

Peer assessment of mathematical understanding using comparative judgement

IAN JONES AND DAVID SIRL

It is relatively straightforward to assess procedural knowledge and difficult to assess conceptual understanding in mathematics. One reason is that conceptual understanding is better assessed using open-ended test questions that invite an unpredictable variety of responses that are difficult to mark. Recently a technique, called comparative judgement, has been developed that enables the reliable and valid scoring of open-ended tests. We applied this technique to the peer assessment of calculus on a first-year mathematics module. We explored the reliability and criterion validity of the outcomes using psychometric methods and a survey of participants. We report evidence that the assessment activity was reliable and valid, and discuss the strengths and limitations, as well as the practical implications, of our findings.

Much summative assessment on undergraduate mathematics courses takes the form of closed book examinations (Iannone & Simpson, 2011). Traditionally, mathematics examinations sample from across a curriculum using a series of short questions that require accurate and precise answers. Such examinations are well attuned to assessing important procedural knowledge (Star, 2005); that is, knowledge of facts, such as definitions, and algorithms, such as differentiating functions. However, they are less appropriate for assessing equally-important conceptual understanding (Rittle-Johnson & Alibali, 1999); that is, understanding foundational mathematical concepts and the inter-relations between them. One reason for this limitation is that open-ended test questions that prompt a wide and unpredictable variety of student responses lend themselves well to evidencing conceptual understanding, but such questions are difficult to mark reliably.

Ian Jones, *Loughborough University*
David Sirl, *University of Nottingham*

In this article we apply a novel method to assessing an open-ended conceptual test question. The method, called *comparative judgement*, enables evidence of conceptual understanding to be assessed reliably, and is well suited to peer assessment activities. We summarise the method and review its application to assessing mathematics in higher education before presenting the study and results. We focus our discussion on how it might be applied more generally to routinely assess a range of mathematical concepts in universities around the world.

Comparative judgement

Comparative judgement (Pollitt, 2012) is a novel method of educational assessment that was first applied to advanced mathematics examinations about 20 years ago (Bramley, Bell & Pollitt, 1998). Rather than marking examination scripts with reference to rubrics, assessors instead make direct, holistic and subjective comparisons of the quality of students' work. Traditionally, the purpose of marking rubrics was to reduce the subjectivity, and therefore low reliability, of holistic judgements. However, by applying a long-established principle of psychophysics known as the *Law of comparative judgement* (Thurstone, 1927), Pollitt (2012) discovered that holistic comparisons can yield reliable results in educational assessment. This is possible because human beings are very consistent when comparing one object with another, even when the property being compared cannot be defined or measured objectively. For example, it is difficult to accurately judge how many grams an object weighs by holding it in one hand. Conversely, it is relatively easy to judge which of two objects is heavier by holding one in each hand. Early psychophysics researchers discovered that such comparative techniques could be used to create accurate scales of both objective properties such as weight and subjective constructs such as social attitudes (e.g. Thurstone, 1954).

Two technological developments have rendered comparative judgement feasible for educational assessment. First, the advent of the internet means that examination scripts can be scanned and presented easily to judges via web browsers, and their judgement decisions instantly collected. Second, developments in testing theories, and notably the increasing prominence of the Rasch model (Andrich, 1988), enable judges' pairwise decisions to be statistically modelled using maximum likelihood estimation (Firth, 2005) to produce a score for each student. In the absence of the Rasch model and computing power, early researchers were able to construct measurement scales of no more than about ten objects. Nowadays, comparative judgement can routinely be applied to hundreds or even thousands of student responses (e.g. Ofqual, 2015).

Pollitt (2012) argues that a third technological development has made the use of comparative judgement for educational assessment a possibility. This is the application of adaptive algorithms, analogous to those used in adaptive computer-based testing whereby test questions are selected for students based on the performance on questions presented so far. In the context of comparative judgement, pairs of student responses are selected for presentation to judges based on the judgements made so far. For example, if one response has been consistently judged as better than another then there is little to be gained from comparing them again. Instead, adaptive algorithms hunt for pairs of responses that are close in terms of perceived quality in order to maximise the information provided to the system by each judgement. In the present study an adaptive algorithm was used as was standard practice at the time of conducting the study. More recently, debate has emerged as to whether adaptivity in fact contributes to the efficiency of comparative judgement exercises (Bramley, 2015; Pollitt, 2015). It seems that the use of adaptive algorithms has little impact on the final scaled scores assigned to student responses (Wheadon, 2015), and we do not discuss adaptivity further in this paper.

