Five go marking an exam question: the use of Adaptive Comparative Judgement to manage subjective bias

Practitioner Research in Higher Education
Special Assessment Issue

University of Cumbria
Vol 11(1) pages 94-100

Jill Barber
University of Manchester

Abstract
Adaptive Comparative Judgement (ACJ) is an alternative to conventional marking in which the assessor (judge) merely compares two answers and chooses a winner. (Scripts are typically uploaded to the CompareAssess interface as pdf files and are presented side-by-side.) Repeated comparisons and application of the sorting algorithm leads to scripts sorted in order of merit. Boundaries are determined by separate review of scripts.

A small pilot of ACJ in the fourth year of the Manchester Pharmacy programme is described. Twelve judges used ACJ to mark 64 scripts previously marked conventionally. 50 students peer-marked their own mock examination question using ACJ.

Peer-marking was successful with students learning from the process, and delivering both marks and feedback within two weeks. There was very good consistency among the students acting as judges, and accuracy (as defined by Pollitt, 2012) of 0.94.

Staff were similarly consistent, but the agreement with marks obtained by conventional marking was disappointing. While some discrepancies could be attributed to conventional marking failing under the stress of marking during teaching term, the worst discrepancies appeared to originate from inadequate judging criteria.

We conclude that ACJ is a very promising method, especially for peer assessment, but that judging criteria require very careful consideration.

Keywords
Adaptive Comparative Judgement; Peer Assessment; Summative Assessment; Marking Schemes; Hawks and Doves.

Introduction
Comparative judgement can be used to manage subjectivity in assessment, leading to demonstrable fairness in the marking of open-ended questions, which are not easily described by detailed marking schemes. 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.

The most rigorous system of comparison will involve every script being compared with every other script (a round robin), yielding an eye-watering amount of assessment for large classes. Numerous sorting algorithms have been designed in different contexts for reducing the number of comparisons required to sort items (such as scripts) into order (Moore et al., No date). In the case of judging student work, we require not only that the scripts are sorted into order, but (actually more importantly) that the separation between the scripts is accurately reflected, so that closely similar

Citation
Barber, J. (2018) 'Five go marking an exam question: the use of adaptive comparative judgement to manage subjective bias', *practitioner Research in Higher Education Journal*, 11(1), pp. 94-100.

scripts are given the same (or very similar) marks and genuinely outstanding or appalling scripts are marked appropriately. Adaptive Comparative Judgement reduces the number of pairs of scripts to be compared, typically to 9-12 complete rounds. It is mathematically complex but has some features in common with a Swiss-system tournament (Scimia, 2018), commonly used in weekend chess tournaments, and designed to yield a points score for every competitor.

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. This level of reliability has been disputed, notably by Bramley (2015), but even the most critical observers do not dispute that the selection of a winner from two pieces of work is much easier than determining an exact mark for any given work.

The motivation for trialling Adaptive Comparative Judgement in the Manchester Pharmacy programme had less to do with accuracy and more to do with other perceived advantages. The most important of these were:

- To achieve reasonable turnaround in assessment and feedback for large classes often requires that several staff members are involved. How do we ensure that marking is consistent?
- To allow students to show off their own thoughts and research requires open-ended examination questions incompatible with rigid marking schemes. How do we manage the inevitable subjectivity when marking this work?
- Students can learn a great deal from conducting peer assessments, yet are very reluctant to assign marks to other students' work. Can we use Adaptive Comparative Judgement to make Peer Assessment more meaningful?

A small one-year trial in the Manchester Pharmacy Programme is presented here.

Methods

The unit of the curriculum chosen for this case study was the "Global Health" subunit of the final year Pharmacy unit called "The Medicine". Students choose two subunits to make up the 30-credit unit; these subunits are examined separately, with the examination making up 50% of the available marks. Coursework, in which the two subunits are integrated, makes up the remaining 50%. In the examination, 25 marks for Global Health are assigned over one hour, made up from one short answer question (5 marks) and two short essays (10 marks each). The Global Health subunit is oversubscribed, with a typical cohort of 65-70 students.

