Peer assessment without assessment criteria

Ian Jones (corresponding author)
I.Jones@lboro.ac.uk
UK 1509 228 217

Lara Alcock
L.J.Alcock@lboro.ac.uk

Mathematics Education Centre
Schofield Building
Loughborough University
Loughborough
LE11 3TU

Abstract
Peer assessment typically requires students to judge peers' work against assessment criteria. We tested an alternative approach in which students judged pairs of scripts against one another in the absence of assessment criteria. First year mathematics undergraduates (N = 194) sat a written test of conceptual understanding of multivariable calculus, then assessed their peers' responses using pairwise comparative judgement. Inter-rater reliability was investigated by randomly assigning the students to two groups and correlating the two groups' assessments. Validity was investigated by correlating the peers' assessments with (i) expert assessments, (ii) novice assessments, and (iii) marks from other module tests. We found high validity and inter-rater reliability, suggesting that the students performed well as peer assessors. We interpret the results in the light of survey and interview feedback, and discuss directions for further research and development into the benefits and drawbacks of peer assessment without assessment criteria.

Keywords
Peer assessment; comparative judgement; validity; reliability; mathematics

Introduction

Peer assessment involves students in making judgements of their peers' work. It presents learners with a complex challenge, typically requiring them to reference peers' work against assessment criteria (van den Berg, Admiraal and Pilot 2006) in a domain they have not yet mastered. Students' success at assessing their peers, is varied (Falchikov and Goldfinch 2000), and literature reviews (Boud, Cohen and Sampson 1999; Dochy, Segers and Sluijsmans 1999; Falchikov and Goldfinch 2000; Gielen et al. 2011; Kollar and Fischer 2010; Topping 2003) have identified design principles that appear to improve students' performance.

One design principle of direct theoretical interest to the present study is that peer assessment should make use of specified assessment criteria (Dochy, Segers and Sluijsmans 1999; Orsmond, Merry and Reiling 1996; Topping 2003). Clear and detailed assessment criteria are considered important for ensuring that teachers and students have a common understanding what is being assessed (Orsmond, Merry and Reiling 2000). Agreement between teachers and students is important for achieving valid assessment outcomes, and study designs in which student participants are familiarised with assessment criteria, either through training (e.g. Topping et al. 200) or through involving them in the development of these criteria (e.g. Orsmond, Merry and Reiling 2000; Sluijsmas et al. 2004), tend to yield more valid, more reliable results (Falchikov and Goldfinch 2000). While the use of assessment criteria does not guarantee sound psychometric measures (e.g. Chang et al. 2011; Tsivitanidou, Zacharia and Hovardas 2011), the overall evidence strongly supports this principle. Moreover, a shared understanding between teachers and students of explicit assessment criteria, and therefore of the learning domain itself, is an inherently desirable pedagogic goal (Orsmond, Merry and Reiling 2000).

The criteria used in peer assessment research vary from study to study, but it is common to have several headings populated by descriptive comments. Chang et al. (2011), for instance, constructed a portfolio assessment rubric that consisted of six dimensions (portfolio creation, learning goal, artifact, reflection, attitude, others, overall). These dimensions comprised 22 individual criteria, such as "accomplishment and growth" and "quality of interaction", each to be scored on a 1-to-5 scale (1309). Similarly, Magin (2001a) described a two-dimensional rubric for assessing "group processes", also scored on a 1-to-5 scale for explicit criteria as follows:

> "*Contribution to discussion*: extent to which student took part in discussion by adding own ideas, experiences, and by taking others' ideas seriously and expanding on them.
>
> *Contribution to development of the group*: extent to which student fostered group development by attending regularly; supporting other members of the group; taking part in the group's activities; and performing group tasks." (143)

In this article we present an alternative approach to peer assessment which, counter to the general trend in the literature, does not require assessment criteria. The aim is not to challenge the evidence in support of using criteria but to explore a different

approach to achieving acceptable psychometrics. In the following section we describe the approach and set out three motivations for the study.

Comparative Judgement

Comparative judgement (CJ) is an approach to assessment that involves presenting assessors with pairs of student scripts and asking them to judge which of the two students has performed "better" (Pollitt 2012). The binary outcomes of many such judgements are then used to create a scaled rank order of scripts from "worst" to "best".

The rationale for the approach is a long-standing psychophysical principle, usually called "The Law of Comparative Judgement" (Thurstone 1927). This principle stands on the robust result that human beings are more reliable at making relative judgements of one sense impression against another than they are at making objective judgements of individual sense impressions in isolation (Laming, 1984). CJ can be used to construct psychological scales by fitting the outcomes of many judgements to statistical models. Early research focused on sense impressions that have physical correlates, such as weight, and later work investigated the construction of scales for subjective properties, such as beauty (e.g. Thurstone, 1954). An early application of CJ to educational assessment was in standards monitoring across equivalent forms of examination papers (Bramley, Bell and Pollitt 1998). More recent studies have tested CJ in a variety of assessment contexts (Kimbell 2012), including school mathematics (Jones, Swan and Pollitt submitted) and peer assessment (Seery, Canty and Phelan 2012).

We had three motivations for studying the use of CJ for peer assessment in undergraduate mathematics.

First, we wished to conduct a feasibility study. We wanted to test whether CJ could yield acceptable validity and reliability (as defined in the following section) when used as the basis for peer assessment, and whether it could therefore offer an alternative approach to the use of assessment criteria.

Second, we wished to study the utility of CJ for assessing subtle and complex constructs that are hard to operationalise in terms of explicit criteria. An example of this difficulty in the context of peer assessment was provided by Topping, Smith, Swanson and Elliot (2000), who reported that criteria focusing on "quality of thought" are less likely to yield high inter-rater reliabilities than criteria focusing on "structural features" (159). Similarly, Orsmond, Merry and Reiling (1996) found that students had difficulty marking to higher-order criteria such as a "clear and justified conclusion". Laming (2004) made the point more strongly, arguing "it has yet to be shown that 'originality' or 'ability to argue persuasively' or 'clarity of thought' can be assessed with any reliability greater than the negligible" (95). A theoretical advantage of CJ for assessing difficult-to-specify constructs is that validity can be defined in terms of the collective knowledge of the assessors, rather than in relation to written criteria. This means that CJ might be better than criterion-based methods when used for assessment of elusive constructs that are central to a discipline and understood by experts, but that are difficult to define accurately and comprehensively. In

mathematics education, such constructs include "communicating mathematics", "creativity" and "problem solving" (e.g. Levesley 2011; Swan and Burkhardt 2012). Moreover, tests designed to assess such constructs are often open-ended, and students' responses can be expected vary in unpredictable ways that are difficult to anticipate in scoring rubrics. The lack of assessment criteria means that CJ can, in principle, cope with such variety in student responses. Therefore an attraction of CJ as an assessment method for undergraduate mathematics is its potential for assessing another important but nebulous construct, that of "conceptual understanding".

Our third motivation for using CJ was pedagogical: we wished to offer students an opportunity to reflect on their own conceptual understanding and communication of mathematical ideas, and thus to promote higher-order learning. The promotion and valuing of peer learning (Slavin 1991) is a common motivation for implementing and studying peer assessment (Topping 2003; Gielen et al. 2011; Kollar and Fischer, 2010; van Steendam et al. 2010; van Zundert, Sluijsmans and van Merriënboer 2010). Peer assessment is often used to generate and evaluate evidence of typical peer learning activities such as team working, reflective thinking without the guidance of a teacher, and domain specific communication skills (Boud, Cohen and Sampson 1999). In the case of the present study our particular motivation was to promote students' awareness of communicating their understanding of sophisticated mathematical ideas to others. Moreover, with regards to the CJ approach in particular, a recent study by Pachur and Olsson (2012) suggested that comparing pairs of items is a more effective learning strategy than comparing individual items against criteria. Again, therefore, CJ has theoretical promise as an alternative to criterion-based peer assessment methods.

Research focus and design

Although our motivations for studying criteria-free peer assessment were both theoretical and pedagogical, our main focus in this paper is the theoretical feasibility of CJ as a peer assessment method. To this end our research design focused on the validity and reliability of peer assessment in our context.

In peer assessment studies, *validity* is commonly operationalised as the extent to which students interpret and apply assessment criteria in the same way as experts. Expert assessments are assumed to provide an "objective" reference, and validity is measured by correlating peer assessments with expert assessments (Falchikov and Goldfinch 2000). However, there are two potential problems with this approach. First, expert and peer assessments are often conducted using different methods or different sets of criteria (e.g. Chang et al. 2011; Seery, Canty and Phelan 2012). In our study, this was not problematic because explicit judging criteria were not used; validity is simply operationalised as the extent to which students understand the global construct "conceptual understanding" in the same way as experts.

Second, expert assessments can be unreliable (Falchikov and Goldfinch 2000; Magin 2001b; Murphy 1982; Newton 1996; Topping 2003), so the assumption that they constitute an objective reference may not be warranted. To address this, our first step in this study was to measure the expert *inter-rater reliability* of using CJ assess advanced mathematical understanding. Inter-rater reliability is a measure of the extent to which an assessment outcome would be identical if repeated with an independent

assessor or group of assessors drawn from the same population. To measure inter-rater reliability, two groups of experts independently used CJ to assess the students' work and the outcomes from the two groups were correlated. Experts, by definition, can be expected to have the knowledge required to make good judgements, so we predicted high expert inter-rater reliability. We then measured the inter-rater reliability of peer use of CJ for assessing advanced mathematics. Assuming that the students, who were first year mathematics undergraduates, are a reasonably homogenous group in terms of the construct being assessed, we predicted a high peer inter-rater reliability. However, peer assessors can be assumed to vary more in mathematical knowledge and ability than experts, so we expected the inter-rater reliability for the peer judgements to be lower than that for the experts.

Having established inter-rater reliabilities for the two groups, we next investigated the validity of the peer assessments in three ways. First, we correlated the outcomes of the peer and expert assessments. Given that the students had been studying material assessed in the written test, we predicted a high validity for the peer assessments. Second, we correlated outcomes from the CJ assessment with those of other summative tests from the module. We anticipated that the peer assessment outcome would be consistent with the outcomes of the module summative tests. Third, because little is known about how judges make their decisions when comparing pairs of scripts (Bramley and Gill 2010), we investigated the possibility that peer judges might be making their judgements on the basis of surface features such as neatness and layout of work rather than on the underlying mathematical concepts. To do so, we recruited a group of novices: social science PhD students who had not studied mathematics beyond secondary school level, and in particular who in had never taken any courses in calculus. Validity of the novices' collective judgements was measured, as before, by correlating the novice scaled rank order with that of the experts. Because novices, by definition, would need to base their assessments on surface features rather than on mathematical understanding, we predicted that the validity of the novices' assessments would be low, and would provide us with further information regarding the comparative validity of the peer judgements.

The above constituted our main analysis, and the results are reported after a detailed discussion of the methods. However, after the main study we also obtained qualitative information by inviting judges to give feedback about their experiences in two ways: via semi-structured interviews and via online open-text comments. Interviewees were self-selecting participants from across the three groups; nine students, seven experts and three novices were interviewed. Online, 50 participants submitted feedback (25 students, 16 experts, 9 novices). This information provided insights valuable for understanding both the participants' judgements and their opinions about peer assessment as implemented in this study; it is discussed after the main results.

Method

Participants

Peers. The students were first-year mathematics undergraduates enrolled on a calculus module. The total cohort was 224; 193 undertook the online peer assessment activity and are included in the analysis reported here.

Experts. The expert judges were 9 mathematics postgraduate students and 11 mathematics lecturers. The postgraduate students were paid for their time and the lecturers volunteered their time.

Novices. The novice judges were 9 non-mathematics postgraduate students who had never studied mathematics beyond high school level and who had never been taught any calculus. The novices were social science students adopting predominantly qualitative research methods in their thesis work. The novices were all paid for their time.

Materials

Written test. A written test was designed for the study by a module lecturer (the second author). The test was intended to assess the students' conceptual understanding of limits, continuity and partial derivatives, by asking them to explain how these concepts apply in relation to a non-typical function as shown in Figure 1. At the time of the study, the students had received definition-based lectures and exercises related to limits, continuity and partial derivatives for functions of two variables, although they had not completed much work on the last. The test required students write their responses on a single side of paper.

***** FIGURE 1 HERE ******

Test preparation. Students were provided with a copy of the question six days before sitting the written test in order that they might prepare to give their responses. In the same document, they also received practical information about the test arrangements. In light of the novelty and conceptual challenge of the test question the students were also advised that they might want to consider these criteria when thinking about how to respond:

- 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?