Comparative judgement has been applied to the assessment of varied topics in a range of contexts, including mathematics (e.g. Jones & Inglis, 2015), design and technology ePortfolios (e.g. Kimble, 2012), written English (e.g. Heldsinger & Humphrey, 2010) and experimental reports in chemistry (McMahon & Jones, 2014), amongst other subjects. The focus of many of these studies was on the feasibility and potential educational benefits of using comparative judgement for assessment, whereas other studies have focussed on the validity and reliability of comparative judgement as an assessment method, including for peer assessment (e.g. Jones & Wheadon, 2015). Recent reviews (Bramley, 2015; Tarricone & Newhouse, 2016) have supported the validity and reliability of comparative judgement for assessing students across a range of subject disciplines. In the following section we consider in more detail the application of comparative judgement to the assessment of mathematical understanding.

Assessing mathematical understanding

Within undergraduate mathematics education, comparative judgement has shown promise for enabling the assessment of conceptual understanding in ways not possible using traditional methods (Bisson, Gilmore, Inglis & Jones, 2016). Key to the approach is the design of open-ended test questions that target a specific concept. For example, one of the three studies reported by Bisson et al. administered the following question to 42 engineering undergraduates enrolled on a first-year mathematics

module: "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." Students were allowed 20 minutes to produce a response on a single side of blank paper. The responses were comparatively judged by paid experts (mathematics PhD students); analysis of correlations of the outcomes across different judges and with independent achievement data suggested the assessment produced valid and reliable student scores. The authors replicated this result in two further studies investigating undergraduates' understanding of p -values on a statistics module, and secondary school students' understanding of the concept of variable. Similar results to those described in this paragraph have also been reported for a range of concepts assessed at secondary level (Jones, Inglis, Gilmore & Hodgen, 2013; Jones & Karadeniz, 2016).

The simplicity of the comparative judgement process, in which judges only need to choose one of two presented scripts based on a global criterion such as "Better understanding of derivative", means it lends itself well to peer assessment arrangements (Jones & Alcock, 2014). Commonly, research into peer assessment, which is mostly situated in the context of university education, seeks to establish the validity and reliability of the outcomes of students assessing one another's work (Falchikov & Boud, 2000). A broad finding of the literature into peer assessment in higher education is that valid and reliable outcomes can only be achieved if students are first trained in the application of detailed rubrics (Topping, 2009). However, Jones and Alcock (2014) reported valid and reliable assessment outcomes without student training. 193 mathematics undergraduates were administered an open-ended calculus question. Their responses were uploaded to an online comparative judgement engine and each student was allocated 20 pairwise judgements. The interrater reliability of the outcomes were high, $r = .72$, and the outcomes correlated strongly with the outcomes of experts who were paid to judge the same responses, $r = .77$, and correlated significantly with the scores of independent assessments, $r = .20$. These figures compare favourably with the meta-analysis of peer assessment studies by Falchikov and Goldfinch (2000), suggesting the students had produced valid outcomes.

However, the Jones and Alcock study suffered from three key weaknesses. First, technological problems meant that some responses were not clearly displayed via the web browser, meaning some of the students' judgements were likely to be guesses. Second, experts were paid in order to moderate the scores generated by the peer judging, an expense that presents a barrier to lecturers wishing to adopt the method for routine assessment purposes. Third, the set up and analysis of the procedure was burdensome, requiring psychometric expertise on the part of the

researchers, thereby presenting a second barrier to routine take up by lecturers.

