Student peer assessment using adaptive comparative judgment: Grading accuracy versus quality of feedback

Practitioner Research
In Higher Education
Copyright © 2019
University of Cumbria
Vol 12(1) pages 50-60

Constantinos Demonacos, Steven Ellis & Jill Barber
University of Manchester

Abstract

In this study we explored the potential of adaptive comparative judgement (ACJ) as a medium for peer assessment and for the giving and receiving of peer feedback. ACJ is a marking protocol in which the assessor (or judge) merely compares two answers and chooses a winner. Repeated judgements and a suitable sorting algorithm allow marks to be derived from a rank order of scripts. Feedback can be added to each script. In this case study (a 500 word report in year 3 of a pharmacy programme) each student gave feedback to 10 others and the overall feedback standard was high, but, as judges, students were inconsistent with one another and with staff assessment. This contrasts with a previous exercise, in which a robust assessment was achieved but feedback was less good. A hierarchical marking scheme and explicit feedback guidelines may be key to optimising ACJ-based student peer assessment.

Keywords
Adaptive Comparative Judgement, Peer Assessment, Summative Assessment, Marking Schemes.

Introduction

For academic staff it is often the case that "an hour spent in marking is an hour less of life" (John Sargeant, 2004). Students may, however, gain much from assessing and giving feedback on one another's work; in particular, they can see what very good work looks like, and reflect on their own practice in the light of their peers' experience. Barriers to the use of peer assessment include the perception that student marking is unreliable, exacerbated by the fact (in our experience) that students are reluctant to apply mark schemes to other students' work, tending to award very high marks without discrimination. Software to support the use of adaptive comparative judgement is now available and, by simplifying the marking process to repeatedly answering the single question "Which is better?" has the potential to bring student peer assessment into mainstream practice. The aim of this study was to evaluate the use of adaptive comparative judgement in student peer assessment.

Adaptive Comparative Judgement (ACJ) is an alternative to conventional marking in which the assessor (or judge) merely compares two answers and chooses a winner (Thurstone, 1927; Pollitt, 2012). The use of a suitable sorting algorithm means that repeated comparisons lead to scripts sorted in order of merit. Boundaries are determined by separate review of scripts. In practical terms, the assessor sees two scripts arranged side-by-side; based on the criteria given, (s)he decides which is better and clicks the appropriate "winner" button. The next pair of scripts is then displayed. Scripts of about 500 words appear as a single page, but scrolling is necessary for reading longer pieces of work. The method does much to remove subjective bias and to allow "hawks" and "doves" to mark successfully in a team. (Daly et al., 2017; Barber 2018).

There is a considerable body of literature devoted to the reliability of adaptive comparative judgement (and of comparative judgement generally). Reliability is mathematically complicated but is essentially a measure of reproducibility (will the order be the same if a different group of expert judges are

Citation
Demonacos, C., Ellis, S., Barber, J.(2019) 'Student peer assessment using adaptive comparative judgment: grading accuracy versus quality of feedback', *Practitioner Research in Higher Education Journal*, 12(1), pp. 50-60.

involved, or if the same group judges on a different day?). Politt (2012) and others (Kimble et al., 2007, Kimble et al., 2009) have claimed very high reliability in the final order obtained by ACJ (typically > 0.93)[1], compared with 0.6 – 0.7 using conventional marking schemes although this very high level of reliability has been disputed, notably by Bramley (2015). Getting the scripts into the correct rank order is only part of the problem, however. It is equally important that the separation is accurate. In student work, the marks tend to be normally distributed, so that there is very little separation between scripts in the middle of the group, but there might be a large difference in marks between (for example) the top and the second script. Scale separation has been explored in detail by Verhavert et al. (2017).

While a great deal of attention has been paid to the accuracy achievable with adaptive comparative judgement, rather little has been paid to the accuracy *required*. This will vary considerably from exercise to exercise and setting to setting, but can be quite low. A typical assessment in a UK university (like the one described in this paper) might carry 10% of the marks in a 20 credit third year unit, and no attempt would be made to mark to closer than 0.5 or 0.25 marks out of 10. The most common marks (typically 6/10 or 6.5/10) are likely to be awarded to more than 10% of the cohort, meaning that a precise order is not necessary unless the cohort is very small.

We were especially keen to explore the use of ACJ as a vehicle for peer assessment and feedback. The literature points to the value of peer assessment and feedback. Nicol et al (2014) and Cho and MacArthur (2011) note that students benefit from both giving and receiving feedback. The language used by fellow students tends to be accessible (Topping, 1998; Falchikov, 2005). Bloxham and West (2004) indicate that peer feedback can help students to understand the assessment process. Another potential benefit of peer assessment by ACJ is that students comment on average on 10-15 pieces of work and receive 10-15 pieces of feedback from different people, a variety that staff seldom match. Our preliminary data (Barber, 2018) indicated that students are capable of sufficiently accurate marking by ACJ and the students reported finding the process of judging very helpful to their learning.

The aim of the present study was to assess the value of adaptive comparative judgement in a summative assessment in the third year of a four year undergraduate pharmacy programme, using students as the judges. Both the reliability of student judgements and the quality of student feedback were assessed. The exercise chosen (a case study on the biology and treatment of a patient with complex needs) was one that students historically found difficult and responded to badly.

Methods

The unit of the curriculum chosen for this study was “Integrated Research Skills” which carries 20 credits in the third year. Students carry out a number of exercises designed to prepare them for the final year project module (30 credits). The other exercises are an audit report (40%), a critique (40%) and a laboratory practical (10%).

The present exercise carries 10% of the final mark. As well as appearing on the Virtual Learning Environment, the exercise was presented to students in detail in the course of a lecture, as shown in Figure 1. In addition to preparing students for their final year project, this unit is intended to illustrate the integration of science with pharmacy practice. Thus, the exercise includes explaining the underlying biology of a condition (in this case multiple sclerosis) to a patient. The student requires both an understanding of the disease and its therapy, and an ability to empathise with a patient and to express their knowledge in plain English.

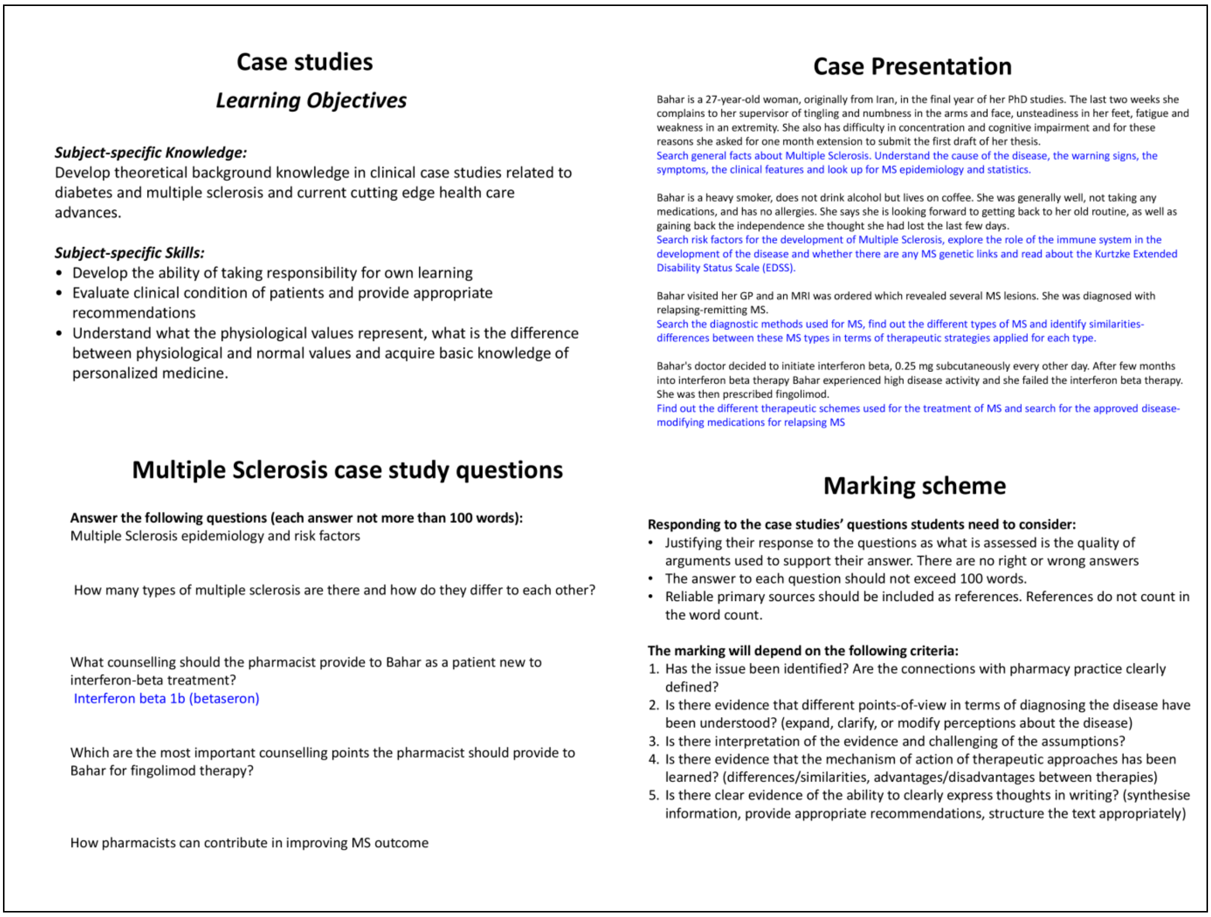
[1] The “reliability” measure (the square of standardised residuals) is mathematically complex and beyond the scope of this report. It is explained in Pollitt (2012) and in Bramley (2015).

The exercise consisted of five parts, each with a maximum of 100 words. It was submitted through a Turnitin dropbox so that any plagiarism would be detectable.

ACJ was carried out using the commercially available software CompareAssess (https://compareassess.com). The complete collection of student file submissions was downloaded from Blackboard / Turnitin as a zipped archive. It is possible to assign each file a unique identifier manually, and we have done this in the past. In the present exercise, some automation was achieved as follows. The filenames of those files included system assigned student ID codes set alongside user submitted title text. To standardise file naming and to minimise propagation of identifying information within a 3rd party system like CompareAssess, these filenames were manipulated before each file was submitted for judgement. To do so the listing of filenames was saved to disk and then modified programmatically so that universally unique identifiers (uuid) became associated with each line of that file listing. This produced a tab separated collection of information that was used as the source in a file renaming step and retained for the final collation of feedback. Renamed files were submitted to CompareAssess and used for the comparative analysis judgement sessions. 19 rounds of judging were completed, so that each student made 10 judgements and commented on about 15 scripts.

Once all rounds of judgements had concluded the data available within the CompareAssess system was regularly and tightly associable with the assigned uuids. Complete anonymity was therefore assured and no third party could associate comments or scripts with individual students. Students were further informed of publication and given the opportunity to withdraw the use of their anonymous comments.

On this occasion, staff also provided marks and some feedback for each script.



Case studies

Learning Objectives

Subject-specific Knowledge:
Develop theoretical background knowledge in clinical case studies related to diabetes and multiple sclerosis and current cutting edge health care advances.

Subject-specific Skills:
- Develop the ability of taking responsibility for own learning
- Evaluate clinical condition of patients and provide appropriate recommendations
- Understand what the physiological values represent, what is the difference between physiological and normal values and acquire basic knowledge of personalized medicine.

Multiple Sclerosis case study questions

Answer the following questions (each answer not more than 100 words):
Multiple Sclerosis epidemiology and risk factors

How many types of multiple sclerosis are there and how do they differ to each other?

What counselling should the pharmacist provide to Bahar as a patient new to interferon-beta treatment?
Interferon beta 1b (betaseron)

Which are the most important counselling points the pharmacist should provide to Bahar for fingolimod therapy?

How pharmacists can contribute in improving MS outcome

Case Presentation

Bahar is a 27-year-old woman, originally from Iran, in the final year of her PhD studies. The last two weeks she complains to her supervisor of tingling and numbness in the arms and face, unsteadiness in her feet, fatigue and weakness in an extremity. She also has difficulty in concentration and cognitive impairment and for these reasons she asked for one month extension to submit the first draft of her thesis.
Search general facts about Multiple Sclerosis. Understand the cause of the disease, the warning signs, the symptoms, the clinical features and look up for MS epidemiology and statistics.

Bahar is a heavy smoker, does not drink alcohol but lives on coffee. She was generally well, not taking any medications, and has no allergies. She says she is looking forward to getting back to her old routine, as well as gaining back the independence she thought she had lost the last few days.
Search risk factors for the development of Multiple Sclerosis, explore the role of the immune system in the development of the disease and whether there are any MS genetic links and read about the Kurtzke Extended Disability Status Scale (EDSS).

Bahar visited her GP and an MRI was ordered which revealed several MS lesions. She was diagnosed with relapsing-remitting MS.
Search the diagnostic methods used for MS, find out the different types of MS and identify similarities-differences between these MS types in terms of therapeutic strategies applied for each type.

Bahar's doctor decided to initiate interferon beta, 0.25 mg subcutaneously every other day. After few months into interferon beta therapy Bahar experienced high disease activity and she failed the interferon beta therapy. She was then prescribed fingolimod.
Find out the different therapeutic schemes used for the treatment of MS and search for the approved disease-modifying medications for relapsing MS

Marking scheme

Responding to the case studies' questions students need to consider:
- Justifying their response to the questions as what is assessed is the quality of arguments used to support their answer. There are no right or wrong answers
- The answer to each question should not exceed 100 words.
- Reliable primary sources should be included as references. References do not count in the word count.