Note that these criteria were provided to the students to aid in their preparation for the test and were not reinforced or restated for the peer assessment exercise itself. Students did not use these test criteria to score peers' work on separate headings, but instead compared pairs of responses in relation to the global construct "conceptual understanding". As such there may have been some common, implicit criteria that the students used when assessing peers but, in contrast to the peer assessment literature described in the introduction, there was an explicit absence of clear and defined criteria for the point-wise scoring of peers' work.

Procedure

Test administration. The test was administered in a lecture under examination conditions, and students were allowed 15 minutes to complete it. Their responses were then collected, anonymised and uploaded to the online CJ system.

Comparative Judgement. The arrangement of peer assessment used was summative, anonymous, asynchronous and non-reciprocal. For the CJ part of the study we used *e-scape*, developed by TAG Developments (Derrick 2012). The *e-scape* system presents pairs of scripts online via a browser. The system uses an adaptive algorithm based on collective judging history so far to select which pair of scripts to present to judges (this selection is designed to maximize the efficiency of the ordered scale construction, not to personalise the scripts received by any given judge) (Pollitt 2012). At the judging stage, students were not reminded of the suggested test preparation criteria described above and were simply asked to make a global judgement about each pair of scripts. Students received no training beyond instruction on using the CJ website (described later) and were not required to provide or respond to specific qualitative feedback.

The day after the written test, a researcher explained the paired judgements activity and demonstrated the CJ website in a lecture. The students were told that they would be presented with pairs of scripts and that they should, for each pair, decide which student had demonstrated the better conceptual understanding. The students were then told to complete 20 judgements over the course of the following week. They were advised that judgements should take on average about three minutes and that the total work should take no more than an hour.

The postgraduate experts and novices attended demonstration workshops that lasted about 30 minutes and were led by a researcher. One expert postgraduate could not attend and received a one-to-one demonstration session with the researcher the following day. The lecturer experts received small group or one-to-one training sessions in person or online.

The experts were given the test question one week prior to the training and were asked to complete it themselves in order to familiarise themselves with the test. With the same goal of familiarisation, the novices were also given the test question one week prior to the training. However, as the novices were by definition incapable of completing the test, they were instead provided with three randomly selected student responses and were asked to rank the responses in order of conceptual understanding.

Grading. The written test and comparative judging replaced a calculation-based online test. Grading was criterion-referenced and was decided by placement of each student's script according to a scaled rank order produced by experts. The grades contributed 5% towards the students' overall module grades.

Data preparation

The output from the *e-scape* system was a complete judgement history that recorded, for every pair of scripts presented, which script was judged better, and who made the judgement.

The total number of written tests completed was 201. For internal assessment purposes, all 201 scripts were comparatively judged by the students and by the mathematics lecturers. However, 33 students opted out of their scripts being used for research purposes. Consequently, the data analysed here includes only the remaining

168 scripts. Judgements that included any of the 33 excluded scripts were removed from the data before analysis.

For technological reasons we could not prevent students from occasionally seeing their own script paired with that of a peer. In practice we found that in 29 judgements a student was presented with her or his own script. The focus of the study is on peer assessment, not self assessment, so these 29 judgements were removed from the data before analysis. (We note for interest that in all 29 cases the student favoured her or his own script over the other.)

After the removals described above, the students completed a total of 2813 judgements. To obtain our measure of the inter-reliability of the peer assessment we randomly allocated each undergraduate and all of their corresponding judgements to one of two peer groups called "Peer 1" and "Peer 2".

The experts completed a total of 2965 judgements. One postgraduate completed most judgements in fewer than ten seconds and so was deemed not to have taken the work seriously; this person's judgements were removed before the analysis. This left a total of 2797 completed expert judgements. To obtain our measure of expert reliability we randomly allocated each expert and all of their corresponding judgements to one of two groups, called "Expert 1" and "Expert 2"; each group included both lecturers and postgraduates.

The novices completed a total of 1217 judgements and were treated as one group. For analysis purposes, the number of assessments made tends to increase reliability, so it was important that every group included the same number of judgements. Accordingly, we randomly selected 1217 judgements from each of the two peer groups and from each of the two expert groups for use in the analysis.

Analysis and Results

Internal consistency of scaled rank orders

The judgements for each of the five groups of participants were fitted to the logistic form of the Rasch model (for details on applying Rasch analysis to CJ data see Bramley 2007). The Rasch analysis produced, for each script, a standardised parameter (*z*-score) and a standard error. The parameters were then used to construct, for each group, a unidimensional scaled rank ordering of scripts.

Fitting CJ data to the Rasch model produces three measures that can be used to check the internal consistency of the scaled rank order. The first is the Rasch sample separation reliability, which many consider directly analogous to Cronbach's alpha (e.g. Bond and Fox 2006; Wright and Masters 1982). The second is each judge's information mean square, or "misfit", figure, which indicates how well each judge's individual decisions matched those of the overall consensus. The convention is to consider those judges whose misfit is greater than two standard deviations above the mean to be inconsistent (Pollitt 2012). The third is each script's "misfit" figure, which indicates the consistency of relative decisions about the quality of the script across all judges. A misfit figure greater than two standard deviations above the mean suggests a lack of agreement regarding that script.

The internal consistency measures of the scaled rank orders produced by the five groups are shown in Table 1. For all five groups, the Rasch sample separation reliability was acceptably high. From a total of 224 judges across the five groups, there were only nine misfitting judges, suggesting that judges within each group behaved with acceptably high mutual consistency. Most of the misfitting judges (7 out of 9) were students, which is unsurprising given the large number of participants in the two peer groups. The large number of student participants also meant that each contributed fewer judgements (after randomly selecting the 1217 judgements per group, this ranged from 0 to 29) than did the experts and novices. Therefore the peer misfit figures were based on far fewer data points than those of the experts and novices, and so are likely to be more error prone.

Across the five groups, the number of misfitting scripts ranged from 4 to 7 out of a total of 168 scripts, suggesting that the scripts were judged with acceptable consistency by the judges within each group. Overall, these three measures strongly suggest that the five rank orders were of acceptably high internal consistency.

***** TABLE 1 HERE ******

Reliability and validity of peer assessment

To measure the inter-rater reliability of the expert and peer assessments we used the Pearson product-moment correlation coefficient. Our first predictions were borne out by this analysis. For the two expert groups, the inter-rater reliability was high (r = .86), indicating that the expert assessments can be used as a baseline for measuring the validity of the peer assessments. The inter-rater reliability was also high for the two peer groups (r = .72), though, as predicted, significantly less so than the inter-rater reliability for the experts (Fisher r-to-z transformation $z = 3.50$, $p < .001$).

To measure the validity of the peers' assessments, we took three approaches, as outlined in the research design section. For the first two of these, we needed a mean peer score and a mean expert score for each script, which we obtained directly from the scaled rank ordering produced by the Rasch analysis. We then proceeded as follows.

First, we used the Pearson product-moment correlation coefficient to compare the mean peer and mean expert scores. By this measure, validity was high (r = .77), in line with our prediction.

Second, we correlated outcomes from the CJ assessment with those of other summative tests from the module. For this purpose we obtained five assessment scores (two computer-based multiple-choice tests, two courseworks, one written test) for 147 of the 168 students. To provide a baseline for comparison, we calculated the mean Pearson product-moment correlation coefficient between the five assessment scores, which was .35. The mean correlation coefficient between the mean expert scores and the five assessment scores was similar (r = .31, $p < .001$, two-tailed). The mean correlation coefficient between the mean peer scores and the five assessment scores was lower, but still significant (r = .20, p = .015, two-tailed). We note that all of these correlations are lower than might be expected, which might be due to the fact that the different assessments measure different calculus-related skills. Nevertheless, these outcomes support the view that peer assessment led to valid results.

Third, we investigated the possibility that the peer judgements were based on non-mathematical criteria by compared the expert, peer and novice assessments. Validity for the novice group was measured by correlating their assessments with the mean expert scores. The novice validity was higher than expected ($r = .64$), although significantly lower than the peer validity (Steiger's $z = 3.33$, $p < .001$). The novices' unexpectedly strong performance suggests that surface features of the students' work, such as the length, neatness and layout of responses, correlate with conceptual understanding. To explore this possibility, we conducted a further analysis and considered responses from our interview data.