In the remainder of the article we present a study that was designed to overcome these limitations and so produce a peer assessment approach that can be readily adopted by practitioners. To address the first and third limitations we used a new and freely available online comparative judgement engine, *nomoremarking.com*, that does not suffer technical limitations with displaying student responses over a web browser. The website also generates accessible output data that can be easily downloaded and understood by non-expert users. To address the second limitation, the lecturer and one volunteer judged the student responses for moderation purposes, overcoming the expense and time-consuming process of recruiting expert judges.

Research focus

In making these improvements, and deploying the assessment in a more typical teaching context than in Jones and Alcock (2014), we sought to address two research questions.

1. Can peer assessment using comparative judgement in undergraduate mathematics produce outcomes valid and reliable enough for use in summative grading?
2. How do students judge the quality of one response to be better than another when making pairwise decisions?

To address the first question we investigated the reliability and criterion validity of the assessment outcomes, as is advised for peer assessment studies (Topping, 2010). Reliability refers to the consistency of judgement decisions across participants and across test responses. Three statistical procedures were conducted to evaluate the reliability of the findings, as described in the analysis section. Criterion validity is the extent to which outcomes from an assessment predict outcomes from independent assessments of the same or similar constructs (Newton & Shaw, 2014). Two statistical procedures were conducted to evaluate criterion validity, also described in the analysis section later in the article.

To address the second question, which relates to the validity of the peer assessment exercise, we conducted an online survey of students once the comparative judgement activity was completed. The survey and its analysis are described below.

Method

Participants

The research was conducted on a first year mathematics undergraduate Calculus module. The total number of students enrolled on the module was 161. The assessment task reported here was a compulsory requirement and worth 5% of the overall module mark, however 20 students declined permission for their data to be used for research purposes leaving a total of 141 participants reported here.

Testing procedure

The test question was written by the module lecturer, who gave the students a copy of the question one week before the test in order to enable them to prepare. Participants were provided with a single sheet of paper on which to write their answer, and were allowed 15 minutes under examination conditions to complete the test. The test question and an example student response are shown in the appendix.

Judging procedure

The completed tests were collected, anonymised, scanned and uploaded to the comparative judgement engine nomoremarking.com. An adaptive algorithm (Pollitt, 2012) was employed to select pairings of test responses for presentation to students. Students were required as part of the module assessment to complete at least 19 pairwise judgements online within a week of the test. This number was a balance between collecting enough judgements to produce reliable scores for the students, and a pedagogic decision by the lecturer based on the number of peer assessments he felt appropriate for the students. For each pairing of test responses, students had to decide which evidenced "the better understanding" of multivariate calculus. The system was set up such that no student saw their own script when judging.

Following this, two experts contributed an additional 50 judgements each for moderation purposes.

Participant survey

After one week, when students had completed their judgements, a paper-based survey was handed out in a lecture. Completing the survey was an optional research activity and 71 students did so. The survey presented participants with eight "features" of test responses and asked them how

influential each was on their judging decisions. Participants responded to each feature using a five-point Likert scale where 0 = "not at all influential", 2 = "moderately influential" and 4 = "extremely influential". The features are shown in figure 1, and were adapted from previous survey and interview findings into the processes of comparatively judging mathematics work (Jones & Alcock, 2014; Jones & Inglis, 2015). The survey also contained an optional comment box with the prompt "Please state any other features you think may have influenced you when judging pairs of scripts".

Results

Judging outcomes

One-hundred and thirty two students contributed a total of 3258 pairwise judgements. The students were requested to complete 19 judgements each although some did not undertake any judging ($n=9$), some completed fewer than 19 judgements ($n=33$, range 6 to 18 judgements) and some completed more ($n=61$, range 20 to 110 judgements). The number of judgements made on each test response was approximately normal and ranged from 28 to 67, mean = 46.9. The students' decision data was statistically modelled using the BradleyTerry2 package in R (Firth, 2005) to produce an estimate (in the statistical sense) of the "quality" of each test response. Note this step was not necessary and the estimates can instead be downloaded from the website by a non-expert user.

Reliability was measured using three techniques. First, the Scale Separation Reliability (SSR), which is considered analogous to Cronbach's alpha for the case of comparative judgement estimates, was calculated (Bramley, 2007). This was found to be acceptably high, SSR=0.89. Second, misfit statistics were calculated to identify whether any judges were inconsistent with the others, or whether any test responses gave rise to judgement decisions that were inconsistent with the overall dataset. A typical rule of thumb is to consider any statistic more than two standard deviations above the mean to be a misfit (Pollitt, 2012). Only 6 judges (4.5%) and 4 test responses (2.9%) were identified as misfits, suggesting that overall the judging was consistent across judges and across test responses¹. Third, inter-rater reliability was estimated by randomly splitting the judges into two groups, calculating new scores for each group, and calculating the Pearson correlation coefficient of the two sets of scores². This process was repeated 100 times and the median correlation coefficient taken as an estimate of inter-rater reliability, which was found to be high, $r=0.80$.

Criterion validity was estimated in Jones and Alcock (2014) by the student responses being rejudged by paid experts (mathematics lecturers and PhD students). The outcomes of the peer and expert judgements were then correlated. In the present study we sought instead to moderate the peer judgements using a quicker and costless method that can be adopted by practising lecturers. To this end, two expert judges, the module lecturer and a PhD student volunteer, contributed an additional 50 pairwise judgements each. The decision data was statistically remodelled to produce a score for each student, using the procedure described above. To investigate the level of agreement between the peers' and the experts' judgement decisions we recalculated judge misfit figures as described above. For the expert-moderated scores only 5 judges were identified as misfits (compared with 6 judges for the unmoderated scores). Importantly, neither of the expert judges were identified as misfits, suggesting that the students and experts were in broad agreement when making pairwise decisions. In addition, the level of agreement between students and experts suggests that had the experts provided more than 100 judgements there would have been no substantial changes to the scores. This provides indirect support for criterion validity, and does so in a way that could be implemented routinely by lecturers. Moreover, misfit figures are generated by the *nomoremarking.com* comparative judgement engine, meaning moderation can be conducted without expertise in psychometric techniques.

Criterion validity was investigated further through calculating correlation coefficients between scores from different assessments on the module. Students were required to complete eight different assessments, including the comparative judgement exercise reported here. Five of these were "courseworks" (written tests on specific topics undertaken as homework and each worth 5% of the final module mark), one was a class test (an exam-like test covering a sample of topics and worth 10%),

Table 1. *Correlation coefficients between assessment scores on the module (n=137)*

	CJ	CW1	CW2	CW3	CW4	CW5	Test
CW1	0.26						
CW2	0.26	0.22					
CW3	0.10	0.17	0.50				
CW4	0.33	0.48	0.24	0.32			
CW5	0.21	0.04	0.12	0.16	0.07		
Test	0.35	0.40	0.29	0.32	0.47	0.13	
Exam	0.34	0.46	0.31	0.31	0.56	0.16	0.69

and the other was a synoptic exam (sampling from across the entire curriculum and worth 60%). Complete data was available for 137 of the students and the correlation coefficients between these assessments are shown in table 1. The mean correlation coefficient between comparative judgement scores and other assessments was $r=0.26$, and the mean correlation between the other (non-comparative judgement) assessments was $r=0.31$. This suggests the criterion validity of the comparative judgement outcomes was in line with the other assessments used on the module.

Survey outcomes

Seventy-one students, a self-selected sample, completed the survey following the peer assessment activity. Mean scores for the eight features that may have influenced judging decisions are shown in figure 1. A one-way repeated-measures ANOVA revealed a main effect for feature scores, $F(7, 70) = 36.37$, $p < .001$. Post-hoc tests (Bonferroni corrected) revealed that "Accuracy of answers" ($M=3.42$) was the highest rated influential feature. "Quantity of ink used" ($M=1.62$) and "Flair and originality" ($M=1.59$) were the lowest rated features and were not significantly different from one another. The other five features were rated between these two extremes and were not significantly different from one another.

We also turned to the students' open-text responses to the optional survey prompt asking them to suggest other influential features missed from the Likert questions. Twenty-nine students volunteered responses

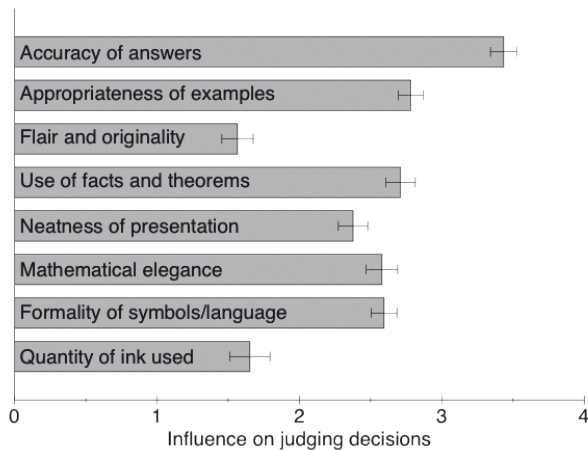


Figure 1. Student ratings of influences on their judging decisions ($n=71$). Error bars show the standard error of the mean

Table 2. Responses from a self-selected sample of 29 students who responded to the open-text question in the survey. Responses have been formatted and spellings corrected for presentation purposes

PRESENTATION	
1	Good/reasonable use of the English language, and able to explain well.
2	Good use of diagram.
3	Presentation. Handwriting. Layout.
4	Diagrams & explanation of how they got to their answer.
5	Diagram.
6	Demonstration of understanding rigorousness of proof.
7	Being able to read the script.
8	The size of font, the organisation of script, use of graphs.
9	Whether or not I could read the persons writing. How they linked diagrams to their explanations.
10	Handwriting. Layout.
11	Accuracy of diagrams. Labelling of diagrams. Layout.
12	Graphs and illustrations.
13	Sketches.
UNDERSTANDING	
14	Some of the scripts were both correct. Some were both wrong. Sometimes, I could not decide which one was better.
15	Layout of answers. Does one argument followed logically from the previous and relating the working to a graph.
16	Qualitative and descriptive scripts with written explanation to supplement the mathematical equations etc.
17	Understanding the question. Care in actually answering the question.
18	I judged most of them compared to my answers which I gave as we weren't given told the right answer.
19	Preciseness and "to the pointness" of answers i.e. no wasteful words etc.
20	Legibility. Sensible ordering of arguments.
21	Thorough analysis.
22	The accuracy of the graph.
23	I found that a lot of the scripts I had to compare had large chunks missing off their answers and this meant I found it difficult [sic] to see if an answer was complete in some circumstance, this led to me choosing the other script. Quality of graphs, if graphs weren't used then I found it difficult to see that the person truly knew if the limit was continuous and everywhere.
TECHNICAL	
24	I had a problem with the system. If I wanted to keep the 1st one of the two scripts, I pressed the button but the other one was kept at the end. Also, it there wasn't any window to press "finish", so I did 33 scripts.
25	Legibility of the photocopy. Quality of the diagram. Flow of the script.
26	Found it difficult reading scripts online, making it difficult mask [sic: presumably should be "task"].
27	Functionality of online system I marked an answer incorrectly but couldn't go back.
OTHER	
28	None I guessed most of them.
29	None. I guessed them all.

to this question. Scrutiny of the comments led us to categorise each as focussing on "Presentation" (thirteen comments), "Understanding" (ten comments), "Technical" (four comments) or "Other" (two comments). The responses are listed in full in table 2.

Some of the responses appear to reflect influences listed in the Likert survey, for example comment 3 in table 2 refers to neatness and comment 22 refers to accuracy. Others suggest influences not directly listed in the Likert survey, including rigour (e.g. comment 6), legibility (e.g. comment 7), layout (e.g. comment 8), relevance (e.g. comment 19) and flow of reasoning (e.g. comment 15). Such comments can be used to inform the design of surveys for future studies, helping to ensure that a redesigned survey does not contain too many ill-defined and overlapping influences.

In contrast to the Jones and Alcock (2014) study, which suffered from some technological problems, here such problems were largely avoided. The only evidence of technological difficulties in the present study was the four "technical" comments in table 2. However none of these comments suggest any critical problems. Indeed two seem to reflect expectations on the part of the students that were not supported by the technology: one comment reflects a misunderstanding of the online judging system (comment 24), and another felt there should be an "undo" facility (comment 27). The other two comments referred to legibility, which was a feature of the scripts themselves rather than how they were rendered online.

Discussion

Overall, we found that the comparative judgement peer assessment procedure was robust enough to be used for summative assessment. Specifically, the peer assessment outcomes were reliable: that is, independent groups of students sampled from within the module cohort produced approximately the same scaled rank order. Indeed the reliability coefficient reported above, $r = .80$, is higher than that reported by Jones and Alcock (2014), $r = .72$, suggesting that the improvements made to the design of the assessment process produced more reliable outcomes. We note however that the difference between these two reliability coefficients was not significant, $z = 1.65$, $p = .10$. It is not possible to compare this reliability to that of other comparative judgement assessment studies because, as Bramley (2015) found in his review, reliability coefficients are not usually estimated or reported.

The assessment outcomes were also valid: expert judgements were aligned with those completed by the students and scores correlated as expected with scores from independent module assessments. The mean

criterion validity coefficient reported here, $r = .26$, is higher than that reported by Jones and Alcock (2014), $r = .20$, although again this difference was not significant, $z = 0.53$, $p = .60$. As with reliability, it is not possible to compare this coefficient with other comparative judgement studies because criterion validity as measured against independent assessments of achievement tend not to be reported (Bramley, 2015).

The survey data provided insights into how students judged the quality of one response to be higher than another when making pairwise decisions. Overall, the findings from the survey were unsurprising. Those features we might expect to contribute to a response that is perceived to be of higher quality were rated highly. In contrast, the feature we might hope is not as influential, "quantity of ink used", was less influential. The exception was "flair and originality" which might be something we hope students aspire to and admire in mathematical work. We do not know why this was not rated highly. It may be that the test question did not lend itself to flair and originality, or that the item should have asked only about flair or originality rather than conflating the two. It is also possible that the low mean rating for this item reflects an algorithmic or test-oriented approach to learning and thinking about mathematics.

The open text responses provided further insights into judging processes, at least for the case of the self-selected sample of 71 students who completed the survey. The majority of the comments related to issues of presentation and understanding, as shown in table 2. It is perhaps the comments related to understanding that are the most enlightening. For example, comment 14 in table 1 alludes to the difficulty of deciding between student responses that are similarly strong or similarly weak. This is certainly the case, and in general we advise judges (expert or peer) that when faced with a tough decision to make the best call they can. Another participant raised a challenge for peer assessing (comment 18), namely that in the absence of a provided "correct" answer judgements had to be made relative to students' own conceptions of a "correct" answer. Based on this, we might expect weaker students, whose responses were judged less favourably, to also be weaker judges. We investigated this possibility by correlating students' misfit figures (the extent to which their judgements were consistent with other students) and their comparative judgement scores. A high correlation would suggest a systematic relationship between performance on the written test and performance when judging peers' response. Perhaps surprisingly this was not borne out: the correlation coefficient was negative, as expected, however it was not strong and not statistically significant, $r = -.16$, $p = 0.06$.

Strengths and limitations

The present study was cost effective, with no experts being paid to establish the validity of the findings. This is not ideal for research purposes, and the compromise was that we provided only a proxy for estimating the match between peer and expert assessment outcomes. Nevertheless, the present study offered a practical and free method for moderating the results as a step towards routine use. The total expert time required for making 100 judgements was about two hours, substantially less time than would be required for marking all the responses were a viable rubric available.

Regarding the survey data, we acknowledge that caution must be exercised when interpreting the results of participant recall of cognitive processes. Participants may respond sincerely and yet are often unable to accurately recall how decisions were made (Nisbett & Wilson, 1977). Moreover, we emphasise here this was a self-selected sample of students who volunteered their time to complete the survey. Accordingly, survey data following a comparative judgement study can help reassure us as to the validity of the outcomes in terms of judges' engagement and post-hoc perceptions of how they made their decisions. However, it would be inadvisable to conclude that the features most highly rated or suggested in the text comments genuinely were those that most influenced pairwise decisions. Alternative methods are required to understand the cognitive processes of comparatively judging mathematical work. Kelly's Repertory Grids (Johnson & Nádas, 2012) and eye-tracking studies (Rayner, 1998) offer possible avenues for future research.

