has been shown here to enhance aspects of observation (students understanding what is required) and of self-regulation (planning actions). By contrast, comparative judgement requires students to make judgements in the absence of absolute performance criteria. This could be an advantage to inexperienced markers but could also lead to judgements based on tangential criteria. For example, this study shows students became distracted by the legibility of the writing as a proxy for the level of explanation.

This small-scale study shows a positive impact on student learning. Further research is needed to investigate the integration of peer-assessment with other forms of evaluative judgement, such as self-assessment, over a longer period of time in a mathematics learning context. The present study is a pilot of a much larger project in mathematics and another university-wide initiative to see the development of longer-term approaches to developing students' evaluative judgement in multiple subject areas.

5. References

Bidgood, P. & Cox, B., 2002. Student Assessment in MSOR. *MSOR Connections*, 2(4), pp.9-13.

Boud, D., Ajjawi, R., Dawson, P. & Tai, J., eds., 2018. *Developing evaluative judgement in higher education: Assessment for knowing and producing quality work.* Abingdon, Oxon: Routledge.

Brookhart, S.M., 2018. Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, 3, 22. https://doi.org/10.3389/feduc.2018.00022

Dawson, P., 2017. Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*, 42, pp.347-360. https://doi.org/10.1080/02602938.2015.1111294

Evans, C., 2013. Making sense of assessment feedback in Higher Education. *Review of Educational Research*, 83, pp.70-120. https://doi.org/10.3102/0034654312474350

Iannone, P. & Simpson, A., 2011. The summative assessment diet: How we assess in mathematics degrees. *Teaching Mathematics and its Applications*, 30, pp.186-196. https://doi.org/10.1093/teamat/hrr017

Jones, I. & Alcock, L., 2014. Peer-assessment without assessment criteria. *Studies in Higher Education*, 39, pp.1774-1787. https://doi.org/10.1080/03075079.2013.821974

Jones, I. & Sirl, D., 2017. Peer assessment of mathematical understanding using comparative judgement. *Nordic Studies in Mathematics Education*, 22, pp.147-164.

Jönsson, A. & Svingby, G., 2007. The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review*, 22, pp.130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Mertler, C.A., 2001. Designing scoring rubrics for your classroom. *Practical Assessment, Research and Evaluation*, 7, pp.1-10. Available at: https://pareonline.net/getvn.asp?v=7&n=25 [Accessed 4 September 2019].

Messick, S., 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, pp.13-23. https://doi.org/10.3102/0013189X023002013

Newton, P.E., 1996. The reliability of marking of GCSE scripts: mathematics and English. *British Educational Research Journal*, 22, pp.405-420. https://doi.org/10.1080/0141192960220403

Panadero, E. & Broadbent, J., 2018. Developing evaluative judgement: a self-regulated learning perspective. In D. Boud, R. Ajjawi, P. Dawson & J.Tai, eds, *Developing evaluative judgement in higher education: Assessment for knowing and producing quality work*. Abingdon, Oxon: Routledge.

Panadero, E. & Jönsson, A., 2013. The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, 9, pp.129-144. https://doi.org/10.1016/j.edurev.2013.01.002

Price, M., Rust, C., O'Donovan, B., Handley, K. & Bryant, R., 2012. *Assessment literacy: The foundation for improving student learning.* Oxford Centre, for Staff and Learning Development.

Reddy, Y.M. & Andrade, H., 2010. A review of rubric use in higher education. *Assessment and Evaluation in Higher Education*, 35, pp.435-488. https://doi.org/10.1080/02602930902862859

Sadler, D.R., 2009. Indeterminacy in the use of preset criteria for assessment and grading. Assessment & Evaluation in Higher Education, 34, pp.159-179. https://doi.org/10.1080/02602930801956059

Swan M. & Burkhardt H., 2012. Designing assessment of performance in mathematics. Educational designer, 2, pp.1-41. Available at: https://www.educationaldesigner.org/ed/volume2/issue5/article19/ [Accessed 4 September 2019].

Winstone, N.E., Nash, R.A., Parker, M. & Rowntree, J., 2017. Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist*, 52, pp.17-37. https://doi.org/10.1080/00461520.2016.1207538

6. Appendices

Appendix 1 – Homework questions

1 Write the system of linear, simultaneous equations

$$\begin{aligned} x + 2z &= 10 \\ 2x + 3y + z &= 5 \\ y + z &= 3 \end{aligned}$$

in matrix form. Use the Gauss-Jordan (and no other) method to find the inverse of the matrix and hence find the solution to the system.

2 By performing suitable row and column operations show that

$$\begin{vmatrix} x & y & z \\ y & x & x \\ z & z & y \end{vmatrix} = (x-y)(y-z)\alpha(x,y,z)$$

where $\alpha(x,y,z)$ is a linear term in x, y, z which you should determine.

3 By performing suitable row and column operations show that

$$\begin{vmatrix} x+2 & 3 & 3 \\ 3 & x+4 & 5 \\ 3 & 5 & x+4 \end{vmatrix} = 0$$

has solutions $x = 0, 1, \beta$ where the value of the constant β is to be calculated.

Appendix 2 – Questionnaire

BEFORE TAKING PART IN THE WORKSHOP TODAY

Please circle your answers.

1 Please rate how nervous you feel about the coursework

Very nervous 1 2 3 4 5 Not nervous at all

2 Completing the coursework seems…

Very difficult 1 2 3 4 5 Very easy

3 Please rate how confident you feel about assessing your own coursework

Not confident at all 1 2 3 4 5 Very confident

4 Please rate how confident you feel about how to go about writing good solutions to problems

Not confident at all 1 2 3 4 5 Very confident

5 In your own words, what makes a problem solution excellent?

Please complete the rest of the survey at the end of the workshop

AFTER COMPLETING THE WORKSHOP

6 Please rate how nervous you feel about the coursework

Very nervous 1 2 3 4 5 Not nervous at all

7 Completing the coursework seems…

Very difficult 1 2 3 4 5 Very easy

8 Please rate how confident you feel about assessing your own coursework

Not confident at all 1 2 3 4 5 Very confident

9 Please rate how confident you feel about how to go about writing good solutions to problems

Not confident at all 1 2 3 4 5 Very confident

10 Have you had any specific ideas on how to improve your solutions to problems in the future?

YES NO

If YES, please describe these briefly: