Peer assessment using comparative and absolute judgement

Ian Jones

Mathematics Education Centre, Loughborough University

Chris Wheadon

No More Marking Ltd.

Author note

Ian Jones, Mathematics Education Centre, Lougborough University; Chris Wheadon, No More Marking Ltd., Guildford UK.

Correspondence concerning this article should be addressed to Ian Jones,

Mathematics Education Centre, Schofield Building, Loughborough University,

Loughborough, LE11 3TU, UK. Email: I.Jones@lboro.ac.uk

ABSTRACT

Peer assessment exercises yield varied reliability and validity. To maximise reliability and validity, the literature recommends adopting various design principles including the use of explicit assessment criteria. Counter to this literature, we report a peer assessment exercise in which criteria were deliberately avoided yet acceptable reliability and validity were achieved. Based on this finding, we make two arguments. First, the comparative judgement approach adopted can be applied successfully in different contexts, including higher education and secondary school. Second, the success was due to this approach; an alternative technique based on absolute judgement yielded poor reliability and validity. We conclude that sound outcomes are achievable without assessment criteria, but success depends on how the peer assessment activity is designed.

INTRODUCTION

Assessment involves judging a student's achievement within a subject domain on the basis of a piece of evidence such as a test response. Peer assessment is an arrangement in which students are required to make this judgement about other students (Falchikov & Goldfinch, 2000; Topping, 2010). There exist a broad range of motivations for implementing peer assessment, as well as purposes to which peer assessment outcomes are applied. Gielen, Dochy, Onghena, Struyven and Smeets (2011) listed five common goals of peer assessment: as a social control tool; as an assessment tool; as a learning tool; as a 'learn-how-to-assess-tool'; as an active participation tool. In this paper our focus is on peer assessment as an assessment tool. Gielen et al. state that this goal usually involves a focus on validity and reliability. Moreover, Kane (2013) argues that investigating validity should take account of the purposes of an assessment. In the peer assessment literature, investigating the validity of an assessment tool typically involves comparing peer outcomes to those of teachers or other experts, and to existing achievement data. Reliability is typically measured by comparing the outcomes of two or more groups of peers undertaking the assessment activity independently. These approaches to evaluation were adopted here. Longer-term goals for this programme of research are peer assessment as a learning tool and as an active participation tool. However, these were not explicit goals for the study reported here, although we consider their implications in the discussion.

There are published design principles recommending how best to ensure particular goals are realised and evaluated (Dochy, Segers & Sluijsmans, 1999; Falchikov & Goldfinch, 2000; Topping, 2003; van Zundert, Sluijsmans & van Merriënboer, 2010). These principles include clarifying goals, training students on how to assess, and familiarising students with explicit and detailed assessment criteria. The focus here is on the latter: the role of assessment criteria for securing valid and reliable outcomes of a peer assessment activity. Our results suggest that there are contexts in which this recommendation does not apply, but only if the assessment procedure adopts a carefully-designed *comparative* approach to peer assessment.

The role of criteria

The literature provides numerous examples of criteria that might be used in peer assessment, including those generated by students. Sadler and Good (2006), for instance, reported seventh-grade student-generated criteria for peer marking of a biology test. They provided example criteria for one of the test items as follows, with each bullet worth two marks.

> *Compare and contrast the classification systems of Aristotle and Linnaeus.*
>
> Similarity:
>
> - used only observable characteristics of organisms.
>
> Differences:
>
> - Aristotle used where animals live (air, land, water) or plant size and structure;
> - Linnaeus used body structure, color, ways of getting food;
> - Linnaeus named using binomial nomenclature: genus-species in Latin;
> - Linnaeus used many taxonomic levels: Kingdom, phylum or division, class, order, family, genus, species. (p. 12)

The students in the Sadler and Good study were experienced at generating and applying marking criteria, and the criteria were displayed on classroom walls during the peer marking exercise. The student scores were used to allocate a grade (A to E) to each test, and the tests were independently graded by a teacher. Sadler and Good found a high correlation between grades awarded by students and the teacher, r = .905. In summary, the authors provided evidence that high agreement between students and teachers is possible, and argued that detailed criteria generated by students who were experienced at assessing peers contributed to the success of the exercise.

More generally, the wider literature makes clear that explicit and well-understood assessment criteria are important for ensuring that peer assessment outcomes are reliable and valid (Chang, Tseng, Chou & Chen, 2011; Dochy, Segers & Sluijsmans, 1999; Falchikov & Goldfinch, 2000; Orsmond, Merry & Reiling, 1996; Topping, 2003). This is often stated in no uncertain terms. Dochy, Segers and Sluijsmans

(1999, p. 342), for example, wrote that the "development of criteria through active cooperation between teacher and students seems to be a critical success factor for self- and peer-assessment". Orsmond, Merry and Reiling (1996) entitled a paper "The importance of marking criteria in the use of peer assessment". Falchikov and Goldfinch (2000, p. 292) considered study designs to be faulty where "students [were] not provided with criteria or structure", and were expected instead to provide a "global rating". Similarly, Topping (2009) emphasised the need to "involve participants in developing and clarifying assessment criteria" (p. 25).