Implications for practice

A common concern about comparative judgement is the assumption that grades must be normative rather than criterion-based; that is, a given student judged amongst a high-achieving cohort will get a lower score than if judged amongst a low-achieving cohort. In fact this concern is unwarranted. The scores from a comparative judgement exercise can be used for assessment procedures, such as criterion-based grading, just like a set of scores generated by traditional marking methods. Grading comparative judgement scores has been described in detail in Jones and Alcock (2014) and McMahon and Jones (2014), and will not be repeated here.

Another commonly expressed concern is that students receive no written feedback from comparative judgement assessments. This is true, and may be perceived as a barrier in light of the increasing expectations

on lecturers to provide detailed, qualitative written feedback to students on all work submitted. However, while no feedback is presented in the traditional sense, we argue that the judging process engages students with meaningful comparisons of the quality of answers, and thereby provides a novel and beneficial form of feedback about their own performance. This learning benefit of peer comparative judgement can be enhanced by preceding a summative assessment with practice assessments. Our approach is to provide students, prior to the administering of a summative test, with two or three dry runs at open-ended tests and comparatively judging the responses.

The development of comparative judgement for educational assessment has been a long and painstaking journey that goes back over a century. In recent decades, technological and theoretical developments have made it possible, but it generally requires the input of specialist researchers and psychometricians to operationalise the approach in practice. The present study, while undertaken by specialists, marks a stepwise shift towards routine use by lecturers in mathematics.

References

- Andrich, D. (1988). *Rasch models for measurement*. London: Sage.
- Bisson, M.-J., Gilmore, C., Inglis, M. & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education*. Online first. doi: 10.1007/s40753-016-0024-3
- Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgment* (Cambridge Assessment Research Report). Cambridge: Cambridge Assessment.
- Bramley, T. (2007). Paired comparison methods. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 264–294). London: QCA.
- Bramley, T., Bell, J. & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives*, 25, 1–24.
- Falchikov, N. & Goldfinch, J. (2000). Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70, 287–322.
- Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical Software*, 12 (1), 1–12.
- Heldsinger, S. & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, 37, 1–19.

- Iannone, P. & Simpson, A. (2011). The summative assessment diet: how we assess in mathematics degrees. *Teaching Mathematics and Its Applications*, 30, 186–196.
- Johnson, M. & Nádas, R. (2012). A review of the uses of the Kelly's Repertory Grid method in educational assessment and comparability research studies. *Educational Research and Evaluation*, 18, 425–440.
- Jones, I. & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education*, 39, 1774–1787.
- Jones, I. & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, 89, 337–355.
- Jones, I., Inglis, M., Gilmore, C. & Hodgen, J. (2013). Measuring conceptual understanding: the case of fractions. In A. M. Lindmeier & A. Heinze (Eds.), *Proceedings of the 37th conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 113–120). Kiel: PME.
- Jones, I. & Karadeniz, I. (2016). An alternative approach to assessing achievement. In Csikos, C., Rausch, A. & Sztányi, J. (Eds.), *The 40th conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 51–58). Szeged: PME.
- Jones, I. & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies in Educational Evaluation*, 47, 93–101.
- Kimbell, R. (2012). Evolving project e-scape for national assessment. *International Journal of Technology and Design Education*, 22, 135–155.
- McMahon, S. & Jones, I. (2014). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice*, 22, 368–389.
- Newton, P. & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: Sage.
- Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review*, 84, 231–259.
- Ofqual (2015). *A comparison of expected difficulty, actual difficulty and assessment of problem solving across GCSE maths sample assessment materials* (Report No. Ofqual/15/5679). London: HMSO.
- Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19, 281–300.
- Pollitt, A. (2015). *On "reliability" bias in ACJ* (Technical report). Cambridge: Cambridge Exam Research.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372–422.
- Rittle-Johnson, B. & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology*, 91, 175–189.
- Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, 36, 404–411.