The marking will depend on the following criteria:
1. Has the issue been identified? Are the connections with pharmacy practice clearly defined?
2. Is there evidence that different points-of-view in terms of diagnosing the disease have been understood? (expand, clarify, or modify perceptions about the disease)
3. Is there interpretation of the evidence and challenging of the assumptions?
4. Is there evidence that the mechanism of action of therapeutic approaches has been learned? (differences/similarities, advantages/disadvantages between therapies)
5. Is there clear evidence of the ability to clearly express thoughts in writing? (synthesise information, provide appropriate recommendations, structure the text appropriately)

Figure 1. Slides introducing the case study to students. A list of references was also given.

The CompareAssess systems presents several useful data fields downloadable as summary reports, however those that contained peer feedback were presented on a document by document basis and might have required laborious manual activity to retrieve. (We understand that this will be improved in a later version of the software.) To avoid this, data scraping routines were implemented using web browser automations which iterated through each of these reports in order to extract peer feedback into structured objects. This information was then readily available to be imported by other systems were it became re-associated with identifying information retained above.

Data Analysis

The software provides information if any of the judges or the scripts were misfits. It also gives an average judgement time for each judge. Data for misfit judges and scripts were analysed visually.

Results

It is a sad fact that among 130 students there will usually be some who do the minimum necessary to pass an assignment, irrespective of any negative effects on their peers. It was therefore necessary to check the parameter "average time per judgement". Two students spent a little less than one minute on each judgement. The quality of feedback from all students was checked in this exercise, but hindsight and further exercises have taught us that it would be sensible to check these two explicitly.

Student marking

Next, we examined the "Judge misfit" panel (Figure 2A). Six judges appeared above the red line, indicating that their judgements are not consistent with those of the other judges. Previous peer assessment exercises have generally seen only one or two misfit judges. Again, hindsight teaches us to examine those judgements explicitly. The "Script misfit" panel (Figure 2B) allows scripts that are difficult to mark to be identified, and these should be (and were) moderated by staff.

The reliability (Figure 2C) was disappointing at 0.55, and the error bars indicate considerable uncertainty about the final rank order. This level of reliability is similar to that of two other peer-assessed exercises, although one previous exercise, carried out by 64 fourth year students, gave a reliability of 0.92.

In this case scripts 39 – 101 have overlapping error bars, and we cannot rigorously exclude the possibility of their being wrongly ordered. These scripts, however, by either manual or CompareAssess marking span a mark range of only 15% (6.0 – 7.5 /10), corresponding to 1.5% of the unit mark or 0.09% of the final degree mark.

Figure 3 shows the correlation between staff marks and student marks awarded by adaptive comparative judgement. There is no doubt about which script is the best, and the general trend is as expected. Nevertheless, there is considerable scatter, with some disagreements of more than 2 marks between staff marks and student marks by ACJ.

On this occasion, the staff marks were used to contribute to the overall unit mark.

Feedback

There was some poor constructed student feedback that was removed from the final text delivered to students. There was also evidence, in the case of one or two students, of copying and pasting the same feedback to all students. Generally, however, the quality and reliability of the feedback was excellent. Examples are given in Table 1.

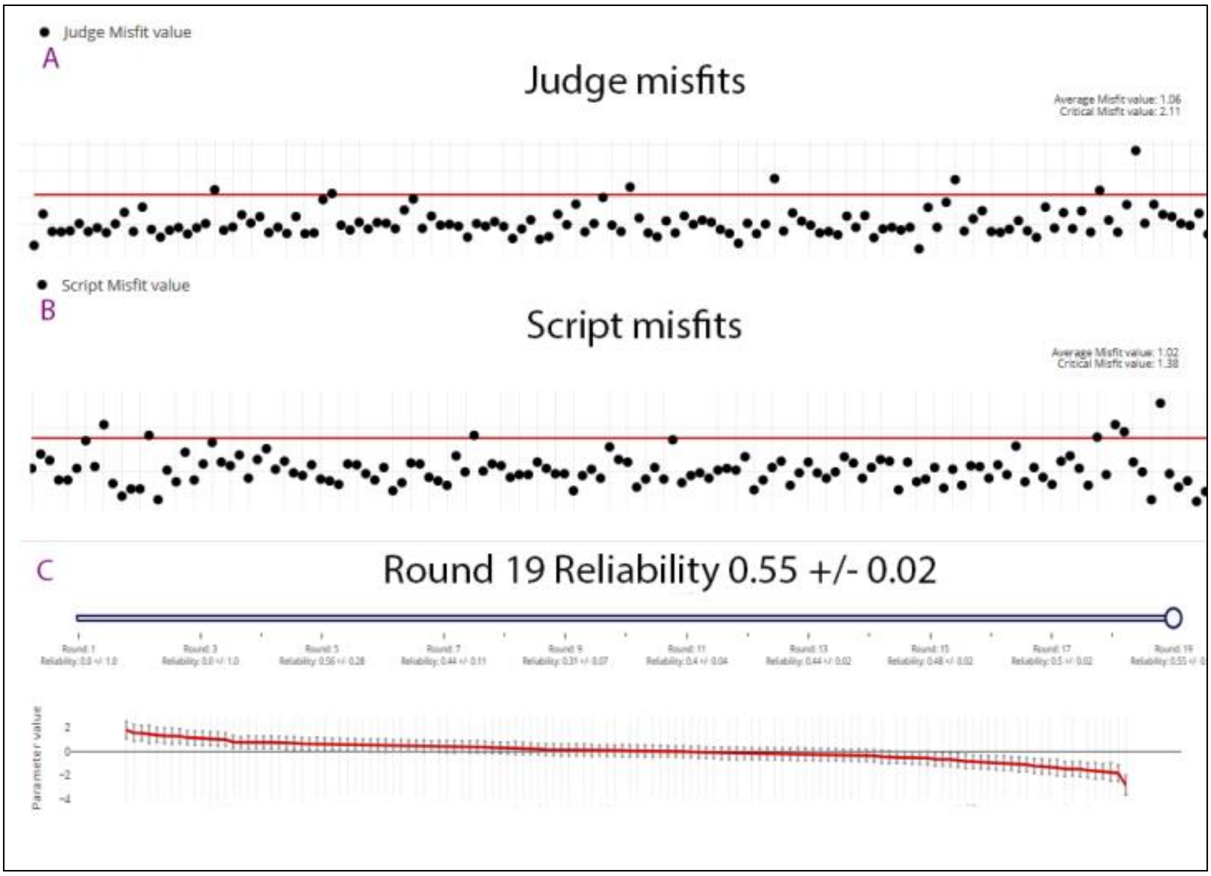


Figure 2. A Judge misfit panel showing 6 judges above the critical misfit line. These judges gave judgements somewhat inconsistent with the rest of the cohort, and their judgements merit closer inspection. B Script misfit panel, showing six scripts above the critical misfit line; these scripts proved difficult to judge, and should certainly be considered by the staff moderating the exercise. C Reliability after 19 rounds.

What students think

We have previously determined student views on feedback from adaptive comparative judgement (Barber 2018), and received very positive responses, including requests to use the system with work from previous year groups. We did not formally assess students' reactions to this exercise, but the unit as a whole is better received by students than in previous years, and, informally, students report the value of seeing their peers' work. The most positive comments about this exercise were received a semester after its completion. Students beginning their fourth year project work have reported (at a scheduled feedback session) that they now see the value of this exercise.

The feedback on the best script is especially detailed, suggesting that students studied this script carefully.

Table 1. Examples of student feedback. Top: in response to the best script scoring 9.75/10. Bottom: in response to a script scoring 6.25/10.

very detailed answers. good referencing sources Questions answered very well with all points covered. Good presentation and obvious research has been conducted and used to back statements with evidence from the literature. Case taken into consideration when discussing side effects and lifestyle changes in questions 3, 4 and 5 Clearly differentiates between the type of multiple sclerosis. The counselling point is suitable for Bahar. There is a good connection for pharmacy practice such as smoking cessation. The therapeutic approach is well justified. Student has given statistics illustrating how prevalent the disease is and which ages are most likely to be affected. A variety of risk factors have been given which gives a better understanding as to how MS could potentially increase its risk. Types of MS were clearly defined and explained showing clearly the differences between all types and what occurs in each. Counselling points were patient specific to Bahar and were well explained, ensuring patient knew how to avoid these symptoms from occurring and what to do if they arose. Contributions the pharmacist can make were well explained and appropriate which showed quality for patient care. This student has been able to include a vast amount of information regarding the condition despite a very low word count being available. Questions were answered very well however, I believe these questions could have been answered in a more patient friendly manner. Positives: Good explanation of the different types of MS. Good at relating back to pharmacy practice. Good referencing and use of journals. Improvements: Need to advise Bahar on taking contraception whilst taking fingolimod Issue has been identified clearly, there is interpretation of the evidence and challenging of the assumptions, the mechanism of action of therapeutic approaches has been described. The similarities, differences, advantages and disadvantages between the therapies have been described. There is clear evidence of the ability to clearly express thoughts. The text is structured well.
Poor range of references No mention of the dose regimen, sterile injection techniques or rotating injection site for interferon Basic counselling points for fingolimod Regarding patient centred counselling perhaps discussions regarding pregnancy and contraception could have been included. Smoking cessation could have been included within the role of the pharmacist section. Epidemiology is well talked about and risk factors have been mentioned well in the first question. The types of MS are clearly written and distinguished between them. The last question is poorly written with no emphasis on the use of Pharmacy services such as MUR, NMS or smoking cessation. Referencing also needs to be improved as there's very little references and poor execution of referencing style. Nicely written but could focus more on the patient directly in advice Small number of references so not much support for the information given

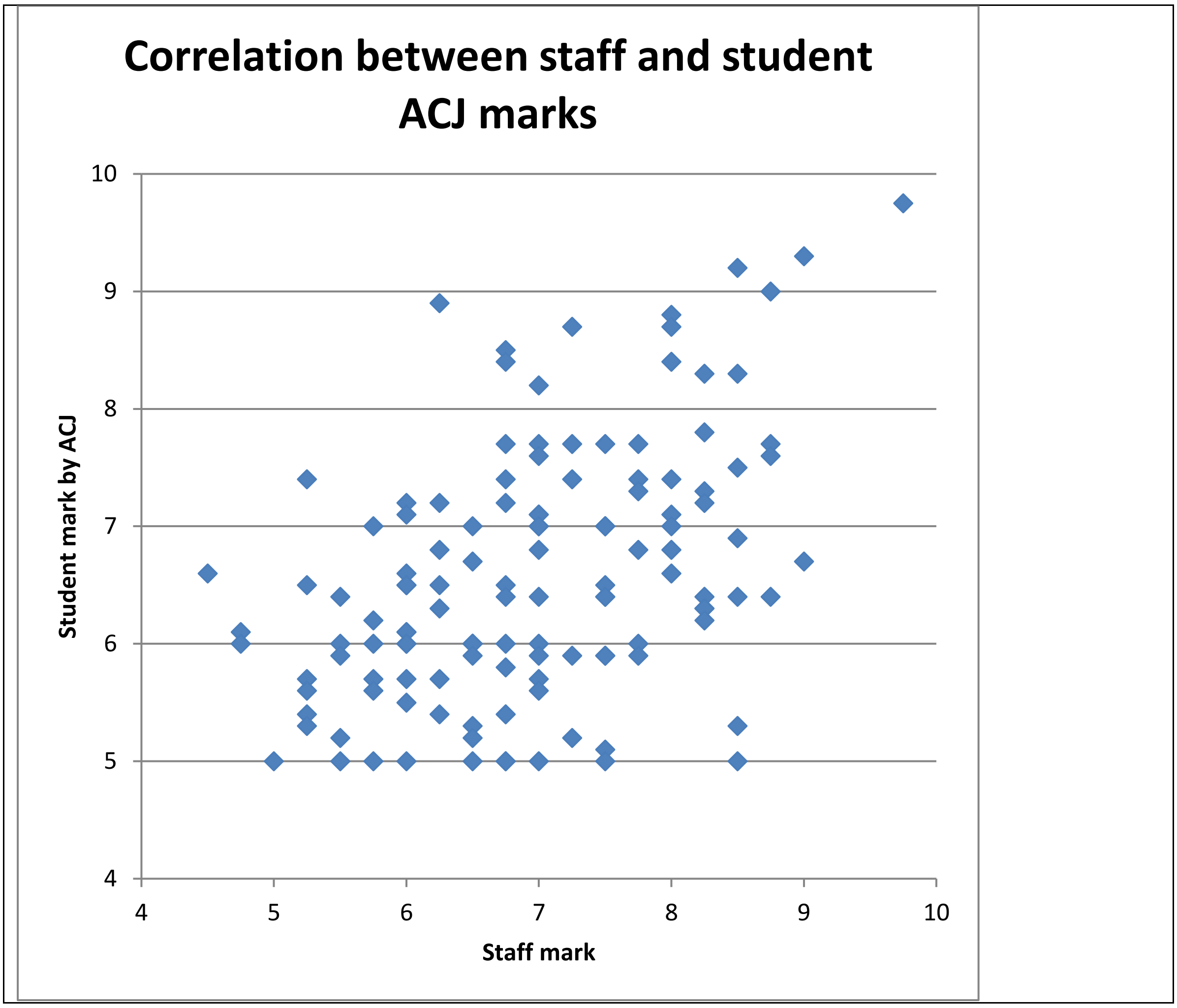


Figure 3. Correlation between staff mark and mark obtained by adaptive comparative judgement with student judges.