The course objectives (learning outcomes) are as follows:

- You will understand the main causes of death worldwide.
- You will understand how poverty and social issues influence the causes of death worldwide.
- You will develop a detailed understanding of therapies and prevention for two of the major life-threatening diseases worldwide, and one non-disease cause of death.

The causes of death considered include poverty, diarrhoeal disease, pneumonia, cancer, suicide, road traffic accidents and several infectious diseases. The10 mark questions are designed to be very flexible, to allow students to focus on the causes of death that most interest them.

[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).

Trial assessment by staff

Initial trials of Adaptive Comparative Judgement focussed on a question set for the 2015 cohort of final year students:

> Imagine you have obtained a grant from a charity of £10 million to be spent within five years. There are no restrictions on your use of the money, except that you must save as many lives as possible. Outline how you would use the money. (a) Specify the disease or other cause of death you wish to address and why. (2 marks) (b) Describe the intervention you would make, including details of where you spend the money, what you would spend it on and why that is important. (5 marks) (c) Justify your costs. (3 marks) [10 marks in total]

This was tackled by 64 students and marked initially using a conventional mark scheme as indicated, by a single academic who was familiar with the taught content. The same academic has for many years been posting "All Student Feedback" in which marked student answers are presented in the form of a spreadsheet after the removal of all identifiers (Ellis and Barber, 2016) and has received no complaints about fairness in marking.

The mark scheme was adapted for the purposes of Adaptive Comparative Judgement. Judging criteria were:

- The disease or other cause of death has been defined, together with the reason for wanting to address it.
- The intervention has been described, together with what the money will be spent on.
- The costs have been justified.

Student answers, in pdf or Word format, were assigned unique, anonymous IDs and uploaded to the CompareAssess interface. On this occasion the upload was completed by Digital Assess staff, but it is now possible to upload scripts in-house. The mechanics of marking are described on the Digital Assess website (Using CompareAssess to support Formative Assessment and Using CompareAssess to support Summative Assessment).

The judgement statements were applied sequentially by 12 judges. Thus, if one script of a pair defined the problem and other did not, the one that defined the problem won, irrespective of other factors. If the three judging criteria failed to produce a result, other factors considered appropriate by the judge (spelling and grammar are examples) were used as tie-breakers.

Peer assessment

A similar question was used in a mock examination in 2016. This was taken by 51 students:

> 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. (10 marks: description of the problem 2 marks, description of action, in detail, 5 marks, costing 3 marks).
>
> Please address the question. There are no marks for brain dumps"

All but one of the students participated in judging their answers, and they also answered a short questionnaire about the judging process. It was a requirement that the students leave a sentence of feedback for each script that they judged.

Results

12 judges took part in the trial assessment of the initial 64 scripts. Most were academic staff members teaching on the Pharmacy programme, although only one taught on Global Health. There was one member of the Professional Support Staff and one student. Figure 1 shows the results of ranking the 64 scripts after 8 rounds of judging. The marks were derived from the rank order by marking four scripts (top, bottom and two close to the middle) manually and deriving other marks by scaling using the difference between scripts estimated by the software. It can be seen that after 8 rounds of judging the standard errors are still significant – in the worst cases a mark is accurate only to ±0.5. The software vendors (Digital Assess Ltd) recommend 12 rounds of judging, but 10 rounds might well be sufficient to provide acceptable marking accuracy for each script. The overall reliability (see Introduction) was 0.95.

It is possible to identify (and remove the judgements of) misfit judges, which can act as a deterrent to students tempted to judge randomly in peer assessments. No judges were flagged as misfits in this exercise. The student and member of the professional support staff were very close to the norm for the group.

Estimates of the time taken to judge a batch of scripts are currently quite informal and relate to the experiences of judges who are new to Adaptive Comparative Judgement. Staff believe that ACJ is time-neutral or takes a little longer than conventional marking (up to a factor of 2).

The shock came when comparing the results obtained by Adaptive Comparative Judgement to those obtained using a conventional marking scheme (see Figure 2).

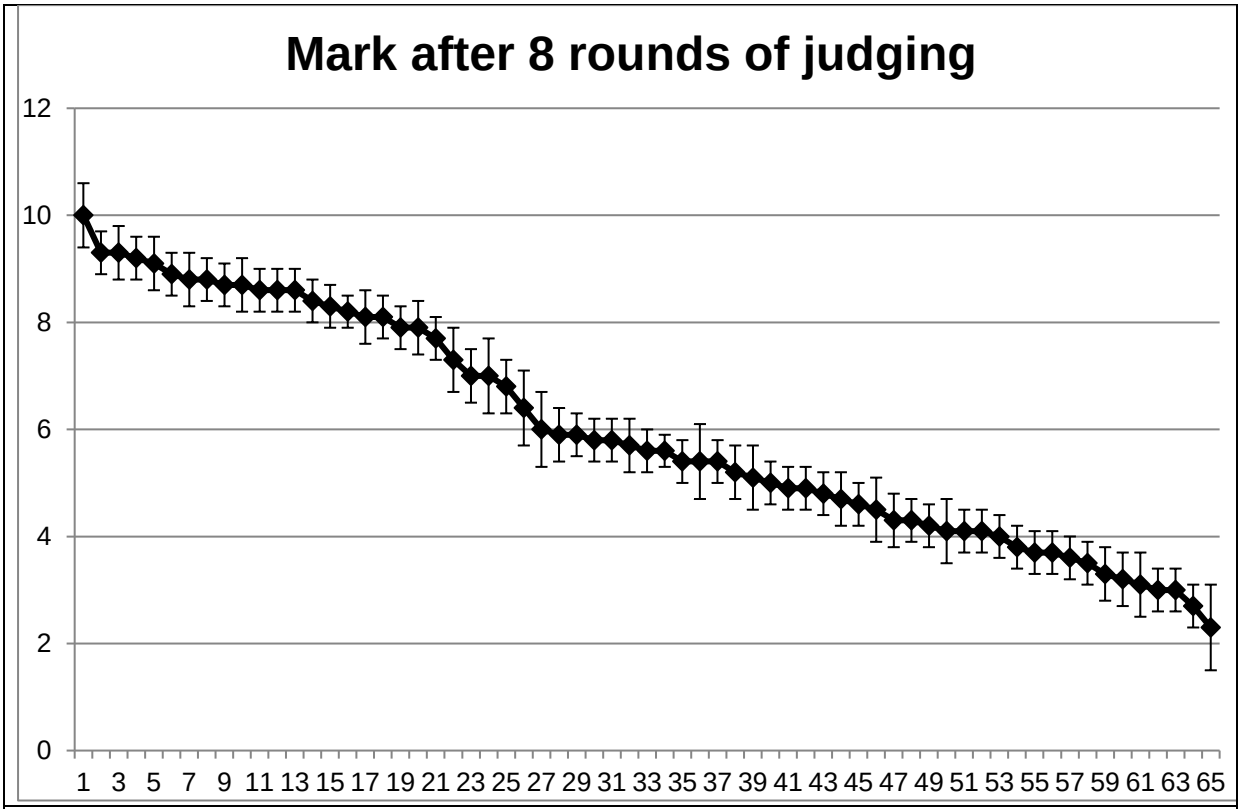

Figure 1. Mark out of 10 plotted against script number for 64 scripts after 8 rounds of judging using Adaptive Comparative Judgement. Error bars represent standard errors.

While the trend line in Figure 2 shows that there was a correlation between the marks obtained by the two methods, it was very far from perfect.

Inspection of the worst outlier scripts revealed that the worst discrepancies between the marking methods arose in large part from differences in the marking schemes. It was difficult to mirror a conventional marking scheme in Adaptive Comparative Judgement. So one script rather poorly described the problem to be addressed but then outlined a good programme of work, yielding a poor mark by Adaptive Comparative Judgement where defining the problem was of overriding importance, but a much better mark by conventional marking. Other factors contributing to the discrepancies included a prevailing culture within the subunit where originality, problem solving and being prepared to get involved were highly rated; these had not translated into the ACJ marking scheme.

