Mapping University Mathematics Assessment Practices

Edited by

Paola Iannone
University of East Anglia

Adrian Simpson
Durham University

Chapter 17
Summative Peer Assessment of Undergraduate Calculus using Adaptive Comparative Judgement

Ian Jones and Lara Alcock

Abstract Adaptive Comparative Judgement (ACJ) is a method for assessing evidence of student learning that is based on expert judgement rather than mark schemes. Assessors are presented with pairs of students' work and asked to decide, for each pair, which student has demonstrated the greater proficiency in the domain of interest. The outcomes of many pairings are then used to construct a scaled rank order of students. Two aspects of ACJ are of interest here: it is well suited to assessing creativity and sustained reasoning, and has potential as a peer-assessment tool. We tested ACJ for the case of summative assessment of first year undergraduates' conceptual understanding of a specially designed calculus question. We report on the relative performance of peer and expert groups of assessors, and the features of student work that appear to have influenced them. We consider the implications of our findings for assessment innovation in undergraduate mathematics.

17.1 Introduction

This project involved implementing and evaluating an innovative assessment of undergraduate calculus students' conceptual understanding of properties of two-variable functions. The innovation replaced a traditional computer-based test and contributed 5% of each student's grade for a first year calculus module. It comprised two parts. First, students completed a written test designed to assess conceptual understanding that was specially designed for the innovation, shown in Figure 17.1. Second, students assessed other's responses to the test online using an Adaptive Comparative Judgement (ACJ) method.

ACJ is an approach to assessing student learning that is based on holistic judgements of work rather than aggregated item scores (Pollitt, 2012). As such it offers promise for assessing conceptual understanding and for use as a peer assessment tool. It has been demonstrated to be effective in a variety of settings, from technol-

Ian Jones, e-mail: i.jones@lboro.ac.uk
Lara Alcock, e-mail: l.j.alcock@lboro.ac.uk
Mathematics Education Centre
Loughborough University

Conceptual Test Question

Consider the function $f : \mathbb{R} \to \mathbb{R}$ given by:

$$f(x,y) = \begin{cases} 0 & \text{if } x < 0 \\ x^2 & \text{if } x \geq 0 \text{ and } y \geq 0 \\ -x & \text{if } x \geq 0 \text{ and } y < 0 \end{cases}$$

Describe the properties of this function in terms of limits, continuity and partial derivatives. You should explain and justify your answers, and you may do so both formally and informally, using any combination of words, symbols and diagrams.

Fig. 17.1 Written test question designed to assess conceptual understanding

ogy teacher training (Seery, Canty and Phelan, 2011) to GCSE mathematics (Jones, Swan and Pollitt, in progress).

ACJ is derived from a well-established psychophysical principle (Thurstone, 1927) that people are far more reliable when comparing one thing with another than when making absolute judgements. Assessors are presented with pairs of scripts and asked to decide which student is the more able mathematician. The judgements of many such pairings are then used to construct a final rank order. This is usually done using a Rasch model which produces residuals for each judgement, thereby allowing the linearity and coherence of the final rank order to be explored in detail.

Until recently comparative judgement was not viable for educational assessment because it is tedious and inefficient. The number of required judgements for producing a rank order of n scripts would be $\frac{(n^2-n)}{2}$, meaning that for the 168 scripts considered here, just over 14000 judgements would be needed. However the development of an adaptive algorithm for intelligently pairing scripts means the number of required judgements has been slashed from $\frac{(n^2-n)}{2}$ to $5n$, so that 168 scripts now require only 840 judgements.

17.2 Implementation

17.2.1 Test design and administration

The written test was developed by the course lecturer (the second author) specially for this project. We considered various practicalities when deciding on the precise test structure and administration. First, timing of the test in relation to the course meant that students had been provided with definition-based lectures and exercises related to the concepts of limits, continuity and partial derivatives for functions of two variables, but that they had done only minimal work on the last of these. Second, we wanted to ask a question that would prompt students to think more deeply about these concepts, which are known to challenge students in different ways and to different extents (Pinto and Tall, 2001), than would routine exercises or even variants

of routine proofs: such routine work is required in other tests within the module. Third, we wanted an individual written test in order to fit with the requirements of the ACJ system but, because it would replace an online test, we did not want something that would take up a lot of lecture time.