Discussion

It is essential that healthcare professionals work as teams, and it is counter-intuitive for their undergraduate study to be unduly competitive. Peer assessment and feedback can be used to foster a sense of community among students. The concept of the community of learners is explicit in the advertising of some MOOCs (Massive Open Online Courses) (https://www.futurelearn.com) and has been explored by commentators such as Tinto (2003 & 2008), but the culture of individual learners competing with one another remains strong. The use of student peer assessment and feedback is facilitated by adaptive comparative judgment, and provides opportunities for students to share responsibility for the learning of the group.

The authors have also been seeking methods of helping students to acquire soft skills, such as critical thinking, fair judgement and personal responsibility, in order to advance their professionalism. The high quality of the feedback offered suggests that these objectives were achieved, at least for the majority of students.

Another objective of this study was to enable students to appreciate what the best work looks like. Carless et al. (2018) are among those who have argued for the use of exemplars of best student work to aid the learning of other students. Our study is much less managed than theirs; students are not told which is the best work. Nevertheless, students are exposed to 15 or 20 examples of student answers, some good, some much less so.

Adachi et al. (2018), in a comprehensive review of the literature, identify these objectives as the major advantages of peer assessment.

The present exercise was not designed for a study of adaptive comparative judgement; rather it was chosen for the study on the basis that ACJ had the potential to improve the exercise. Adaptive comparative judgement was first used as a method of assessment in this School in 2016 and every assessment allows more lessons to be learned. We have not, however, used ACJ to replace assessments that work well; it has been used either to improve the student or staff experience of summative assessment or to allow formative assessments to take place where pressures on staff time would not permit a conventional assessment. The assessment described here involved a detailed analysis of both marks and feedback generated by peer assessment.

The peer-marking was not wholly successful, students failing wholly to agree with one another or with the academic staff marker. Two other peer assessments in the School saw similar reliability (0.5-0.6). Another (Barber, 2018) has been very successful and is shown in Figure 4.

Imagine that you have been awarded a grant of £1 million to reduce premature deaths anywhere in the world. Choose a specific disease, group of diseases or other major cause of death and describe how you would spend the money to reduce the death rate.

This will be marked by CompareAssess with the criteria:

Description of the problem
Description of the action
Costing

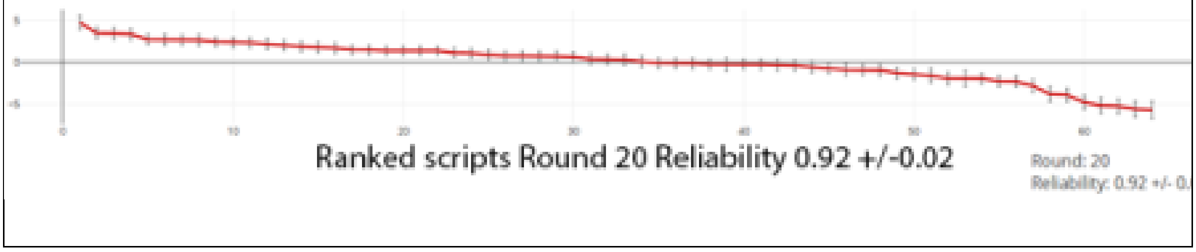


Figure 4. Example of question marked with a hierarchical marking scheme (top) with reliability statistics after round 20 (bottom).

The present exercise, however, gave rise to rich and varied feedback of a quality and quantity we very rarely see in peer feedback. Indeed, the brief feedback notes left the staff member marking the exercise were almost all in complete agreement with the feedback provided by the students. Mortier et al. (2015) have examined students' perceptions of the comparative judgement process in great detail, finding them positive overall.

This is at first sight hard to reconcile. Staff and students agree about feedback but there is no very strong agreement about marks. Close inspection of the feedback offers one possible clue. Whereas the staff member was especially concerned that the students knew the mechanism of action of the drugs, many students offering feedback focused on the language used and whether or not it was patient friendly. There is the added complication that the patient is a PhD student; she is clearly intelligent, and quite probably a scientist, so that patient-friendly language is not necessarily simple language. Adaptive comparative judgement requires that judges choose between two scripts; unless they are given very clear guidance, they will choose on the basis of what they consider to be most important.

The key difference between this exercise and the one shown in Figure 4 is therefore probably the latter assessment employed a hierarchical marking scheme. Judges were asked to address the first criterion "Has the problem been clearly described?" and if one script defined the problem and the other did not, the first must win, irrespective of the rest of the answer. If both scripts (or neither script) contained a good description of the problem, the judgement was made on the basis of the action described, and only if this also tied was the costing considered. Students were free to comment on the action or the costing but it was clear that the description of the problem took precedence in the judging process.