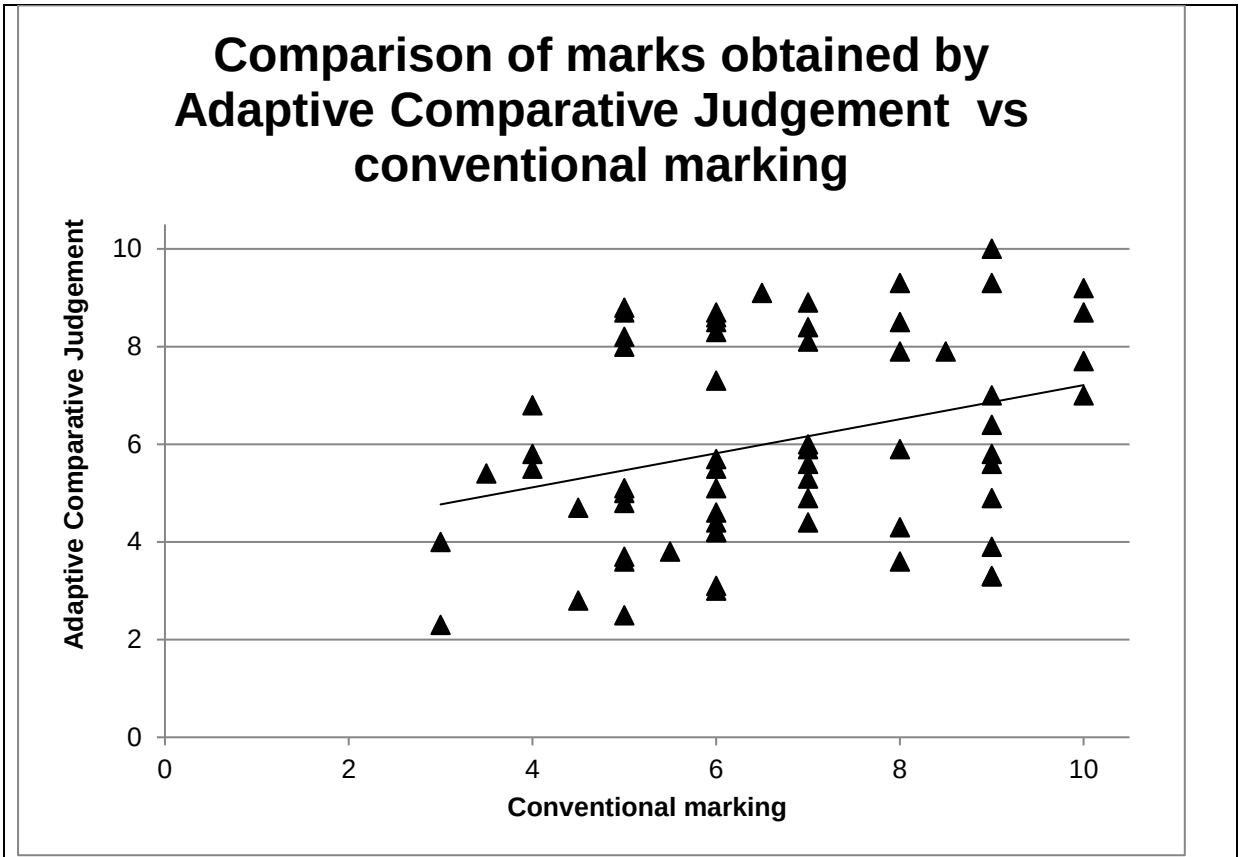

Figure 2. Marks obtained by ACJ and by conventional marking for the same set of scripts.

Peer Assessment

In this exercise, students were required to carry out 9 comparisons each, after which they had obtained an overall accuracy of 0.94, which is very close to that achieved by staff. A short questionnaire about the process was answered by the students, and the results obtained are as shown in Table 1.

The results shown in Table 1 were generally very encouraging, with students learning from reviewing their peers' work. In this preliminary exercise their opinions of the value of their peers' feedback was not sought. Most students were content that the software was easy to use and that the process was fair.