We argue here that the importance of explicit criteria for producing sound peer assessment outcomes is overstated. There are two grounds to this argument. First, the data reported in a widely cited meta-analysis by Falchikov and Goldfinch (2000) are, on closer inspection, equivocal on the role of criteria for achieving sound peer assessment outcomes. The authors identified three approaches in the literature: aggregated scores based on individually-marked criteria as exemplified above; global judgements informed by detailed criteria; and global judgements without criteria. Falchikov and Goldfinch compared the means of the reported correlations between peer and tutor assessment outcomes for each approach. They found a high mean correlation for studies that used global judgements with criteria ($N = 18$, $r = .77$)[1] or global judgements without criteria ($N = 17$, $r = .72$), and a lower mean correlation for studies that used aggregated scores of individually marked criteria ($N = 18$, $r = .53$). They also compared the mean effect sizes (Cohen's d) of the three approaches, based on the means of marks produced by peers and tutors, where the smaller the effect size the better the agreement between the assessment outcomes of peers and tutors. Peers assessed more harshly than tutors (negative effect size) when using global judgements without criteria ($N = 2$, $d = -.32$), and more generously when using global judgements with criteria ($N = 10$, $d = .17$); peers and tutors were in close agreement when using aggregated judgements across discrete criteria ($N = 12$, $d = .03$). The effect size analysis does support the use of discrete criteria, but Falchikov and Goldfinch acknowledged the small number of studies involved, notably only two studies for global judgements without criteria. Moreover, they excluded a problematic study (Butcher, Stefani & Tariq, 1995) from the effect size analysis and noted that

[1] The reported mean correlation for global judgements with criteria excluded a problematic study by Bumett and Cavaye (1980). When this study was included the mean correlation was higher, $r = .85$.

aggregated judgements resulted in the largest mean effect size when it was included ($N = 13$, $d = .34$). In sum then, a case can be made based on the correlational analysis that aggregated judgements across discrete criteria are inferior to global judgements with or without criteria. Conversely, a case can be made based on effect size analysis that discrete criteria are markedly superior and global judgements without criteria are markedly inferior. As such, the evidence provided by Falchikov and Goldfinch is equivocal regarding the role and nature of criteria in peer assessment studies.

Our second reason for questioning the importance of explicit criteria is a study by Jones and Alcock (2014) that investigated a novel approach to using global judgements without criteria. 193 mathematics undergraduates peer-assessed a conceptual calculus test using a technology-enabled comparative judgement technique, described later. The peer assessment outcomes were compared with those of 20 expert mathematicians who assessed the same test responses using the same technique. The correlation ($r = .77$) was higher than the overall mean reported in the meta-analysis of Falchikov and Goldfinch ($N = 56$, $r_m = .69$)[2], supporting the validity of the outcomes. The inter-rater reliability of the peer assessment outcomes were estimated and also found to be acceptable ($r = .72$). Jones and Alcock argued that if assessment arrangements are devised appropriately and carefully, good outcomes can be achieved without criteria. More generally, there may be contexts in which the aims of a peer assessment exercise are best served using global judgements without criteria.

RESEARCH AIMS

In this article we set out to replicate and extend the findings of Jones and Alcock (2014). We report a study in which secondary school students undertook a computer-based peer assessment exercise in comparative and absolute judgement conditions, and the inter-rater reliability and validity of the outcomes were estimated. There were empirical and theoretical motivations to the research.

The empirical motivation was to explore whether the assessment measures reported for the case of undergraduates' understanding of calculus (Jones & Alcock, 2014)

[2] The nature of judgement (global/dimension and criteria/no criteria) was not specified in three studies, hence this overall correlation coefficient is based on 56 rather than 53 studies.

were replicable for lower secondary students' understanding of fractions. This motivation is consistent with Topping's (2010) call for further research into how the arrangement of peer assessment interventions interacts with outcomes; little is known about how the "age and nature of institution of participants" (p.342) might impact on peer assessment activities. The successful replication of Jones and Alcock's main findings for secondary school students would provide support as to the generality of the approach. Given the successful use of comparative judgement in a variety of educational contexts (Bramley, 2007; Heldsinger & Humphry, 2013; Kimbell, 2012; Seery, Canty & Phelan, 2012) we predicted that Jones and Alcock's results would be replicated.

The theoretical motivation was to investigate the extent to which the sound assessment measures reported by Jones and Alcock can be attributed to the particular comparative judgement technique, described below, rather than to some alternative implementation of global judgements without criteria. To this end we used an experimental design in which students were allocated to a *comparative* or *absolute* judgement condition, and the assessment outcomes for each group were compared to those of experts and student achievement data. Given the long-standing theoretical rationale for comparative judgement, described in the next section, we predicted that students in the absolute judgement condition would not produce sound assessment outcomes.

It should be emphasised that the use of the comparative and absolute conditions was not intended as a test between two competing approaches to peer assessment. We do not advocate implementing the absolute condition described here for practical assessment purposes, nor do we consider it a proxy for non-comparative approaches to peer assessment reported in the literature. Rather, the purpose was to investigate experimentally the role that comparison plays in yielding sound outcomes in the absence of explicit and detailed criteria. This required a control condition as closely matched as possible to the comparative condition except for the use of absolute judgements.

In the following section we describe the technology-enabled comparative judgement technique adopted, setting out its theoretical rationale and highlighting those attributes that make it particularly promising for use in peer assessment exercises.

COMPARATIVE JUDGEMENT

Comparative judgement has a long history as a method for constructing psychological scales of sense stimuli (Thurstone, 1927). A key finding is that human beings are very reliable at comparing one stimulus with another, but very poor at judging the value of a single stimulus in isolation (Laming, 2004a). However, comparative judgement has only recently become a viable option for educational assessment due to technological advances that enable the automation of the required logistics and statistical modelling (Pollitt, 2012).

