Comparative Judgement in Undergraduate Mathematics

i.hamilton

23 June 2021

1 Abstract

This document relates to comparative judgement used for the assessment of mathematics at a university level, especially with respect to peer assessment. It seeks to present the most relevant papers and their findings, as well as relevant personal conversations and some concluding remarks.

2 Introduction

Comparative Judgement is a method of educational assessment where instead of marks being awarded based on a marking scheme, items are assessed pairwise with a preference selected within each pair and then an aggregation of these pairwise judgements leading to a rating for each item, typically through the application of the Bradley-Terry model (which the education literature interprets as a conditional Rasch model (Andrich, 1978)). The key insight on which the method is often sold is that humans are generally better at comparative than absolute judgement (Laming, 2003). These forms of pairwise judgement began to be used in an educational setting about twenty years ago, initially as a means of comparing exams over time and exam board moderation rather than for assessing pieces of work (Bramley et al., 1998; Bramley, 2007). The method was then further promoted and popularised by Alastair Pollitt, then Director of Research at Cambridge Assessment, reflected in a number of publications (Pollitt, 2004, 2012a,b; Pollitt and Whitehouse, 2012). Over the last decade the method has garnered steadily more interest in the academic literature and in wider practice.

The largest current use case is in school-level, especially primary, education, where it is employed as an efficient means for schools to get a reliable and nationally consistent assessment of writing. While the name of the largest supplier of these services — `nomoremarking.com` — hints at the nature of the initial selling point of the method, a number of other benefits have become apparent. Being able to compare pairwise means that items can be compared to those from other cohorts and years in a reliable manner, so that robust estimations of the distribution of quality for any year group can be made, and a translation can be made for each piece of work to an equivalent achievement age. Additionally because little training is required for assessors, schools have found that they can employ a wider cohort of teachers in the assessment exercise, reducing the burden on any particular individuals and engaging a wider section of the teaching staff in the literacy of the students at any given level.

The use of comparative judgement for peer assessment is more recent. It relies on the same principle as in the school setting, that when seeking assessment reliability from a large number of judges of varying expertise, comparative judgement could be an effective approach. It also may benefit the judges themselves in exposing them to a variety of approaches, arguments, and fallacies that they must assess. In this way it is potentially preferable to other peer assessment approaches, which emphasise a clear and unambiguous marking scheme (Falchikov and Goldfinch, 2000), the application of which may provide less of a learning opportunity and be excessively burdensome to produce and train markers on. In a Mathematics context much of the research has been conducted by Ian Jones at the University of Loughborough and co-authors. The works described below investigate questions such as reliability, validity, absolute vs comparative judgement, and the nature of what judges value. The inclusion criteria are rather loose but I think all, or certainly most, works specifically involving comparative judgement of undergraduate mathematics are included, with a few additional items looking at either comparative judgment of mathematics or comparative judgement for undergraduate peer assessment. The length of each account is commensurate with how useful each is for our purpose.

A brief note on terminology. In what follows the terms *reliability* and *validity* will be used. Reliability refers to the degree to which under the same assessment protocol but with potentially different or permuted assessors a grade awarded to an item would be the same. Validity refers to the degree to which the grade attained corresponds to a quality of interest which the

assessment is seeking to measure. Hence an assessment can be reliable without being valid (e.g. if assessors were asked to count the number of words in the answers then this would be a reliable but invalid measure of quality) or valid without being reliable (e.g. if each student was asked to give an absolute mark out of 100 for the conceptual understanding demonstrated by the answer of one other student having been given very clear guidelines of what constituted conceptual understanding in that context).

3 Jones and Alcock (2012, 2014)

These two articles appear to discuss the same assessment exercise. The question is consistent and the exact number of students participating in the research $n = 168$ is the same, as is the number of students who took part in the assessment exercise but decided not to participate in the research (33). The takes are slightly different though and the number of assessors also changes between the two, suggesting that some additional assessments were done later. (it would be good to check this with the authors).

3.1 Method

In the assessment exercise, undergraduate calculus students were asked to respond to the following prompt.

> Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by:
>
> $$f(x,y) = \begin{cases} 0 & \text{if } x < 0 \\ x^2 & \text{if } x \geqslant 0 \text{ and } y \geqslant 0 \\ -x & \text{if } x \geqslant 0 \text{ and } y < 0 \end{cases}$$
>
> 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.[1]

The question was designed specifically for this exercise. It was given to the students six days before the assessment, along with some practical guidance and the suggestion that they might want to think about the following criteria:

[1] There appears to be a typo in Jones and Alcock (2012) which defines the function as $f : \mathbb{R} \to \mathbb{R}$.