Our further analysis involved ranking the students' responses, which were scanned as single-page pdfs, by file size, to provide a crude approximation of response length. The Spearman rank-order correlation coefficient between file size and the expert assessments was .45, and the Spearman coefficient between file size and the novice assessments was also .45. This suggests the validity measure achieved by the novices may indeed have been due to a correlation between mathematical quality of the responses and surface features such as length.

The post-main-study interview data shed further light on this issue. A researcher presented interviewees with pairs of student responses and asked them to decide which demonstrated the better conceptual understanding, and to explain their decisions. We found that experts, and to a lesser extent peers, justified decisions in terms of mathematical understanding as well as presentation, whereas novices justified decisions in terms presentation only, as exemplified in Table 2.

**** TABLE 2 HERE ****

Discussion

Logistical issues and limitations

The implementation of peer assessment within a large mathematics module was broadly successful, but some logistical issues did arise. We discuss these issues briefly here, illustrating where appropriate with information from the interviews and online feedback, and noting likely effects on our reliability and validity measures.

First, some students complained about the illegibility of the responses on their computer screens. One student commented in the online feedback, "I think it is extremely unfair that some of the scripts were very difficult to read since they were blurry". We believe this arose because the *e-scape* system allows the user to toggle between low and high resolution modes, with low resolution as the default. Participants were instructed to switch to high resolution before starting their assessments, but some students may not have done so because the resolution toggle button is not prominent or intuitively obvious within the interface.

Second, the groups received different amounts of training. The expert and novice participants received up to 30 minutes of website training and assessing practice in workshops, small groups or individually. In contrast, due to practical constraints, the peers received a five-minute demonstration in a lecture of around 200 students (this

may explain why some peers may not have remembered to switch to high-resolution mode before beginning).

Third, the experts and novices completed far more judgements each (100 or more) than the peers (20). The reason for this difference was that we wanted the peer assessment exercise to take each student up to an hour, and we estimated this at about 20 judgements per student. However we had far fewer experts and novices, and required them to complete at least 100 judgements each to ensure we had enough data to construct rank orders. As such, the experts and novices had substantially more opportunity to develop experience with making pairwise comparisons.

These issues would be worth consideration in future studies: one could ensure that training in using the comparative judgement software is equivalent across all groups and provide students with judging practice. However, we note that the existence of these issues do not undermine our main results on reliability and validity of peer assessment, since they are all likely to have deflated rather than inflated our measures of these constructs.

Student perceptions

Because our study was designed to assess feasibility of peer assessment, eventual student grades were awarded on the basis of expert judgements. However, in general it would be desirable for peer assessment to form part of the awarded grades. Thus it is important to consider student perceptions of the fairness of this approach, which we do now.

First, predictably and sensibly, some students were concerned that some of their peers lacked the understanding to act as assessors. One articulate student expressed this concern as follows:

> "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, because that's what they believe the answer is."

This is astute observation. If the peers' collective understanding of the construct is weak, then they cannot be expected to perform adequately as assessors. Whether or not peer assessment validity is high enough is of course a matter of judgement.

Second, some students expressed concern about how seriously their peers took the exercise. For example, one commented in the online feedback that "I do feel that some people may not have to judged the tests accurately as it made no difference to there (sic.) work". One tutor informed us a student reported disliking the exercise and deliberately chose the poorer response each time in order to undermine the outcomes. In fact, our conservative approach, basing grades on expert judgements, may have contributed to such problems: because students' grades were entirely dependent on the quality of their written response and there was no extrinsic motivation to take the comparative judgement component seriously. Ensuring that learners are motivated to take peer assessment seriously is an important design principle (e.g. Cho, Schunn and

Wilson 2006; Topping 2010). In future studies, students could receive a modification to their grade based on the consistency of their judgements with the final rank order.

Third, students felt uncomfortable if presented with their own work onscreen (recall that we removed all such occurrences from the analysis), or if they felt they recognised a friend's handwriting. In future studies, an obvious improvement would be to ensure students are never presented with their own scripts.