The use of comparative judgement in marking large bodies of student work has, until recently, been hindered by the lack of appropriate software. Now that the software is available, teachers are in a position to build expertise in its use. In the Manchester undergraduate pharmacy programme, we have successfully carried out assessments with staff as judges. We have also used the adaptive comparative judgement method to deliver successful peer assessment and successful peer feedback; the next challenge is to deliver both in the same exercise!

Conclusion

Adaptive comparative judgement has enormous potential in peer assessment and in enabling the giving and receiving of peer feedback. The method requires, however, the use of specialised, hierarchical marking schemes and there is, as yet, no body of literature or consensus about how to develop these. We will continue to try to develop protocols that combine good quality peer assessment and feedback, as part of our strategy to create communities of learners.

Acknowledgements

We thank Matt Wingfield and Declan Lynch of Digital Assess for their collaboration in this ongoing project. We also thank many students at the University of Manchester for their participation as judges.

References

Adachi, C., Tai, J. H.-M., Dawson, P. (2018) 'Academics' perceptions of the benefits and challenges of self and peer assessment in higher education', *Assessment & Evaluation in Higher Education*, 43, PP. 294-306.

Barber, J. (2018) 'Five go marking an exam question: the use of Adaptive Comparative Judgement to manage subjective bias', *Practitioner Research in Higher Educatio*n, 2018, 11, PP.94-100.

Bloxham, S., West, A. (2004) 'Understanding the rules of the game: marking peer assessment as a medium for developing students' conceptions of assessment', *Assessment & Evaluation in Higher Education*, 29, pp. 721-733.

Bramley, T. (2015) Investigating the reliability of Adaptive Comparative Judgment. Cambridge Assessment Research Report, 23 March 2015. Available at http://www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf (Accessed: 30 March 2018).

Carless, D., Chan, K.K.H., To, J., Lo, M., Barrett, E. (2018) Developing students' capacities for evaluative judgement through analysing exemplars, in Boud, D., Ajjawi, R., Dawson, P and Tai, J. (eds.), Developing Evaluative Judgement in *Higher Education: Assessment for knowing and producing quality work*. London: Routledge.

Cho, K and MacArthur, C (2011) 'Learning by reviewing', *J. Educational Psychology*, 103, pp. 73-84.

Daly, M.; Salmonson, Y.; Glew, P.J. and Everett, B. (2017) ‘Hawks and doves: The influence of nurse assessor stringency and leniency on pass grades in clinical skills assessments’, *Collegian,* 24, pp. 449-454.

Falchikov N (2005) *Improving assessment through student involvement*. London: Routledge- Falmer.

Kimbell, R., Wheeler, T., Miller, S. and Pollitt, A. (2007) *e-scape portfolio assessment: Phase 2 report*. London: Technology Education Research Unit, Goldsmiths, UL. Available at: http://www.gold.ac.uk/media/e-scape2.pdf. (Accessed: September 2017 and unavailable March 2018).

Kimbell, R., Wheeler, T., Stables, K., Sheppard, T., Martin, F., Davies, D., Pollitt, A. and Whitehouse G. (2009) *e-scape portfolio assessment: Phase 3 report*. London: Technology Education Research Unit, Goldsmiths, UL.Available at: https://www.teachertoolkit.co.uk/wp-content/uploads/2014/08/e-scape_phase3_report.pdf (Accessed: 30 March 2018).

Mortier, A., Lesterhuis, M., Vlerick, P. and De Maeyer, S (2015) ‘Comparative judgment within online assessment: exploring students feedback reactions’, *Communications in Computer and Information Science,* 571, pp. 69-79.

Nicol, D., Thomson, A., Breslin, C. (2014) ‘Rethinking feedback practices in higher education: a peer review perspective’, *Assessment and evaluation in higher education,* 39, pp.102-122.

Pollitt, A. (2012) ‘The method of Adaptive Comparative Judgement’, *Assessment in Education: Principles, Policy & Practice*, 19, pp. 281-300.

Thurstone, L.L. (1927) A law of comparative judgement. *Psychological Review*, 34, 273-286.

Tinto, V. (2003) Learning better together: The impact of learning communities on student success. Higher Education monograph series, 2003 - nhcuc.org. Available at: http://www.nhcuc.org/pdfs/Learning_Better_Together.pdf (Accessed 07 January 2019).

Tinto, V. (2008) Learning better together: The impact of learning communities on the persistence of low-income students. Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.472.8470 (Accessed: 07 January 2019).

Topping, K. (1998) ‘Peer assessment between students in colleges and universities’, *Review of Educational Research,* 68, pp.249-276.

Verhavert, S., De Maeyer, S., Donche, V. and Coertjens, L. (2018) ‘Scale Separation Reliability: What Does It Mean in the Context of Comparative Judgment?’ Applied Psychological Measurement 42, pp. 428-445.