The mechanics of the comparative judgement approach to peer assessment are simple. Students first sit a test, which in the research reported here was a question designed to elicit conceptual understanding of fractions, shown in Figure 1. The test responses are anonymised, scanned and uploaded to a website (two example test responses are shown in Figure 2). Students are then presented with pairs of responses via an internet browser and have to decide which of the two is "better" according to some agreed global descriptor, in this case "understanding of fractions". Ties are not allowed, and the students are required to complete a certain number of pairwise judgement decisions. Once complete, the judgement decisions are statistically modeled to generate a parameter estimate and standard error for each test response (Bramley, 2007; Pollitt, 2012). The parameter estimates and standard errors can then be used to construct a scaled rank order of the test responses (see Figure 3).

Write down these fractions in order of size from smallest to largest.
Underneath, describe and explain your method for doing this.

$\frac{3}{4}$ $\frac{3}{8}$ $\frac{2}{5}$ $\frac{8}{10}$ $\frac{1}{4}$ $\frac{1}{25}$ $\frac{1}{8}$

Figure 1: The test question used to generate responses for the peer assessment exercise.

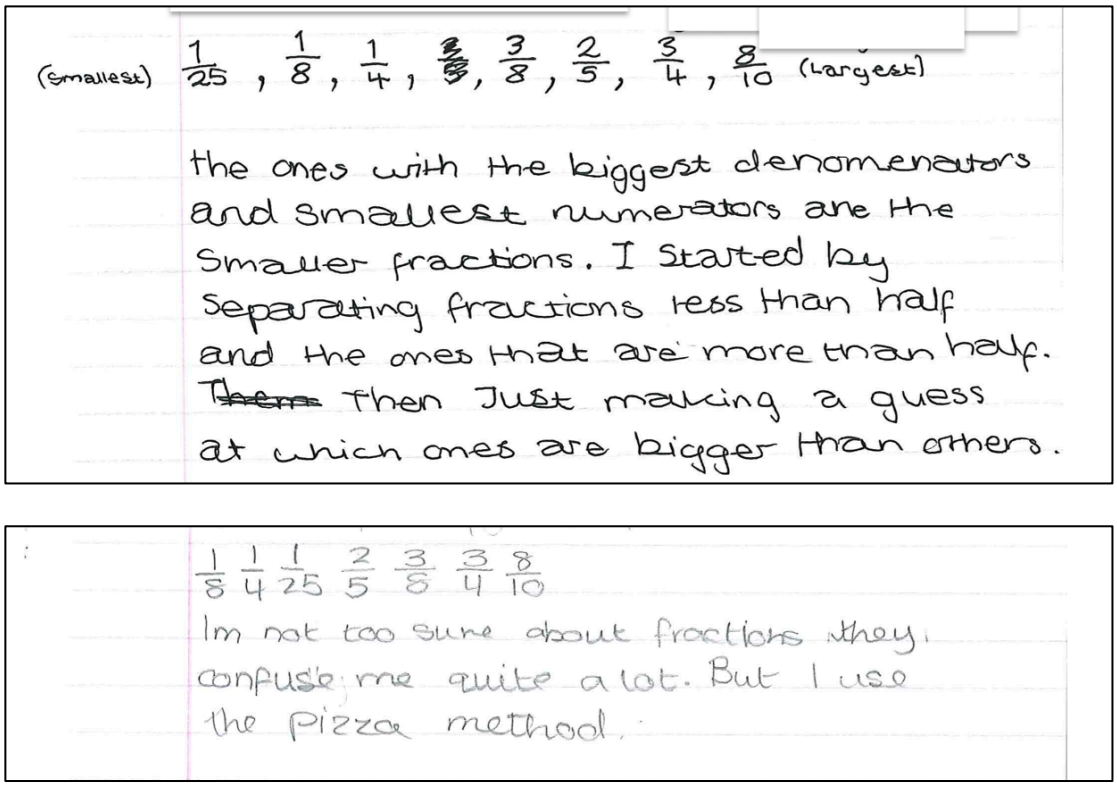

Figure 2: Two example test responses to the question shown in Figure 1.

Comparative judgement does not require explicit, detailed assessment criteria. Validity is therefore grounded not in detailed rubrics or marking schemes, but rather in assessors' collective interpretation of a global descriptor. Jones and Alcock (2014) claimed that the comparative judgement approach to validity offers three potential affordances for peer assessment exercises.

First, it is 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 (e.g. Anderson, 2014; Suto, 2013; Walport et al., 2010), but their assessment tends to yield relatively poor inter-rater reliability and validity (Chang et al., 2011; Laming, 2004b; Orsmond et al., 1996; Topping, Smith, Swanson & Elliot, 2000; Tsivitanidou, Zacharia & Hovardas, 2011). Comparative judgement might help overcome this difficulty because it makes no attempt to capture nebulous constructs to the letter, and instead assumes that a given construct can be recognised by assessors ("you know it when you see it") in terms of a global descriptor.

A second and related potential benefit is that comparative judgement is suited to assessing a wide variety of unpredictable responses that would be difficult to anticipate comprehensively and precisely in rubrics. Even if two test responses are disparate in content or style, then so long as both provide evidence relevant to the global descriptor, a pairwise judgement decision can be made. This is helpful in certain assessment contexts because the kinds of nebulous constructs discussed above generally lend themselves to open-ended, unstructured test questions (Swan & Burkhardt, 2012).