These concerns are no doubt worthy of consideration. However, they should also be balanced against the potential of CJ for enhancing student learning. As discussed in the introduction, a pedagogical motivation for implementing CJ in this context was to allow students to reflect on their own conceptual knowledge and mathematical communication. Feedback from the students indicates that, in keeping with other findings in the literature (e.g. Dochy, Segers and Sluijsmans 1999), many did find the process challenging but worthwhile. For example, one student commented:

> "It is hard to judge other people's work. It is always the lecturers who judge whose work is better. It is a very good idea to get students to think about how hard it is to judge people's work. Sometimes we as student 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. It is a good experience for me."

Of course, self-report cannot be relied upon to establish whether higher-order learning has taken place (e.g. Russo, Johnson and Stephens 1989). A key theoretical focus of our future work will be to measure the learning gains associated with using CJ for peer assessment.

Conclusion

In this study we investigated whether students were able to assess their peers' conceptual understanding of advanced mathematics reliably and validly in the absence of assessment criteria. We found that the students performed well, and our results compare favourably to those summarised in the meta-analysis by Falchikov and Goldfinch (2000). Further, the high inter-rater reliability of the two expert groups provides assurance that assessment in the absence of criteria can be a useful approach. Indeed, the expert inter-rater reliability was higher than might be expected for an unstructured written task assessed against criteria (e.g. Murphy 1982; Newton 1996).

This finding stands in contrast to empirical results reported in the peer assessment literature, which support a correlation between the use of clearly understood criteria and good validity and reliability. Our theoretical claim is that CJ enabled peers to perform well as assessors and that this was due to the absence of criteria rather than despite the absence of criteria. The explanation for our result is that CJ draws on the psychological principle that human beings are better at compare one object against another than they are at comparing an object against specified criteria (Thurstone 1927). Moreover, CJ can be expected to be particularly advantageous for difficult-to-specify constructs such as conceptual understanding of advanced mathematical ideas, as was the case here.

Our purpose is not to argue against assessment criteria, which are very important to achieving good peer assessment outcomes in many contexts. However, there may be other contexts in which assessment criteria are not appropriate, and where their absence can result in sound peer assessment outcomes. We believe we have reported a study set in one such context here.

Acknowledgements
This work was funded by a Royal Society Shuttleworth Education Fellowship and a grant from the HE-STEM.

References

Boud, D., R. Cohen, and J. Sampson. 1999. Peer learning and assessment. *Assessment and Evaluation in Higher Education* 24: 413–426.

Bramley, T. 2007. Paired comparison methods. In *Techniques for Monitoring the Comparability of Examination Standards*, ed. Newton, P., J-A. Baird, H. Goldstein, H. Patrick, and P. Tymms 264–294. London: QCA.

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

Bramley, T., and T. Gill. 2010. Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education* 25: 293–317.

Chang, C. C., K. H. Tseng, P. N. Chou, and Y. H. Chen. 2011. Reliability and validity of Web-based portfolio peer assessment: A case study for a senior high school's students taking computer course. *Computers and Education* 57: 1306–1316.

Cho, K., C. D. Schunn, and R. W. Wilson. 2006. Reliability and validity of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology* 98: 891–901.

Derrick, K. 2012. Developing the e-scape software system. *International Journal of Technology and Design Education* 22: 171–185.

Dochy, F., M. Segers, and D. Sluijsmans. 1999. The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education* 24: 331–350.

Falchikov, N., and J. Goldfinch. 2000. Student peer assessment in Higher Education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research* 70: 287–322.

Gielen, S., F. Dochy, P. Onghena, K. Struyven, and S. Smeets. 2011. Goals of peer assessment and their associated quality concepts. *Studies in Higher Education* 36: 719–735.

Kimbell, R. 2012. Evolving project e-scape for national assessment. *International Journal of Technology and Design Education* 22: 135–155.

Kollar, I., and F. Fischer. 2010. Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction* 20: 344–348.

Laming, D. 1984. The relativity of "absolute" judgements. *British Journal of Mathematical and Statistical Psychology* 37: 152-183.

Laming, D. 2004. Marking university examinations: some lessons from psychophysics. *Psychology Learning and Teaching* 3: 89–96.

Levesley, J. 2011 Taking control of the assessment agenda. In *Report of the HE Mathematics Curriculum Summit*, ed. P. Rowlett, 21–23. York: Higher Education Academy.