As a result, we decided to set the test question given in Figure 17.1, which we hoped would allow students considerable flexibility in choosing how to respond, and which would prompt them to think about whether and how concepts from the course applied in a non-standard situation. In order to encourage this thinking without taking up excessive lecture time, we distributed copies of the question to the students six days in advance of the lecture in which the written test was to take place. The test was administered in a lecture under exam conditions: students were allowed 15 minutes to complete the test and were told that their answer must fit on the single side of A4 paper as provided.

17.2.2 Peer use of ACJ

33 students opted out of their scripts being used for research purposes and we discuss only the remaining 168 scripts in the report. The scripts were anonymised by removing the cover sheet, and then scanned and uploaded via a secure file transfer protocol to the ACJ website[1].

The day after the written test, a researcher explained the paired judgements activity and demonstrated the ACJ website to the students in a lecture. The researcher told the students that they would log in and be presented with 20 pairs of scripts, and that they should decide, for each pair, which author had demonstrated the better conceptual understanding of the question. A screenshot of the user interface is shown in Figure 17.2. He advised them that each judgement should take on average around three minutes and that the total work should take no more than one hour.

A user guide was provided on the course VLE page to support students with technical aspects of ACJ, and drop-in support sessions were offered in a computer lab during the exercise. In practice, no technical problems were reported and the only help requested were password reminders.

17.2.3 Rank order construction

Once the students had completed the online judging we constructed a rank order of scripts by fitting the judgements to a Rasch model (Bond and Fox, 2007). The outcome of the Rasch analysis was a scaled rank order. Each script was assigned a parameter value and standard error along a logistic curve. The final rank order of scripts produced by the students is shown in Figure 17.3.

[1] The ACJ website is called “e-scape” and is owned and managed by TAG Developments, the e-assessment division of Sherston Software.

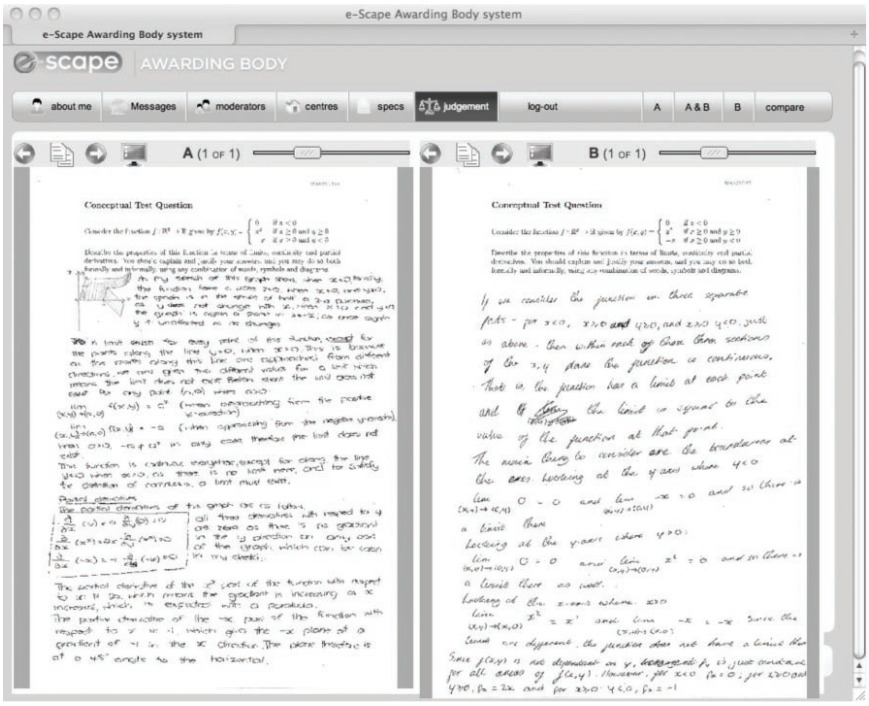

Fig. 17.2 The "e-scape" system's ACJ user interface.

Rasch analysis produces a host of measures that can be used to explore the stability of the rank order. A key measure is the internal consistency, analogous to Cronbach's α, which can be considered the extent to which the students' judgements are consistent with one another. The internal consistency of the students' rank order was .91, an acceptably high figure.

17.2.4 Allocation of grades

A rank order produced by ACJ can be used to allocate grades to students in the standard way. This can be done using norm referencing, for example, allocating the top 20% of scripts a grade 'A' and so on. Alternatively it can be done using criterion referencing. This requires sampling scripts from across the rank order and comparing them against agreed assessment criteria. Boundary scripts within the rank order can then be identified and grades applied accordingly. In our case the students will be eventually be awarded grades using criterion referencing, but that process was not within the scope of this project.

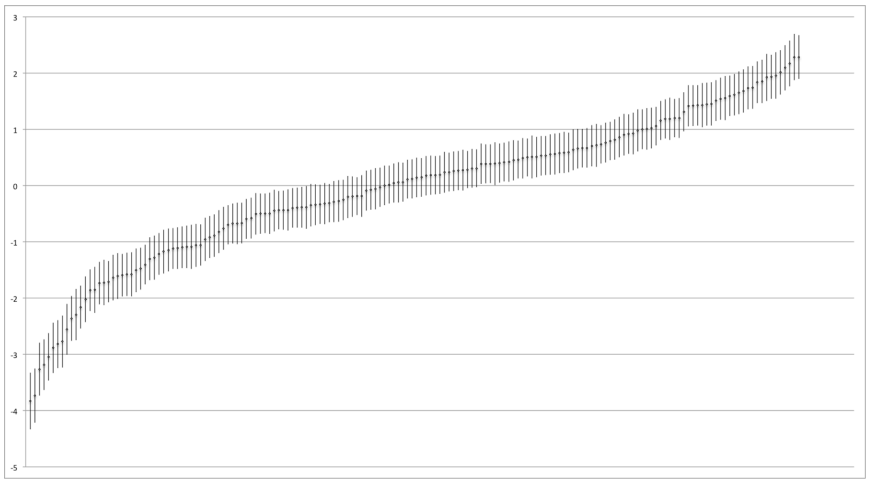

Fig. 17.3 Scaled rank order of student scripts. The horizontal axis shows the 168 scripts from "worst" to "best". The vertical axis is the scripts' parameter values in logits. The standard error of each parameter is also shown.

17.3 Evaluation

We intended to use the students' peer assessment for summative purposes and it was therefore necessary to thoroughly evaluate the process. We undertook a statistical analysis in order to evaluate the consistency and reliability of the rank order of scripts. We also interviewed and surveyed students – and other participants as introduced below – in order to establish which features of scripts influenced them when undertaking pairwise comparisons.

17.3.1 Statistical analysis

To evaluate the students' performance we correlated the rank order they produced with rank orders of the scripts produced by two further groups of participants. One of the groups comprised nine experts (mathematics PhD students) and the other comprised nine novices (social science PhD students with no mathematics qualifications beyond GCSE or equivalent).

The expert group provided a benchmark against which to compare the students' performance. It was expected the expert and student rank orders would correlate very strongly. The novice group provided a control. The participants in the novice group had never studied any advanced mathematics and would thus not be able to use mathematical understanding when making judgements. It was therefore expected the expert and novice groups would correlate weakly at best.

The participants were paid for their time and the procedure was the same for both the expert and novice groups, except for a preparatory activity. The experts were sent the written test and asked to become familiar with it by completing it themselves. The novices, presumably unable to complete the test, were instead sent three student written responses. The novices were asked to inspect the three responses and rank them, as well as they were able, in terms of the students' conceptual understanding of the test question.

Each group then attended a training session lasting 30 minutes. During the training sessions a researcher explained the rationale and theory of ACJ, and demonstrated the "e-scape" website. Two expert participants were unable to attend the workshop and received individualised training instead. The participants practised judging scripts online. Once familiar with the website they were each allocated 94 judgements to be completed within ten days of the training.

Once the judging was complete, the judgements for each group were fitted to a Rasch model. The internal consistency was acceptably high for both the expert group (.97) and the novice group (.99).

17.3.2 Analysis and results of statistical analysis

Spearman's rank correlation coefficients were calculated for the three pairs of rank orders. The outcomes are shown in Table 17.1.

	Peer	Novice
Expert	.628	.546
Novice	.666	

Table 17.1 Spearman rank correlation coefficients between the rank orders produced by the students and the two groups of participants. All correlations are significant at $p < .001$.

The expert and peer rank orders correlated significantly more strongly then the expert and novice rank orders, $Z = 1.670$, $p = .048$. This suggests that the experts and peers were more in agreement with one another about what constitutes a good answer to the question than were the experts and novices. Nevertheless, the significance was marginal and we had anticipated a much more marked difference. We also expected the novice group to correlate much more weakly than it did with either the peer group or the expert group. The relatively strong correlation between the novice and two other groups leads to the counter-intuitive and unexpected conclusion that novices lacking knowledge of advanced mathematics can, to some extent at least, assess understanding of advanced mathematics. Furthermore, it is surprising that the peer and novice rank orders correlate more strongly than do the peer and expert rank orders, albeit this difference falls short of significance, $Z = -.7350$, $p = .231$. Reasons for these unanticipated results are considered later in the report.

17.3.3 Survey

Once the judgement week was complete, the students on the course were sent an email inviting them to complete a short online survey about their experience of completing the judgements (they were informed that two randomly-selected students who completed the survey would each win a book token worth £20). Twenty-five students completed the survey. The same survey was also completed by seven of the expert judges and all nine of the novice judges.

The survey instrument comprised nine items which judges rated using a three point nominal scale. The items were derived from the literature into examiner marking and grading processes (e.g. Crisp, 2010) as well as in consideration of contrasts across the students' responses to the written test. The items were worded as generically as possible such that the instrument could be calibrated for use in future ACJ studies using different test questions and, possibly, in different disciplines. For each item judges were asked to consider whether a criterion had a negative, neutral or positive influence on how they made their decisions when judging a pair of written tests. The nine items are shown in Figure 17.4. The instrument also contained an open response section.

17.3.4 Analysis and results of survey

The results from the students' and participants' responses to the nine items are shown in Figure 17.4. There was no difference between the three groups' mean scores, $F(5,35) = .931$, $p = .473$ and so the groups' responses are combined in Figure 17.4.

Item 5, which asked whether use of colour was influential when judging scripts, was intended as a control item and indeed most responses were "no influence". Items 6 and 8 also addressed surface features, although most respondents were negatively influenced by untidiness. The use of written sentences (item 1), formal notation (item 4), diagrams (item 7) and structure (item 9) were all considered largely positive influences. We were slightly surprised by the uniformity of responses to these items, expecting individual differences such as a preference for formal notation over written sentences. The presence of errors (item 3) was negatively influential and evidence of originality and flair were positive (item 2), as might be expected.

Items 2 and 3 are perhaps the only two that novices were unable to use due to their lack of knowledge of advanced mathematics. This means novices were in fact able to recognise other features when making judgements. This may in part explain why their rank order correlated more strongly than expected with that of the students and experts.

An optional open question asked respondents to "state any other features you think may have influenced you when judging pairs of scripts". The responses revealed three influential features not included in the nine items: completeness (e.g. "whether all parts of the question were answered"), factual recall (e.g. "display

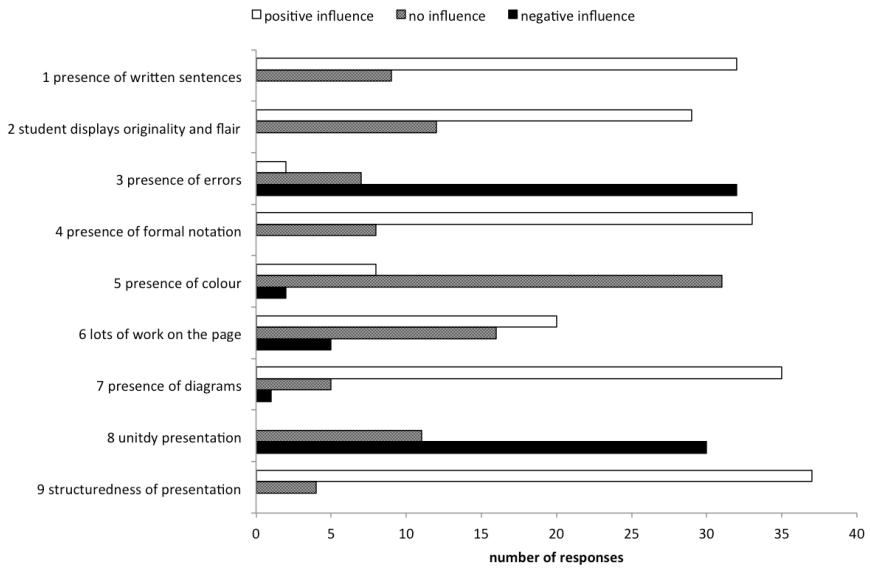

Fig. 17.4 Student and participant responses to the nine items in the online survey.

of knowledge of basic definitions") and vocabulary (e.g. "key words such as flat, smooth, cut had a positive influence"). These items will be included in future adaptations of the instrument.

A second optional open question asked respondents to "comment on your overall experience and feelings about the computer-based part of the conceptual test". Analysis is ongoing but we note here three concerns raised by students. One was that the resolution of the scripts on the screen was too poor to read them properly. Such students presumably did not notice or use the resolution toggle button which overcomes this problem. This feature was demonstrated to the students and highlighted in a support email, and we do not know how many students failed to use it.

Another concern expressed was that not all peers took the activity seriously. One student said, "I do feel that some people may not have to judged the tests accurately as it made no difference to there (*sic*) work. I do understand students should do, however speaking to various students may not have spent the correct time on the computer-based part of the test." This is an astute comment as the quality of the students' judgements had no effect on their final grade. The problem of ensuring undergraduates are properly motivated when assessing one another has been raised in the peer-assessment literature (Topping, 2003), and we return to it later.

Some students commented on the poor quality of some answers, and questioned their peers' ability to assess advanced mathematics. For example, "at least half of the scripts which I read said that the graph was continuous everywhere, when it wasn't. What concerns me is that those people who believe that the graph was continuous everywhere would most probably be marking my own answer wrong." The ability

of the students to assess the test can be addressed by statistical analyses, and we discuss further work in this direction below.

17.3.5 Interviews

Semi-structured interviews were conducted with samples from each group of judges. In total nine students, seven experts and three novices were interviewed. Each interview lasted about 20 minutes and was audio recorded and transcribed.

In the interview, the researcher first presented the judge with three pairs of scripts on laminated card. The judge was asked to decide, for each pair of scripts, which was the better in terms of conceptual understanding of the question. They were also asked to give a confidence rating for their decisions on a three-point scale (not at all confident, somewhat confident, very confident). The researcher then asked the participant to talk about each of their decisions in turn using the following three prompt questions:

- How did you decide which test showed the better conceptual understanding?
- Did anything else influence your decision?
- Any other comments about this pair of tests?

Just before the end of each interview the researcher also asked, "How did you find the experience overall?"