A third potential benefit is that a peer assessment exercise using comparative judgement can be conducted without training. No prior experience is necessary and students need not develop or familiarise themselves with detailed or multidimensional assessment criteria. This potential benefit goes against the finding that more robust outcomes are associated with previous experience and training (Topping, 2009; Tsivitanidou et al., 2011; van Zundert, Sluijsmans & van Merriënboer, 2010). Of course, there are many contexts in which the development and understanding of such criteria is an inherently valuable learning activity (e.g. Hunter & Russ, 1996; Sadler & Good, 2006), but there may exist other contexts in which it is desirable or at least pragmatically useful for students to assess peers' work without the need for preliminary training.

METHOD

Participants

A total of 157 students participated in the peer assessment exercise, drawn from two urban schools in central England (school A and school B) and a rural school in the south of England (school C). The participants in schools A and C were students aged 14 and 15, and in school B the participants were students aged 13 and 14. The students in all schools were streamed for mathematics, and those in the lowest achieving class in each school were omitted. This was due to teacher concerns that some students would struggle with the literacy demands of the assessment task.

Materials

The participants assessed 24 responses to the test question shown in Figure 1. The test responses were obtained from students in England aged 12 to 15 years as part of a previous study (Jones, Inglis, Gilmore & Hodgen, 2013), and two examples are shown in Figure 2. Standardised test data revealed these students ranged from low to high achievers in mathematics. The responses came from students in a school not involved in the present study, and the students who produced them were not involved in the peer assessment exercise.

Procedure

In each school the students were randomly allocated into two groups, and each group received either the comparative or the absolute judgement condition, detailed below. The break down of groups by schools is shown in Table 1. In schools A and C the comparative and absolute groups were taught in two separate lessons. In school B this was not possible due to timetabling constraints, and instead there were two lessons containing participants allocated to each condition.

	comparative	absolute
School A	22	23
School B	27	31
School C	27	27

Table 1: The number of students in each school in each condition involved in the peer assessment exercise.

Each lesson was taught by the same member of the research team (referred to as the 'teacher' in the remainder of this article). In school A, a second member of the research team was present as an observer, and for most of the lessons the head of the school's mathematics department was present too. In school B a different member of the research team was present as an observer, as well as a mathematics teacher from the school. In school C the head of the school's mathematics department was present as an observer.

In each lesson the students were first shown a test question (Figure 1) and told it had been administered to students in another school. The students were told that they were going to play the role of examiners and assess the responses of students from the other school. The teacher then demonstrated the website in either the comparative or absolute assessment mode (schools A and C), or both the comparative and absolute mode (school B).

The comparative judgement condition was implemented using the www.nomoremarking.com website. Each student accessed the online judging using a unique url, and was presented with a series of response pairs. A zoom tool was provided for inspecting responses more closely, and "left" and "right" decision buttons were provided for selecting the response judged to reflect the "better understanding of fractions". The absolute condition was similarly implemented using www.nomoremarking.com. Students were presented with a series of single test responses, along with a zoom tool and slider that could be used to adjust a displayed rating number between the values of 0 to 100. They were instructed to use the slider to rate each response in terms of "understanding of fractions", taking 0 to mean "could not be worse" and 100 to mean "could not be better".

The students were instructed to complete at least 24 pairwise judgements (comparative condition) or 24 single judgements (absolute condition) and were allowed up to 30 minutes to do so. The pairings of test responses in the comparative condition and the orders of presentation of responses in both conditions were randomised between students. Due to an error in setting up the comparative condition the judging session ended after 23 judgements were completed, although some students discovered that refreshing the browser enabled them to continue beyond 24 judgements. The mode of comparative judgements completed per student was 23, and the mean was 26.1. Students in the absolute condition tended to judge more slowly and many did not complete all their judgements. The mode of absolute judgements completed per student was 24, and the mean was 22.4.

Students were each given a strip of paper showing the correct order of fractions for reference. Students had a computer each and were told to complete only their own allocation of judgements, but were allowed and encouraged to talk about their work with peers seated nearby. This decision was made on the basis of pilot work in which students had appeared to enjoy judging, and were keen to talk to one another about

their decisions; it was felt this created a positive and stimulating learning environment that should be fostered in the interests of the participants' education. When students completed their allocated judgements they were told to log off and wait until their peers had finished.

To explore validity the outcomes were compared to that of experts, which is a standard procedure for evaluating the validity of peer assessments (Falchikov and Goldfinch, 2000). For the comparative condition, eight mathematics educators (four teachers, two examiners, two research students) assessed the test responses using comparative judgement.[3] Each expert completed 50 pairwise judgement decisions, which were modeled to produce a parameter estimate and standard error for each test response (for full details see Jones et al., 2013). For the absolute condition, an expert mathematics educator assessed the test responses using absolute judgement. For both conditions, the Pearson product-moment correlation coefficient between the peer and expert assessment outcomes was calculated. A second estimate of validity was explored by comparing the peer assessment outcomes in both conditions with existing mathematics achievement data.

ANALYSIS AND RESULTS

We describe the data preparation and present the inter-rater reliability and validity estimates for the two conditions.