- Tarricone, P. & Newhouse, C. P. (2016). Using comparative judgement and online technologies in the assessment and measurement of creative performance and capability. *International Journal of Educational Technology in Higher Education*, 1, 13–16.
- Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273–286.
- Thurstone, L. L. (1954). The measurement of values. *Psychological Review*, 61, 47–58.
- Topping, K. J. (2009). Peer assessment. *Theory Into Practice*, 48, 20–27.
- Topping, K. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction*, 20, 339–43.
- Wheadon, C. (2015). *The opposite of adaptivity?* (Blog post, Feb. 10, 2015). Retrieved from <https://blog.nomoremarking.com/the-opposite-of-adaptivity-c26771d21d50#2jably9kx>

Notes

- 1 Under the assumption of a normal distribution we would expect around 5% to be misfits.
- 2 This produces an underestimate of the true inter-rater reliability because the "split-halves" technique described here effectively requires throwing out half of the judgement decisions in to order to produce a correlation coefficient.

Appendix

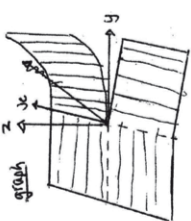
Test question and sample student response.

Conceptual Test Question

Consider the function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ given by

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

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



limits

- For $x < 0$, $f(x, y) = 0$
 $\lim_{(x,y) \rightarrow (a,b)} f(x, y) = f(a, b) = 0$
 with $a < 0$ and $b \in \mathbb{R}$
 $f(x, y) = -x$
- For $x > 0$ and $y > 0$
 $\lim_{(x,y) \rightarrow (a,b)} f(x, y) = f(a, b) = a \cdot b$
 with $a > 0$ and $b < 0$
- For $x > 0$ and $y < 0$ $f(x, y) = -x$
 $\lim_{(x,y) \rightarrow (a,b)} f(x, y) = f(a, b) = -a$
 with $a > 0$ and $b < 0$
 $\lim_{(x,y) \rightarrow (a,b)} f(x, y) = -a$
 $\lim_{(x,y) \rightarrow (0,b)} f(x, y) = 0$

Partial derivatives

For $x < 0$, $f(x, y) = 0$ The function is constant at this interval
 \therefore the derivative is 0

For $x > 0$ & $y < 0$ $f(x, y) = -x$ respect to x $\frac{\partial f}{\partial x} = -1$
 respect to y $\frac{\partial f}{\partial y} = 0$

For $x > 0$ & $y > 0$ $f(x, y) = xy$ $\therefore \frac{\partial f}{\partial x} = y$ $\frac{\partial f}{\partial y} = x$
 $\therefore -1 \neq 0$

Therefore the limit exists at $f(x, y) / \{ (x, y) | x > 0, y > 0 \}$

Continuity

The function is continuous everywhere apart from the semi line where $x > 0$ and $y = 0$

Explanation:
 On the set $\{x > 0, y = 0\}$ we have $\lim_{(x,y) \rightarrow (a,0)} f(x, y) = 0$ (from 3)
 we have $\lim_{(x,y) \rightarrow (a,0)} f(x, y) = f(a, 0) = -a$
 Hence limits are not the same so function is not continuous

Ian Jones

Ian Jones obtained a PhD in Mathematics Education from the University of Warwick and is a Senior Lecturer in the Mathematics Education Centre at Loughborough University, UK. Prior to this he was a Royal Society Shuttleworth Education Research Fellow and taught in primary and secondary schools for ten years. His research interests are in school children's learning of algebra and the assessment of procedural and conceptual understanding of mathematics.

i.jones@lboro.ac.uk

David Sirl

David Sirl is a Lecturer in the School of Mathematical Sciences at the University of Nottingham. He is enjoying spending some time working with education researchers to explore new ways of improving teaching and learning.

david.sirl@nottingham.ac.uk