17.3.6 Analysis and results of interviews

To analyse the interviewees' judgements of the three pairs of scripts we first identified for each pair which script was the "best" based on an independent expert rank order (see below). This enabled us to designate every judgement made in the interviews as correct (i.e., consistent with the expert rank order) or incorrect. The confidence rating for each correct judgement was scored 1 (not at all confident), 2 (somewhat confident) or 3 (very confident), and conversely each incorrect judgement was scored -1, -2 or -3. We then calculated a weighted score for each interviewee by summing their confidence ratings across the three pairs of scripts. The mean weighted scores across the three groups were 2.14 for the expert group ($N = 7$), -0.44 for the student group ($N = 9$), -0.33 for the novice group ($N = 3$). The experts were the only group to score positively while the students and novices scores were close to zero. This suggests the experts were better able than the peers or novices to judge the scripts, although the small number of participants means we cannot claim statistical significance.

Analysis of responses to the three follow up questions is ongoing and will help us to understand the cognitive processes involved in deciding which of two scripts is the better. Early analysis suggests, perhaps unsurprisingly, that experts, and to

an extent peers, focused on mathematical correctness and understanding, whereas novices focused on surface features. To illustrate this, the following responses from each group to scripts A and B are representative:

> *Expert:* First "B" provided more explanation to the answer. "A" just said it is continuous when $x \geq 0$. But "B" said more exactly on the line where $x > 0$ and $y = 0$. And on this line, the function is not continuous and does not have a partial derivative, so I think it confirms "B" is better. And the reason is ok, and also I think "B" said the partial derivative does not exist on the function where it is not continuous.

> *Peer:* It was quite hard as they are similar. They have got a lot of the same information on them. The partial derivatives for "A", she said are all 0, and "B" says they don't exist. So I agree with "A". I think they exist.

> *Novice:* It was very tight. I am not really confident about this one. But I prefer the way they table the answer in "A" in terms of all elements of the question were approached, they set up the limits, and the bit about continuity, and they got to the partial derivative in a logical order to me. "B" had very nice graphs - although one graph had nothing on. It did not seem as coherent to me.

We note that students' responses to the final question, "How did you find the experience overall?", suggest that they found judging peers' scripts challenging, but beneficial for learning. For example:

> It is hard to judge other people's work ... Sometimes we as students, we think we understand, but we have to make sure that if someone else reads who has no clue what the concept is, by looking at the question they should be convinced it answers the question. So it is important to write in a good way. It is an improvement for me for my future writing.

17.4 Discussion

In practical terms, the implementation of this novel assessment approach was a success. The scanning and delivery of the scripts to the e-scape system was unproblematic, and no-one in any of the judging groups reported any technical barriers to using the system. All those students who engaged with both parts of the test thus had the opportunity to formulate their own answer to a conceptual question, and to consider the relative merits of responses provided by their peers. Participation was acceptably high - numbers completing both parts of the test were comparable to what would be expected for any other in-class or online test for this amount of credit. In this sense, the goals of the project were successfully achieved.

In theoretical terms, the picture is more mixed. The correlations in Table 17.1, while in the expected direction and statistically significant, are not as anticipated. We expected the correlation between the peer and expert groups to be very strong ($> .9$) and the correlation between the novice and expert groups to be weak ($< .5$). The correlations are also at odds with the experts' superior performance when judging the scripts presented in the interviews, and with their mathematically more sophisticated explanations of how they made their decisions.

The crucial problem appears to have been that the software's adaptive algorithm may not have been optimal when pairing scripts. In other words, a technical glitch meant that the judgements were not informative enough for constructing stable rank orders, no matter how "correct" or internally consistent the judges' decisions. To explore this hypothesis the expert and novice groups are undertaking further judgements. Early analysis suggests the correlation between peers and experts will increase significantly.

Another reason for the relatively low correlation between peers and experts may be due to some students not taking the exercise seriously, or neglecting to adjust the website resolution, or finding the question too difficult to be able to judge the quality of others' answers. We discuss how we intend to address these issues in the next section.