Data preparation

Comparative condition. The 76 students in the comparative judgement condition completed a total of 1983 pairwise judgements. These were fitted to the Bradley-Terry model (Bardley & Terry, 1952) using a maximum likelihood estimation procedure (Firth, 2005) to produce a parameter estimate and standard error for each

[3] The experts in fact assessed 25 responses rather than the 24 assessed by the peers, but the additional response was removed before undertaking the validity analysis. There are two stages at which a response can be removed, either by deleting it from the parameter estimates as was the case for the analysis reported here, or removing all of its occurrences from the raw judgement data and recalculating the parameter estimates. A check revealed that both approaches produced equivalent outcomes, r = .997.

test response. The parameter estimates were then used to construct a scaled rank order of responses from "best" to "worst", as shown in Figure 3.

Absolute condition. The 81 students in the absolute judgement condition produced a total of 1812 single rating judgements on a scale of 0 to 100. As mentioned above some of the students ($N = 38$) rated all 24 test responses; others rated between 15 and 23 responses. For each test response a mean rating and standard error of the mean was calculated.

For comparison with the outcome of the comparative condition, a scaled rank order of test responses from "best" to "worst" is shown in Figure 3. A statistical analysis of the differences in outcome between the comparative and absolute conditions is included in the remainder of this section.

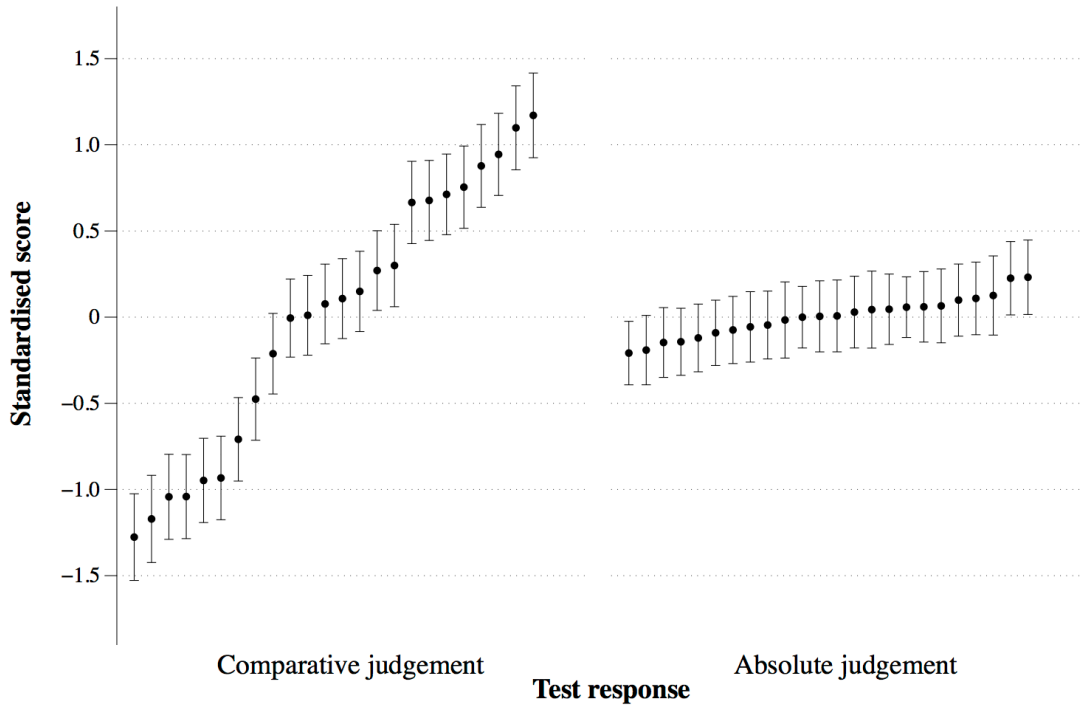

Figure 3: Assessment outcomes of the comparative and absolute conditions, showing parameter estimates of test responses from 'best' (right-most) to 'worst' (left-most). Scores have been converted to standardised *z*-scores and displayed on the same scale for ease of comparison across the conditions. Error bars show standard estimation errors of the statistical modelling procedure (comparative condition) and standard errors of the mean (absolute condition). Test responses are not necessarily in the same order in each graph.

Inter-rater reliability

Comparative condition. The inter-rater reliability estimate provides a measure of the extent to which the same scaled rank order was produced by different groups of peers drawn from the same population. The ideal method for estimating inter-rater reliability is to allocate groupings randomly across all participants, but in our case this would mean comparing possibly non-independent assessment outcomes due to discussion being encouraged during the experimental lessons. Therefore we instead created three groups of peers corresponding to the three schools involved in the study, thereby producing three sets of independent outcomes. The students within each school were able to communicate with each other during the judging, and so this compromise is more likely to deflate rather than inflate the inter-rater reliability estimate; also, comparing the peer assessments of classes from different schools is of direct practical interest even if not theoretically ideal. The judgement decisions of the students in each school (N_A = 720, N_B = 549 and N_C = 714) were statistically modeled to produce three scaled rank orders. The Pearson product-moment correlation coefficients between the three sets of parameter estimates were high, r_{AB} = .93, r_{AC} = .82 and r_{BC} = .85 (scatterplots are shown in Figure 4), suggesting the outcomes were reliable in the sense that they were not dependent on the particular students who undertook the judging.

Absolute condition. A typical technique for estimating the inter-reliability of a peer assessment exercise is to calculate the correlation matrix of scores across different peer assessors (Magin & Helmore, 2001; Topping, 2003). However, this was not appropriate for the present study because fewer than half of the peer assessors (38 out of 81) produced ratings for all 24 test responses, so a correlation matrix would exclude over half the available data. Instead we grouped the students by school and used a technique that enabled the inclusion of all the peers' judgement decisions. This avoided inflating the reliability estimate due to the possible non-independence of within-lesson judgements as discussed above. It also had the advantage of producing reliability estimates directly comparable to the estimates calculated above for the comparative judgement condition.