- Are all the statements correct?
- Are all the statements written clearly?
- Are the explanations and justifications convincing?
- Is the overall answer comprehensive and coherent?
- If diagrams are used, are these accurate and well-labelled?
- If diagrams are used, is it clear how these are related to the text?
- Does the layout help a reader to understand the answer?

. It was then completed in 15 minutes of lecture time, under exam conditions. Students were instructed that answers must take no more than a single side of A4. The mark that they received contributed 5% of their overall module mark, and replaced an online test.

The day after the written test, students were introduced to the peer assessment platform during a lecture, and instructed that they should select the item that "had demonstrated the better conceptual understanding of the question." They were told that they would be presented with 20 pairs each and they should expect to spend about three minutes on each paired judgement, and that the total work should take no more than an hour. The criteria provided in guiding their preparation were not restated, either at the time of the actual assessment or in eliciting the pairwise judgements. These judgements were aggregated using the Bradley-Terry model, providing a rating for each script.[2] In this case the authors chose to convert this to a grade by criteria assessment — identifying a boundary script that was the first to sufficiently meet certain criteria such that higher ranked scripts would also attain at least this grade. Alternatively grades could be created on a normative basis e.g. by allocating top grades to the top 20%.

In order to assess the reliability of the peer marking, the scripts were also comparatively judged by two further groups — 'experts' and 'novices'. The expert group in Jones and Alcock (2012) consisted of nine Mathematics PhD students; in Jones and Alcock (2014) the group had been augmented by a further eleven academics. In both articles the novice group consisted of nine Social Sciences students with no mathematics qualifications beyond

[2]Due to a limitation of the platform, participants could not be prevented from potentially having their own scripts included in the assessments that they were shown. These instances were removed at the analysis stage. In all 29 instances the student ranked their own work as better!

GCSE level, though Jones and Alcock (2012) suggests that the novice group may have done some additional judgements that are included in the results of Jones and Alcock (2014) but not Jones and Alcock (2012), but this was not fully determinable from the descriptions (again, something to be checked with the authors). The experts were asked to familiarise themselves with the exercise by completing it themselves. Both groups attended a training session lasting 30 minutes to familiarise them with the platform, after which each judge was required to complete 94 judgements within the next ten days.

Jones and Alcock (2014) provide some additional analysis in order to assess reliability and validity. They randomly split both the student and expert groups into two and compare the ratings produced by the two subgroups with each other. This constitutes a check on validity when applied to the expert group in the sense that one way we might understand comparative judgement to be less valid is that expert judges may have a differing sense for what is important, and thus rate work differently. Thus if the assessment is reliable within the expert group then it might be thought to be valid.

Finally the exercise was subject to further analysis by means of a voluntary survey (with an associated book token prize), and semi-structured interviews with small samples from each of the judging groups. The survey was completed by 25 students, seven experts and all nine novices. The interviews were with nine students, seven experts and three novices. During the interviews, the researcher presented the interviewee with three pairs of scripts (on laminated card) and for each pair asked them to state which they considered was better, and how confident they were in that assessment, with these being coded $-3, -2, -1, 1, 2, 3$, with the sign being determined by whether the order was consistent with that of the expert judges in the overall assessment, and the absolute value with their confidence levels.[3]

3.2 Results

Jones and Alcock (2012) displays the main results in a table showing the

[3]As David Firth has suggested, while on the face of it this might seem like an appealing construct, perhaps we might better think of it as 2-dimensional in that one might for example think one item was a very small amount better but with a high degree of confidence e.g. suppose you had two pieces of work which were identical except for the use of one adjective, with one being clearly more appropriately descriptive and expressive, then a judge might be very confident that one was better than the other but the difference between them is marginal.

Spearman rank correlation coefficient for the three judging groups as shown in Table 1, with all differences being significant. It should be remembered

	Peer	Novice
Expert	.628	.546
Novice	.666	

Table 1: Spearman rank correlation coefficient between judging groups (Jones and Alcock, 2012)

that the expert judgement here was based on just the nine PhD judges.

In contrast, the results given in Jones and Alcock (2014) use Pearson correlation on the ratings, but do not provide a figure for peer-novice correlation as in Table 2. It should also be noted that Jones and Alcock (2014) use a

	Peer	Novice
Expert	.77	.64

Table 2: Pearson correlation between judging groups (Jones and Alcock, 2014)

different expert set in including 11 additional academics and in removing one of the PhD assessors based on their response times being only 10 seconds per pair.

Continuing with the report in Jones and Alcock (2014), sub-group reliability was high with Pearson correlation of .86 for the experts and .72 for the peers. Additionally it presents statistics for each of the individual sub-groups as presented in Table 3. These are based on statistics that have become standard in the comparative judgement literature and are themselves based on Rasch analysis. They are described in detail in Pollitt (2012b). There are grounds for treating these with scepticism when viewed in isolation (this is something that has been demonstrated in empirical work (Bramley, 2015), but is something I hope to look at more closely in my PhD).

The final piece of quantitative evidence around the validity of the peer assessment as reported in Jones and Alcock (2012) came from the interviews where the mean scores from the confidence exercise for expert (N=7), peer (N=9) and novice (N=3) groups were $2.14, -0.44, -0.33$ respectively.

Based on these results, Jones and Alcock (2012) explain that they had intended to use the peer assessments to attribute the module marks, but

	Peer1	Peer2	Expert1	Expert2	Novice
SSR	.73	.86	.93	.89	.97
Judges	100	93	11	11	9
Misfitting Judges	3	4	0	1	1
Misfitting scripts	6	6	7	4	6

Table 3: Comparative judgement statistics (Jones and Alcock, 2014)

decided instead to use expert judgements. The wording suggests that it was the judgements of the 11 academics that were used for the grading, though it is not explicitly stated. In Jones and Alcock (2014) it is simply asserted that expert judgements were used for the grading.

3.3 Discussion

The discussions in both articles provide further useful insights. As they note, it is important for students to have confidence in the method for it to be accepted as fair if it is to contribute to their grade. There were three concerns regarding this that came up in the feedback. First that the way the platform operated meant that some users may not have seen clear renderings of the work. Second that some fellow students (as evidenced by their own answers) may not have been reliably capable of assessing what constituted correct/incorrect in this context, and therefore their ability to judge superiority of an answer might be questionable. Third that some student assessors did not take the task seriously. All of these seem addressable. The first of these would not be an issue with current platforms. The second may be addressable through explanation of literature that suggests this not to be the case or even perhaps a practical example delivered in the lecture introducing the procedure, or by providing more guidance for the judging, perhaps by highlighting common mistakes. The third could potentially be addressable by making a portion of the marks contingent on their own assessments being sufficiently consistent with the overall assessment. This seems reasonable pedagogically also in that being able to identify superior answers may itself be viewed as a reasonable test of understanding. It might also address an issue that was not mentioned but seems plausible, that of up-ranking the work of friends.

Jones and Alcock (2012) calls out the algorithm by which the pairwise comparisons were scheduled as being sub-optimal. This may well be the

case (another potential topic of my PhD) but there does not seem to have been any explicit attempt to address this in Jones and Alcock (2014) or any subsequent work by the authors other than to employ more assessors in the expert group, and perhaps have the novice group do more judgements. It is also worth noting that `nomoremarking.com`, the largest comparative judgement platform in the U.K. (and possibly globally), has reverted to using random assignment for the pairwise scheduling (as per personal conversation with Chris Wheadon).

The question/prompt used was also highlighted by some expert judges as something worth further consideration. While it was successful in the sense that it elicited a wide variety of responses many with both correct and incorrect parts, the fact that it allowed the students to answer with respect to three distinct questions was seen as something that made assessing answers more challenging. For example, as they ask "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?" Whether this is a bug or a feature of the method may be debated, but it may be possible to mitigate the extent of this effect through a more carefully considered question.

Finally and importantly there was some evidence in the feedback that students had found the exercise challenging but beneficial for learning.

4 Jones and Wheadon (2015)

This paper, which seems to build on Jones et al. (2013) in using the same question but with a larger sample, seeks to look directly at the degree to which comparisons, rather than global criteria, contribute to the effectiveness of the method by having a group of 13 to 15 year olds respond to the question

> Write down these fractions in order of size from smallest to largest.
> Underneath describe and explain your method for doing this.
>
> $$\frac{3}{4} \quad \frac{3}{8} \quad \frac{2}{5} \quad \frac{8}{10} \quad \frac{1}{4} \quad \frac{1}{25} \quad \frac{1}{8}$$

These answers were then comparatively and absolutely judged by both the students themselves and experts, and the results compared. As expected

the absolute judgements (giving a mark on a scale of 0 to 100) proved very unreliable whereas the comparative approach proved very reliable.

Perhaps more interesting for our purpose are some of the notes the authors make on the use of comparative judgement. First that it is well-suited to "assessing constructs that are not readily defined and operationalised in rubrics. Such constructs, for example "creativity", "problem solving" and "clarity of understanding", are increasingly valued in the 21st century". Second (though relatedly) it is well-suited to assessing unpredictable responses that might fall outside of a strict rubric-based approach. Third that a peer assessment comparative judgement can be performed without training. Fourth they discuss its use as a learning tool. As they note, while there have been studies in other subjects where text-based comments could be made on the comparisons these come at the cost of slowing down the assessment considerably and it is not clear that the comments are likely to be revealing. On the other hand as they note "Rather, we are interested in an approach in which the assessment activity generates discussion about what defines a high quality question response, as was the case in the present study. Moreover,this can be continued in follow up lessons by a class discussion of the final rank order, and the properties of test responses judged most highly by students." A more extreme version of this has been used in Design Education (EdCanNetwork, 2019), where an intermediate peer assessment comparative judgement round was used to inform students in the final completion of their assignment. One could imagine a similar use with two somewhat similar mathematics questions.

5 Jones and Inglis (2015)

This paper, building on Jones et al. (2015), looks at GCSE-level mathematics and the assessment of problem-solving. They make a strong case that there is an element of 'value what we can measure, not measure what we value' in mathematics exams, with lip service continually paid to the importance of problem-solving in the mathematics curriculum, but with only limited methods of assessing it.

The study consisted of two parts. The first part took four experienced exam writers and created a process where they were encouraged to write exam questions that did not require a marking scheme. The second part sought to mark these using a mark scheme fitted retrospectively based on

student responses in pilot settings by an independent fifth experienced exam writer and then also by comparative judgement. The idea was to see to what degree questions were different when mark schemes were not a requirement of their production, and secondly to assess the validity of the comparative judgement assessment approach in this context.

It is not mostly relevant to our present considerations but the report of teachers' responses to the final questions makes interesting reading. Perhaps one element of that that could be relevant is that a significant minority noted that they depended on quite high literacy skills and so might disadvantage some students, including those for whom English was a second language.

Of relevance from the Discussion section is that, while time taken was not described or measured directly, this was an 11-page examination and judges were expected to make 50 pairwise judgements per hour! Combined with the fact that the papers contained four questions on different topics (statistics, geometry etc.) then it seems infeasible that those time constraints could have been met. As the authors note, it seems more likely that comparative judgement is better suited to smaller tasks where the question is more unidimensional.

6 Jones et al. (2015)

This paper looks at applying both rubric-based and comparative judgement assessment to both a traditional GCSE exam and a set of questions designed to be more open-ended. It finds that the assessments agree substantially for both methods and on both papers. Perhaps the interesting point to draw from this paper was that comparative judgement performed very consistently with the rubric-based approach on the traditional GCSE questions, which were a 47 page exam! The examiners noted that they applied varying sampling strategies in order to accomplish the comparative judgement, some choosing the questions that they felt were likely to be on the most differentiating topics (algebra, geometry), some choosing the first few, some choosing the longer ones. They consistently noted that they were uncomfortable with this, feeling that such sampling is unfair on candidates. It is interesting then that the results produced were so consistent with grades from the rubric-based method having a Spearman rank correlation coefficient of 0.91 with the comparative judgement assessment. So at least part of the constraint here seems to be the discomfort of judges in taking a sampling approach,

more than the sampling approach itself being a constraint on the method's efficiency.

7 Bisson et al. (2016)

This paper looks at three mathematical questions and seeks to assess the validity of comparative judgement by comparing the marks achieved by comparative judgement with those achieved in relevant validated closed questions and external marks such as relevant A-level or module marks. Two of the questions — around understanding p-values and understanding derivatives — were addressed to undergraduate audiences. The third — understanding the use of letters in algebra — was addressed to 11-12 year-old schoolchildren. All judgements were performed by 'expert' judges, PhD students in the case of the undergraduate studies.

In both the undergraduate studies the question asked was very open. For p-values:

> Explain what a ***p*-value** is and how it is used to someone who hasn't encountered it before. You can use words, diagrams and examples to make sure you explain everything you know about p-values. Write between half a page and one page.