The unexpected results presented us with an immediate practical problem. We had originally intended to use the peers' own judgements for assigning grades. However, the relatively weak correlation between the peer and expert groups caused us to decide not to do this. Instead an independent group of experts, made up of maths and maths education lecturers (and including the course lecturer), has re-judged the scripts and their rank order will be used for grading purposes.

17.5 Further work

Because of the innovative nature of this work, it is currently too early to specify whether and how adaptive comparative judgements will be used as an assessment system in this course or more broadly in the institution. The pressing work required is to test and if necessary improve the adaptive algorithm used to select which pairs of scripts to present to judges. On the basis of previous studies (Jones, Swan and Pollitt, in progress; Kimbell, 2011) we suspect that this alone may go far to improving the peer and expert correlations to acceptable levels. Once improved, we aim to repeat the exercise next academic year.

We will also seek to improve the students' performance by ensuring scripts always load clearly without need to adjust the resolution. Students will also be incentivised to take the exercise seriously by adjusting their grade based on their judging performance. One possibility is to compare their individual judgements with the scripts' positions in a rank order generated by experts. Their performance could then be used to adjust their grade according to agreed levels.

Something that will need to be carefully considered is the question used. On the one hand, the conceptual question used in this instance did challenge the students (many wrote things that were partially correct and partially incorrect), and the responses were very varied so that those students will have seen a wide range of response types. On the other hand, some of the independent experts thought, with hindsight, that this particular question had one problem in particular: the fact that it allowed the students freedom to answer in terms of three different properties meant that it was sometimes difficult to compare two scripts. For instance, how should

one compare one script that provides a clear diagram and a correct and well-argued response about the properties of limits and continuity but no information on partial derivatives, with one that has a similar diagram and information about all three properties but contains minor errors? In planning for future tests of a similar nature, we will give more advance consideration to issues of comparability on multiple dimensions.

Addressing these practical issues will also allow us to make further theoretical developments about the use of ACJ for assessing advanced mathematics. We consider the online survey to be the first step in developing a reliable instrument for evaluating the cognitive processes involved in judging. The items will be adapted and extended according to the results and qualitative feedback. We will also increase the rating scale from three to five points to enable more discriminatory responses.

Further ahead we will wish to explore in detail any potential learning benefits that can arise from a pairwise comparisons approach to peer assessment. Many students, and even some PhD maths experts, reported that they felt the exercise was beneficial for learning. A suitable instrument and method will need to be adapted or developed for future studies.

References

Bond, T.G. & Fox, C.M. (2007) *Applying the Rasch model: Fundamental measurement in the human sciences*. Abingdon: Routledge.

Crisp, V. (2010) Towards a model of the judgement processes involved in examination marking. *Oxford Review of Education*. 36(1), 1-21.

Jones, I., Swan, M. and Pollitt, A., (*in progess*) Adaptive Comparative Judgement for assessing sustained mathematical reasoning.

Kimbell, R. (2011) Evolving project e-scape for national assessment. *International Journal of Technology and Design Education*, [online first issue].

Pinto, M. and Tall, D. (2001) Following students' development in a traditional university classroom. In *Proceedings of the 25th Conference of the International Group for the Psychology of Mathematics Education*. Utrecht, The Netherlands: PME, 4, 57-64.

Pollitt, A. (2012) The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice*. [online first issue].

Seery, N., Canty, D. and Phelan, P. (2011) The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design Education*, [online first issue].

Thurstone, L.L. (1927) A law of comparative judgment. *Psychological Review*. 34(4), 273-286.

Topping, K. (2003) Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.) *Optimising New Modes of Assessment: In Search of Qualities and Standards*. Dordrecht: Kluwer Academic Publishers, 55-87.

Mapping University Mathematics Assessment Practices
Published 2012.
University of East Anglia
ISBN 978-1-870284-01-1