To estimate inter-rater reliability we calculated a mean rating for the 24 test responses for the total judgements from each school group (N_A = 697, N_B = 520 and N_C = 595). The Pearson product-moment correlation coefficients between the sets of parameter estimates were low and two were not in the expected direction, r_{AB} = -.28, r_{AC} = .39 and r_{BC} = -.17, (scatterplots are shown in Figure 4). These estimates suggest that the absolute condition did not lead to reliable assessment outcomes in the sense that these were not replicable across different groups of peers.

The mean reliability correlation coefficient for the comparative condition was significantly higher than that for the absolute condition, mean *r*s = .86 and -.02 respectively, Fisher's $z = 4.13$, $p < .001$, in line with our predictions.

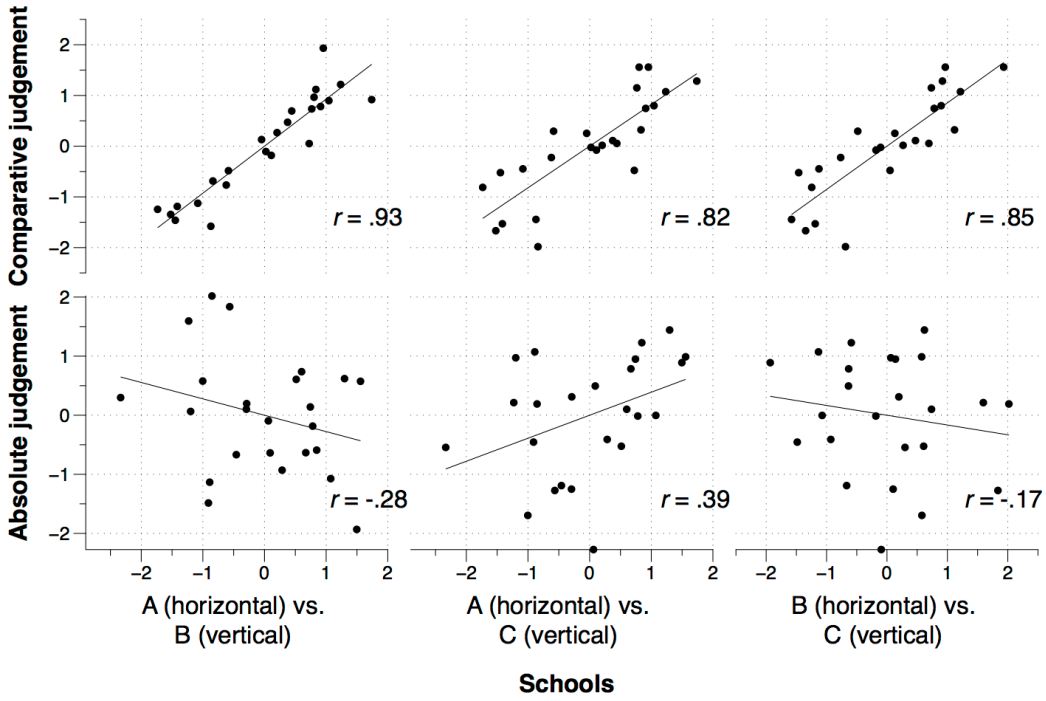

Figure 4: Estimates (standardised *z*-scores) of the inter-rater reliabilities of the peer assessment outcomes across the three schools for the comparative (top) and absolute (bottom) conditions.

Validity

Comparative condition. The validity of the comparative judgement outcome was estimated using two approaches: comparing outcomes to that of experts and to existing mathematics achievement data. The correlation between expert and peer

comparative judgement outcomes was found to be acceptably high, $r = .85$, (a scatterplot is shown in Figure 5), providing support to our claim that the peers assessed the test responses validly. This validity estimate is not significantly different to the expert inter-rater reliability reported by Jones and Alcock (2014), $r = .87$, Fisher's $z = .25$, $p = .401$.

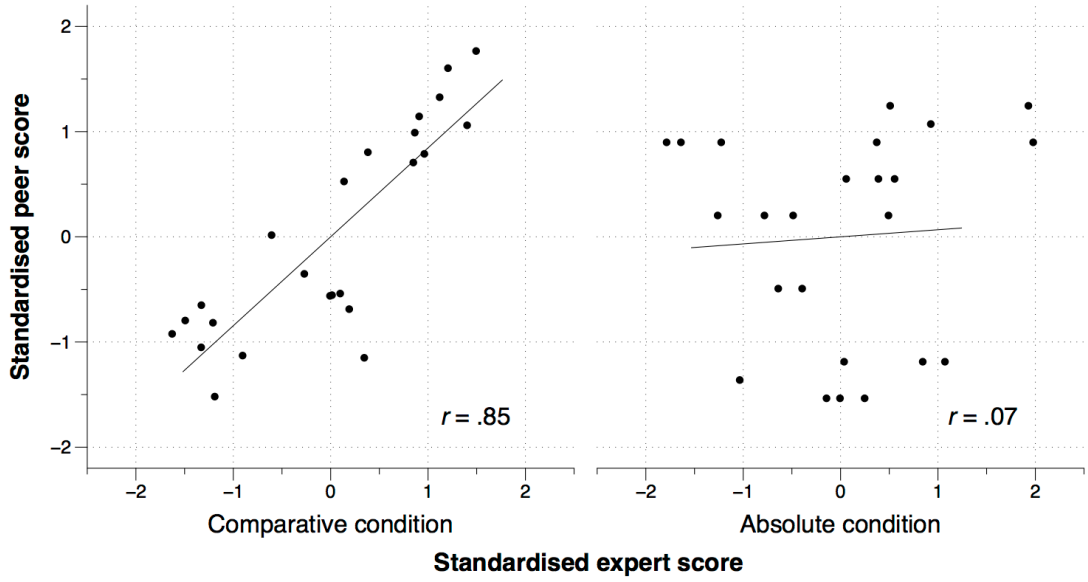

Figure 5: Validity as estimated by the correlation between standardised expert and peer scores for the comparative (left) and absolute conditions.