And for derivatives:

> Explain what a **derivative** is to someone who hasn't encountered it before. Use diagrams, examples and writing to include everything you know about derivatives.

In the case of the p-value question, the comparative judgement assessment had a correlation of 0.46 with the closed assessment, and both had correlations of 0.55 with marks for the overall Applied Statistics module. In the context of the research question this was interpreted as showing that comparative judgement produced similar validity as the validated closed assessments. For the derivatives question, they struggled to get the (supposedly) validated closed questions to provide a meaningful measure, based on internal consistency as measured by Cronbach's alpha, even when they considered subsets of the questions. The conclusion was that comparative judgement produced results moderately correlated with external benchmarks and with high internal reliability. Additionally, in the derivative question the judges were split

into three judging groups, one of which received guidance on how to mark, two of which did not. Of the three groups the correlation of marks was highest between the two non-guided groups, though not significantly so. They do not report any measures for the internal reliability of each of these groups in order to show if the guidance increased consistency, but the statistics they do present suggest that the conclusions of the different judging groups were not materially different, and so any such effect would be small.

This paper is also notable for being the first where they adopt a procedure of randomly splitting the judges into two groups on multiple occasions (20 in this paper, 100 in subsequent ones) and using (the median of) the Pearson correlation of those subgroup's marks as a measure of reliability.

8 Barber (2018)

This is an assessment of a comparative judgement exercise applied to a fourth year Global Health module. I found the write-up a little confusing but there were a number of interesting features. First they tried to apply criteria to the comparative judgement. To my mind it was oddly designed — a hierarchy of criteria that good answers should have such that if A met criteria 1 and B did not then A was judged better even if B was far superior in all other regards. Perhaps not unsurprisingly, these seemed to have the effect of making the comparative judgement less similar to the rubric-based marking. The more surprising element was that the author concluded that more or better criteria were required rather than less. The other interesting element was the positive feedback from students on the peer assessment as a learning experience. For example when asked if it was useful (compared to conventional revision), the responses were: Useful 37; Quite useful 9; Less useful than revision 5. Asked about the fairness of the exercise ("were you convinced the marking was fair?", responses were: Yes 27; Probably a bit uneasy 19; No 4. I think this referred to the peer-assessed grade though it may refer to expert-assessed; it is not quite clear. In this exercise students were also asked to write one line of feedback for each script they viewed. They were not asked whether they found this feedback useful.

9 Demonacos et al. (2019)

I understand this to be a follow up to Barber (2018). Students performed ten pairwise comparisons and provided quite detailed feedback on items reviewed. The ratings were found to have low reliability both internally and when compared to expert grades. Notably however students appreciated the feedback and there was evidence in follow-up interviews completed after the completion of the module that students had found benefit in the exercise for completion of work elsewhere.

10 Jones et al. (2019)

This study used comparative judgement alongside validated closed questions to assess the effectiveness of two school-level algebra learning interventions. The interesting idea here is that comparative judgement can allow an assessment of conceptual understanding and thus an effective evaluation of interventions, in contrast to more procedural questions which may instead test the similarity of the intervention approach to the test.

One of the interesting outcomes in our context was that it found evidence, corroborating the feedback in Jones et al. (2015), that literacy levels were more important in students' success when comparative judgement was used than when the more traditional closed questions were employed.

It also includes an interesting discussion noting that whereas closed question marking will remain the same over time for the same questions, comparative judgement may change as the culture of what constitutes 'good' changes. As the practice matures it will be interesting to see to what degree this is a valid concern (my money is on 'not very' especially with respect to Maths, but even with respect to something like primary school English, though I could see this could be more the case in something like design education as the discussions of tribes in Kimbell (2018) might suggest).

11 Davies et al. (2020)

This paper looks at assessing proof comprehension by providing a proof and then asking students to summarise the proof in 40 words or less. It did not use peer assessment. The interest for present purposes is the question format i.e. providing a proof and then asking students to summarise to see if they

can identify the most pertinent parts. It feels like a nice way to test proof comprehension (very directly) but it is interesting to wonder to what degree this question would be amenable to peer assessment. My sense is that it would be less so than some of the other questions in that the inability of a student to identify the important parts of the proof would be more of a barrier to them appropriately evaluating two pieces of work.

It is also interesting in that part of the method was comparing the comparative judgement results with those from a multiple choice quiz designed (onerously so) to test the conceptual understanding of a particular idea. The multiple choice quiz itself failed to produce internal reliability, suggesting not just that comparative judgement may be the easiest and most efficient way of assessing conceptual understanding, but that it may be the most/only reliable way also.

12 Conversations

12.1 Pete Collison - RM Compare

RM Compare is one of the possible platforms that we could use. RM is an education technology company and RM Compare is one of their products. RM Compare itself does not seem to have a business model at this point. They allow people to use the platform for free (contingent on some T&Cs - not copying the product, provisions around customer data) and seem to be in search of a commercial angle. On the good side I get the impression that they are keen to work with people to try stuff. Their key selling point is that they can take more than just pdfs, which makes sense as I think they have grown out of the Design and Technology field (with the work of Richard Kimbell and Scott Bartholomew), but I do not think is a particular selling point for us.

12.2 Ben Davies - UCL

Ben was the lead author of a couple of the more recent Loughborough comparative judgement papers. His use of comparative judgement is more as a tool to assess pedagogical questions e.g. conception of proof (Davies et al., 2021). Points discussed:

- He uses `nomoremarking.com`, which is free to researchers. He described

the process as being quite simple and that the data returned includes time taken for assessment (which might be useful for some investigations).

- He noted that Ian Jones has released a plug-in for Moodle but he (Ben Davies) has not tested it yet.

- He said that as far as he was aware there were no university Maths departments yet actively using Comparative Judgement as either a learning or assessment tool, including Loughborough (He suggested this was at least partly due to the separation between the Maths department and the Maths education department rather than a perceived limitation in that setting on behalf of Jones et al). [Based on Ian Jones' EAMS talk I think Ben Davies was incorrect on this.]

- He felt that for the type of studies he was doing that Adaptive Comparative Judgement was not necessary.

- He is currently co-authoring with Ian Jones a guide to Comparative Judgement for social scientists, a preprint for which he expects to be available by the end of the month, and may be useful, though the perspective is very much on how to use comparative judgement as a research tool rather than an education tool.

- He noted that he thinks that people approach comparative judgement as a negative marking exercise (in the same way as they often do with rubric-based) in the sense that actually they are not identifying the better piece of work but instead identifying the worse one and then selecting the other. With this being the case it means it might be prone to deductions for features like presentation rather than content, which might be consistent with the concern of disadvantaging students who have English as a foreign language.

- We discussed the dangers of it being a test of literacy. It was not something that he was aware had been looked into.

- He was persuaded of the utility of comparative judgement but noted that in a number of studies that non-expert judges were able as a group to achieve reasonable reliability (though not as high as experts), which could have a number of interesting explanations.

- He likes the idea of formally testing the impact of comparison time by enforcing a particular screen time (or enforcing a minimum or maximum) for each pair (apparently there is a system called PsychoPy that has something like this implemented on it). His suspicion was that this could go really quite low while maintaining reliability.

- He noted that there was a group in Belgium working on comparative judgement, and are looking to publish a comparative judgement special edition of some journal towards the end of the year.

- He is interested in looking at what would happen if people were asked to assess on gradable scale (i.e. marker on a screen that can be pulled to particular point to express degree of superiority) instead of binary, including for example certainty variation by gender.

- He made clear that it was worth going for ethical clearance for any work because there are still relatively few studies so anything would be a useful addition to the literature (and though he didn't mention it, it is the case that there is a certain amount that can quite readily be added afterwards e.g. getting novice or expert judges to do an additional assessment)

- He noted that in his experience (across three universities) what was required by Ethics committees varies greatly so we didn't go into detail of that aspect of his studies.

13 Ian Jones - Loughborough at EAMS

This talk was given as a demo of the Moodle Comparative Judgement plug-in. The YouTube link is `https://www.youtube.com/watch?v=ym_qtl794rg`

- He explained that he uses it in a first year first term module where he uses it for three formative assessments and a final summative assessment.

- In the formative assessments he gets students to complete the assessment in a computer lab with discussion between individuals about which one to select encouraged.