Objectively, the students received their marks and feedback in a timely way (about two weeks after the mock examination) and this would not have been achievable with the course leader marking all the scripts.

Table 1. Results of questionnaire sent to students concerning the Adaptive Comparative Judgement process.

Question	Summary of answers (n=50)
Ease of use (compared with Turnitin Grademark[2])	ACJ easier (28), more difficult (2), similar (20)
Useful (compared with conventional revision)	Useful (37) quite useful (8) less useful than revision (5)
Fairness (were you convinced the marking was fair?)	Yes (27), probably but a bit uneasy (19) no (4)
Nine judgements per student were required to get a result. Was this number OK?	Yes, I was still learning from later judgements (25), prefer fewer (19) last few are a waste of time (6)
How many exercises would be appropriate per semester?	Two or three during Reading Week (27) was the most popular response.
[2]Students were familiar with Turnitin Grademark, so it served as a useful comparator.	

Discussion

The pilot study of Adaptive Comparative Judgement within the Global Health unit was generally successful and a full licence has been obtained for the next academic year.

The method is especially attractive for assessments in which there is an element of subjectivity in marking, and allows academics to set rather open-ended questions without the tyranny of detailed model answers. In the case described here, students can write about cancer, suicide, diarrhoeal disease or road traffic accidents in response to the same question and a very flexible marking scheme is required.

The Global Health unit is oversubscribed and is examined in January. Marking time is typically two weeks while teaching is in progress, requiring that marking is shared among staff. Adaptive Comparative Judgement removes concerns about "hawks" and "doves" (Daly et al., 2017), the subjective bias that can arise when different staff mark different scripts.

Perhaps the most unqualified success is in peer marking. Students benefit from both giving and receiving peer feedback (Nicol et al 2014). Peer feedback is often written in more accessible language than feedback from academic staff and where (as in the present study) a piece of work is reviewed by many peers, students also benefit from the sheer quantity and variety of feedback (Topping, 1998, Falchikov, 2005). The appreciation that students also benefit from giving feedback is more recent (Cho and MacArthur 2011) and is strongly endorsed by this study. Bloxham and West (2004) additionally point to the need for students to understand the assessment process, and this can be facilitated by peer review exercises. Our own experience, however, is that students are reluctant directly to assign marks to their peers, although the mark is perceived as an essential part of the feedback. In Adaptive Comparative Judgement, marks are derived from a series of tasks that students are prepared to do – making pairwise comparisons.

In this pilot, it was clear, however, that the criteria on which scripts are judged need to be very carefully formulated. In the initial exercise, the judges delivered a very precise set of marks, which did not correlate very well with the marks obtained by traditional marking. In many cases, the Adaptive Comparative Judgement marks were reviewed and found to be better than the original marks, obtained in the stressful January marking period. Some of the most marked outliers, however, resulted from inadequacy of the judging criteria. These need to be reviewed carefully to ensure that they genuinely test the required learning.

Conclusion

Adaptive Comparative Judgement is a very promising method for marking open-ended student work. It would be helpful if novel applications of the method were widely shared to enable practitioners to optimise its use.

Acknowledgements

I thank Matt Wingfield and Chris Chambers (Digital Assess) and Steve Ellis (University of Manchester) for their vital collaboration in starting up this project. I also thank many colleagues and students at the University of Manchester for taking part as judges.

References

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).

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.

Ellis, S. and Barber, J. (2016) 'Expanding and personalising feedback in online assessment: A case study in a school of pharmacy', *Practitioner Research in Higher Education, Special Assessment Issue,* 10(1), pp. 121-129.

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).

Moore, K., Pilling, G., Khim, J. (No date) Sorting Algorithms. Brilliant.org. Available at: https://brilliant.org/wiki/sorting-algorithms/ (Accessed: 29 March 2018).

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.

Scimia, E. (2018) What is the Swiss System? Available at: https://www.thespruce.com/the-swiss-system-611537 (Accessed: 29 March 2018).

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

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