The second validity estimate was obtained by comparing the peer assessment outcomes to existing mathematics achievement data for the 24 students who had completed the test. This was done separately for the younger students (12 to 13 years, $N = 10$) and the older students (13 to 15 years, $N = 14$) due to the non-comparability of available achievement data. For the younger students, binary achievement scores (high/low) were available and a Mann-Whitney U test was conducted on the parameter estimates between the high- and low-achieving groups. This revealed that the high-achieving group scored significantly higher than the low-achieving group, mean parameter estimates = .328 and -.354 respectively, $p = .038$, supporting the validity of the peer assessment for the test responses of the younger students.

For the older students, predicted external examination results (General Certificate of Secondary Education in Mathematics grades) were available. The grades were converted to a numerical equivalent ranging from 7 (grade A*, highest) through to 1

(grade F, lowest), and Spearman's rank-order correlation coefficient was used to explore the relationship between predicted grades and peer assessment parameter estimates. The correlation coefficient was high, $\rho = .72$, supporting the validity of the peer assessment of the older students.

Taken together, the correlation with expert assessment and the consistency with existing achievement data provide supporting evidence that CJ offers a valid approach to implementing the peer assessment of secondary mathematics work.

Absolute condition. As for the comparative condition, the validity of the peer assessment was estimated by comparing outcomes to that of an expert and to existing mathematics achievement data. The correlation between the peer and expert absolute judgement outcomes was small and non-significant, $r = .07$, $p = .754$, as shown in Figure 5, suggesting that the collective outcome of the peers' assessments were not valid[4]. The correlation coefficient between the assessment outcomes of peers and experts for the comparative condition was significantly higher than the absolute condition, rs = .85 and .07 respectively, Fisher's $z = 3.80$, $p < .001$, in line with our predictions.

Second, the peer assessment outcome was compared to student achievement data. For the test responses of the younger children, a Mann-Whitney U test revealed no significant difference, $p = .610$, between the mean ratings for the high-achieving students and the low-achieving students, mean ratings = 39.13 and 40.81 respectively. For the test responses of the older children, Spearman's rank-order correlation coefficient between predicted exam grades and peer assessment mean ratings was negative and not significant, $\rho = -.03$, $p = 9.15$.

A further aspect of validity was also explored regarding the suitability of comparative judgement for assessing the test question. Comparative judgement is inappropriate for assessments suited to marking and rubrics such as itemised tests, for which very high validity and reliability estimates can be obtained (e.g. Sadler & Good, 2006). In the present study, the first part of the test required students to sequence a list of provided fractions in order of size (see Figure 1), and it is possible the peer assessors judged the

[4] The Pearson product-moment correlation between the expert assessment in the comparative and absolute conditions was high ($r = .70$, $p = .005$). This suggests that unlike students, experts are able to achieve reasonably valid assessments using absolute judgement.

test responses solely or largely based on the correctness of the sequenced fractions. If so, we would expect the "correctness" of the fractions sequence to correlate strongly with comparative judgement outcomes. To explore this we generated a score for each test response using Levenshtein distance (Levenshtein, 1966). A perfectly ordered sequence of fractions was scored 0 (meaning 0 steps would be required to correct the sequence) and a highly unordered sequence was scored 7 (meaning 7 steps would be required to correct it). We found that the students' comparative judgement outcomes correlated marginally with Levenshtein distance (mean = 2.54, SD = 1.84), r = -.40, p = .052, suggesting the students did not judge test responses solely or mainly in terms of the objective part of the question. (For completeness, there was no significant correlation between absolute judgement outcomes and Levenshtein distance, r = -.15, p = .478.) Therefore it seems the test was suitable for comparative judging, although we intend to avoid objectively markable question parts in future work.

DISCUSSION

Students in the comparative judgement condition produced collective peer assessment outcomes that yielded high inter-rater reliability and validity estimates. Those in the absolute judgement condition produced outcomes that yielded poor inter-rater reliability and validity estimates. We discuss these findings in terms of the theoretical and empirical motivations for undertaking the research, and set out recommendations for when and how a comparative judgement approach to peer assessment might be appropriate.

Theoretical motivation

The theoretical motivation for the study was to determine whether the high reliability and validity reported by Jones and Alcock (2014), in the absence of assessment criteria, was due to the use of comparative rather than absolute judgement. This required isolating and contrasting a single variable (Topping, 2010) – comparative versus absolute judging – between randomly allocated groups. We found that the collective peer assessment outcome of students in the comparative condition was psychometrically sound, but that the collective outcome of those in the absolute condition was not.

This result was predicted on the grounds of research making use of comparative judgement techniques in a variety of contexts. The overarching finding from the literature is that human beings perform better when making judgements of stimuli relative to other stimuli than when making judgements of stimuli in isolation. Therefore the comparative condition did not enhance students' ability to make judgements *per se*, rather it exploited the human capacity for making accurate comparisons.