- Afterwards he might for example take the item ranked top and draw out some of the features that it included.
- He also mentioned sometimes seeding some bad answers so that he can talk about those without embarrassing any individual students.
- He gave the example of what he does for the first one using the prompt: "What is an equation? Give examples of how equations can be useful."
- For the summative assessment (if I understood correctly) he gives them three questions that they can choose to prepare, and then in the test, done in class, they are asked to complete two of them (selected by him) in half an hour.
- These are then marked using comparative judgement both by peers but also with a large number done by himself (and other experts? - I was not clear on this point).
- He emphasises that the peer assessment is a form of feedback (not least so that students reflect this in end of module feedback).
- He has found that over the course of the summative assessments students become more comfortable with not having a mark scheme and he is able to tell them "You know what a good answer looks like."
- He converts these into grades from 1 to 5 or 1 to 10 depending on how many marks this is contributing to overall module. He did not go into details on this.
- The Moodle plug-in is designed to work with pdf or jpeg uploads (not so well with other file types e.g. Word).
- Looking at the code it seems that it standardises the scores in some way.
- it uses R package sirt for the Bradley-Terry output and provides the SSR statistic as reliability.
- It is not clear to me what potential there is for using altered code (for example to allow tie-like determinations) or if it is possible to get decision time information as well (which is interesting to at least some of my other investigations).

14 Concluding remarks

A number of things have become clear in looking at the relevant literature. Perhaps the first and most interesting is how nascent its use is in a university setting. It has been adopted only to a very limited extent and only in a few fields where it has been championed by particular individuals. In Mathematics, Ian Jones has done much of that work, and he has collaborators from departments including Edinburgh, but it is not clear to me how widely it is being used. It would seem not greatly.

The second is what are likely to be its most helpful use cases. Comparative judgement, either peer or expert assessed, should not be considered as a potential replacement for the most typical modes of assessment in Mathematics, given the high reliability that is often achieved in (moderated) Maths examinations. However it does seem that the method enables an assessment of a type of question whose inclusion in the overall module mark could increase the validity of that overall module mark, and also that it is highly likely to be the most effective (broadly understood) form of peer assessment possible, so that if peer assessment is considered desirable in itself then comparative judgement is likely to be the best way to achieve that.

Based on my reading of the literature I would suggest two use cases where it could be reasonably and helpfully employed in university level Maths. The first is as a learning tool in a peer assessment task. This has a couple of nice features. The first is that it exposes students to a more open-ended style of question, encouraging and rewarding the ability to understand and communicate a subject in a more holistic way. Second is the exposure to multiple alternative answers to the same question. This can help students gain a sense of the breadth of possible approaches and the relative quality of their own work, and if marks are made contingent on their judgements being consistent with the whole, then it can be a useful learning opportunity. While comparative judgement is sometimes critiqued for not providing feedback, I would argue that in fact through performing the peer assessment process they are getting a form of feedback, which in being more effortful on their part and in exposing them to a greater breadth of responses may be more useful than standard feedback approaches, perhaps one might call it 'self-directed feedback' or some such. This could be further guided by discussing the features of higher and lower-ranked answers in a subsequent lecture. (It might even be an interesting piece of research to test this directly e.g. do a group of students who are required to do the peer assessment perform better

on some later related test than a group who are given conventional feedback and/or a group who are given no feedback but are able to access a range of submitted answers?) Within this framework it might also be interesting to set up two such assessments with the first compulsory but not contributing to module mark, but the second with a similar open-ended question contributing to the module mark, so that the first is taken to be a lesson in answering more open-ended questions, with the second being the assessment of that. [This was written before I became aware of Ian Jones' three formative assessment + 1 summative assessment approach, but seems very consistent with that.]

The second use case that I think the department may want to consider is an expert comparative judgement, especially of a basic component of a mandatory first year course. This offers two benefits. The first is that it feels reasonable to say that this type of assessment could assess a valuable aspect of mathematical understanding that is somewhat distinct from those captured in the typically more closed questions students face, either in exams or assignments. The second is that it allows the assessment to be readily shared between judges while exposing each individual judge to a reasonable spread of students' work. This could reduce the marking burden on one individual but perhaps more valuably it may be beneficial to the department to have a wider scope of the staff who are aware of the sort of level of understanding that students are working at, especially near the start of the course, without having to trawl through whole exam scripts. [As I understand it, Ian Jones does this in his summative assessment but augmented by peer judgements as well.]

I would therefore see these as having different aims — the first being one of primarily a learning experience but with an assessment as a corollary; the second as being an assessment but with a departmental education as a corollary. If the two-stage comparative judgement assessment approach was adopted with the first peer-assessed and the second expert-assessed (and with marks awarded for judging consistency in the first and/or second exercise and for the expert rating in the second) then perhaps both can be incorporated in a usefully complementary manner.