Magin, D. J. 2001a. A novel technique for comparing the reliability of multiple peer assessments with that of single teacher assessments of group process work. *Assessment and Evaluation in Higher Education* 26: 139–152.

Magin, D. J., and P. Helmore. 2001b. Peer and teacher assessments of oral presentation skills: how reliable are they? *Studies in Higher Education* 26: 287–298.

Murphy, R. 1982. A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology* 52: 58–63.

Newton, P. 1996. The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal* 22: 405–420.

Orsmond, P., S. Merry, and K. Reiling. 2000. The use of student derived marking criteria in peer and self-assessment. *Assessment and Evaluation in Higher Education* 25: 23–38.

Orsmond, P., S. Merry, and K. Reiling. 1996. The importance of marking criteria in the use of peer assessment. *Assessment and Evaluation in Higher Education* 21: 239–250.

Pachur, T., and H. Olsson. 2012. Type of learning task impacts performance and strategy selection in decision making. *Cognitive Psychology* 65: 207–240.

Pollitt, A. 2012. The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice* 19: 281–300.

Russo, J. E., E. J. Johnson, and D. L. Stephens. 1989. The validity of verbal protocols. *Memory and Cognition* 17: 759–769.

Seery, N., D. Canty, and P. Phelan. 2012. The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design Education* 22: 205–226.

Slavin, R. 1991. Synthesis of research of cooperative learning. *Educational Leadership* 48: 71–82.

Sluijsmans, D., S. Brand-Gruwel, J. van Merriënboer, and R. Martens, R. 2004. Training teachers in peer-assessment skills: Effects on performance and perceptions. *Innovations in Education and Teaching International* 41: 59–78.

Swan, M., and H. Burkhardt. 2012. Designing assessment of performance in mathematics. *Educational Designer* 2: 1–40.

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., E. F. Smith, I. Swanson, and A. Elliot. 2000. Formative peer assessment of academic writing between postgraduate students. *Assessment and Evaluation in Higher Education*, 25: 149–169.

Topping, K. 2003. Self and Peer Assessment in School and University: Reliability, Validity and Utility. In *Optimising New Modes of Assessment: In Search of Qualities and Standards*, ed. M. Segers, F. Dochy, and E. Cascallar, Vol. 1, 55–87. Dordrecht: Kluwer Academic Publishers.

Topping, K. 2010. Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction* 20: 339–343.

Tsivitanidou, O. E., Zacharia, Z. C., and Hovardas, T. 2011. Investigating secondary school students' unmediated peer assessment skills. *Learning and Instruction* 21: 506–519.

van den Berg, I., W. Admiraal, and A. Pilot. 2006. Design principles and outcomes of peer assessment in higher education. *Studies in Higher Education* 31: 341–356.

van Steendam, E., G. Rijlaarsdam, L. Sercu, and H. van den Bergh. 2010. The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learning and Instruction* 20: 316–327.

van Zundert, M., D. Sluijsmans, and J. van Merriënboer. 2010. Effective peer assessment processes: Research findings and future directions. *Learning and Instruction* 20: 270–279.

Figure 1: The test question used in the study.

Conceptual Test Question

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

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.

Table 1: Internal measures of the scaled rank orders produced by each group.

	Peer 1	*Peer 2*	*Expert 1*	*Expert 2*	*Novice*
Rasch sample separation reliability	.73	.86	.93	.89	.97
Total judges	100	93	11	11	9
Misfitting judges	3	4	0	1	1
Misfitting scripts (out of 168)	6	6	7	4	6

Table 2: Example justifications for judgement decisions by an expert, peer and novice.

Expert	[Response E] is very, very good. Very nice illustration. His diagram, his table of values is very good. [Response F] showed the contour plot, but his concept of continuity was not portrayed here. He said the function was continuous on all points, which is not true.
Peer	I prefer the layout of [Response C], it is easy to read and easy to understand. However, the content in Response D is more accurate. I have seen the function already so it is easier to understand the context. On [Response C] it has a picture, and describes it pretty well. What [Response D] is saying and describing is more accurate I think.
Novice	I was looking at the order. I think [Response B] used diagrams more clearly to illustrate what he was talking about. While I don't really understand the maths of it, it helped me to understand more what he was trying to achieve, and I find it more of a strategic structure.