One thing to note is that the reliability of peer-assessed comparative judgement as a means of attributing marks in the university Mathematics setting has only a small evidence base. Those results are encouraging and perhaps if it were to make up only a small proportion of the overall module mark, say 5% or less, then this would be acceptable, but others might reasonably consider that the reliability of peer-assessed comparative judgement

in the context of a particular setting (e.g. question and module) needs to be assessed before it can be used in that way. One of the difficulties in providing this evidence is answering the question "how reliable is reliable enough?". The evidence, such as it is, suggests that peer-assessed comparative judgement is reliable and clearly moreso than other constructs attempting to assess the same conceptual understanding (Bisson et al., 2016; Davies et al., 2020), but less so than expert-assessed comparative judgement, and perhaps uncomfortably close to 'novice' assessment (Jones and Alcock, 2012).

One aspect that I have not proposed here but I would consider to be a very useful element of the approach in general is the ability to get a reliable assessment on the relative distribution of the quality of students over time, as answers from previous years/cohorts can be assessed in direct comparison by including them in the same assessment exercise. I am not quite sure how to operationalise this in the university setting, but one way might be to take an open-ended question, or variety thereof, that would be suitable to ask both at the start of the course and at some intermediate point (for example at the end of first term). For example perhaps "What is mathematical proof?" or "What is a derivative?" or something like the question in Jones and Alcock (2012). Having a reliable distributional estimate of the strength of students as they come in and progress would at minimum be very interesting. Examples of how this has been used in the primary education setting can be found on the NoMoreMarking blog `https://blog.nomoremarking.com/`.

References

Andrich, D. (1978). Relationships between the thurstone and rasch approaches to item scaling. *Applied Psychological Measurement*, 2(3):451–462.

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

Bisson, M.-J., Gilmore, C., Inglis, M., and Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education*, 2(2):141–164.

Bramley, T. (2007). Paired comparison methods. In Newton, P., Baird, J.-A.,

Goldstein, H., Patrick, H., and Tymms, P., editors, *Techniques for monitoring the comparability of examination standards*, pages 246–294. London: Qualifications and Curriculum Authority.

Bramley, T. (2015). Investigating the reliability of adaptive comparative judgment. *Cambridge Assessment, Cambridge*, 36.

Bramley, T., Bell, J., and Pollitt, A. (1998). Assessing changes in standards over time using thurstone paired comparisons. *Education Research and Perspectives*, 25:1–24.

Davies, B., Alcock, L., and Jones, I. (2020). Comparative judgement, proof summaries and proof comprehension. *Educational Studies in Mathematics*, 105(2):181–197.

Davies, B., Alcock, L., and Jones, I. (2021). What do mathematicians mean by proof? a comparative-judgement study of students' and mathematicians' views. *The Journal of Mathematical Behavior*, 61:100824.

Demonacos, C., Ellis, S., and Barber, J. (2019). Student peer assessment using adaptive comparative judgment: Grading accuracy versus quality of feedback. *Practitioner Research in Higher Education*, 12(1):50–59.

EdCanNetwork (2019). How it can be used to enhance teachers' formative assessment skills and students' learning. https://www.edcan.ca/articles/comparative-judgment/.

Falchikov, N. and Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research*, 70(3):287–322.

Jones, I. and Alcock, L. (2012). Summative peer assessment of undergraduate calculus using adaptive comparative judgement. *Mapping university mathematics assessment practices*, pages 63–74.

Jones, I. and Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, 39(10):1774–1787.

Jones, I., Bisson, M., Gilmore, C., and Inglis, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, 45(3):662–680.

Jones, I. and Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, 89(3):337–355.

Jones, I., Inglis, M., Glimore, C., and Hodgen, J. (2013). Measuring conceptual understanding: The case of fractions. In Lindmeier, A. and Heinz, A., editors, *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education (PME 37)*, volume 3, pages 113–120. International Group for the Psychology of Mathematics Education (IGPME).

Jones, I., Swan, M., and Pollitt, A. (2015). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, 13(1):151–177.

Jones, I. and Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation*, 47:93–101.

Kimbell, R. (2018). Constructs of quality and the power of holism. *PATT36 Research and Practice in Technology Education: Perspectives on Human Capacity and Development*, pages 181–186.

Laming, D. (2003). *Human judgment: The eye of the beholder*. Cengage Learning EMEA.

Pollitt, A. (2004). Let's stop marking exams. In *IAEA Conference, Philadelphia*.

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2):157–170.

Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: principles, policy & practice*, 19(3):281–300.

Pollitt, A. and Whitehouse, C. (2012). Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment. *AQA*.