We do not claim that comparative judgement as implemented here is the only way to arrange peer assessment exercises that lack criteria, and variations on the design of both conditions can readily be conceived. For example, the absolute judgement condition could have used a discrete Likert-type scale instead of a slider from 0 to 100, which is effectively continuous. Conversely, various comparative approaches to estimating "quality" values for stimuli has been published in the literature, in peer assessment contexts (e.g. Lejk and Wyvill, 2001), general educational assessment contexts (Black & Bramley, 2008), and psychological research contexts (Goffin & Olson, 2011). Further research is required to explore the possible variation of comparative approaches to using criteria-free global judgements for peer assessment.

Empirical motivation

The empirical motivation for the study was to investiagate whether the sound psychometrics reported by Jones and Alcock (2014) could be replicated with younger learners in a different context to inform the generality of the approach. In Jones and Alcock's study, undergraduate mathematics students peer assessed 168 conceptual calculus tests using a comparative judgement technique. The assessment outcomes yielded a high inter-rater reliability estimate (r = .72), and correlated well with expert assessment outcomes (r = .85). These findings are not significantly different from the results reported here for inter-rater reliability (Fisher's z = 1.66, p = .096), and the correlation between peer and expert assessment outcomes (Fisher's z = 1.02, p = .308). This indicates that the comparative judgement approach to peer assessment can yield high reliability and validity across different age ranges, types of institution and topic domains (albeit within the discipline of mathematics). There is an emphasis on higher education contexts within the peer assessment literature (Dochy, Segers &

Sluijsmans, 1999; Gielen et al. 2011; Falchikov & Goldfinch, 2000; Topping, 1998; van Zundert, Sluijsmans & van Merriënboer, 2010), and the work presented here suggests that peer assessment techniques developed higher education institutions can be relevant to schools. However, we acknowledge that the evidence presented here and in Jones and Alcock (2014) did not include low achieving students. In the present study this was due to teacher concerns about some students' literacy levels, and in Jones and Alcock students were mathematics undergraduates with strong qualifications in mathematics. We might expect that the inclusion of low achieving students would deflate the reported estimates of reliability and validity.

Recommendations

In the remainder of the discussion we set out recommendations for when and how comparative judgement might be deployed for peer assessment based on the findings reported above. We draw where appropriate on recommendations set out in the literature (Falchikov & Boud, 1989; Falchikov & Goldfinch, 2000; Topping, 2009, 2010) to structure the discussion.

Nature of assessment evidence. A comparative judgement approach lends itself to assessment activities in which the evidence of achievement is difficult to assess using traditional methods. In this case these were open-ended tests that invited a range of response-types that would be difficult to anticipate in rubrics. Other examples in the peer assessment literature include ePortfolios (Chang et al., 2011; Tsivitanidou, Zacharia & Hovardas, 2011), posters (Orsmond et al., 1996) and case study reports (Topping et al., 2000). The reliability and validity estimates reported here compare favourably to the use of non-comparative approaches using rubrics and marking for such assessments (Falchikov & Goldfinch, 2000).

Age of students. Much of the literature on peer assessment focuses on higher education contexts, and little is known about how the age of the students involved might impact on outcomes (Topping, 2010).

We found no systematic difference in agreement with expert judgements for students aged 13 to 15 years compared to undergraduates (Jones & Alcock, 2014). However, age of students was not a controlled variable across the two studies and therefore this lack of difference is suggestive rather than conclusive. Possible confounding variables

include type of institution, topic domain and arrangements for the peer assessment activities (in class verses homework).

Complexity of material. Some scholars have suggested that peer assessment outcomes are more reliable and valid the more advanced the level of study (Falchikov & Boud, 1989), although the evidence is mixed (Falchikov & Goldfinch, 2000). We have reported here that a lower-level topic (fractions) and a higher-level topic (calculus) led to similar peer assessment outcomes, supporting the robustness of the comparative judgement approach across different levels of complexity. Again, this lack of difference is suggestive rather than conclusive due to possible confounding variables as mentioned above.

Number of peer assessors. The number of peer assessors per test response may influence the validity of outcomes. It has been suggested both that a greater number of peer assessors leads to greater agreement with experts (Falchikov & Boud, 1989), and conversely that too many peer assessors reduces agreement (Falchikov & Goldfinch, 2000).

In Jones and Alcock (2014), 193 students peer assessed 167 test responses, a ratio of just over one assessor per response, and the modal number of judgements completed was 20 per student; in the present study 76 students peer assessed 24 test responses, a ratio of about three assessors per response, and the modal number of judgements completed was 23 per student. If the three schools are considered separately (see Table 1) then the ratio of peer assessors per assessment, as well as the modal number of judgements per student, are roughly comparable to those of Jones and Alcock. The inter-rater reliability of peer assessors across the three schools reported earlier suggests the same findings would have been reported with just one third of the total assessors. Overall, this suggests that the comparative judgement approach is not sensitive to the number peer assessors involved. (This conclusion assumes a minimum threshold is reached at which each test response receives enough judgements to calculate a scaled rank order, see Bramley, 2007; Pollitt, 2012).

Training of students. The participants were not trained prior to the study, and had not experienced peer assessment activities as part of their normal mathematics lessons. Therefore comparative judgement approaches to peer assessment can yield sound outcomes without the need for training. However, a lack of training is not necessarily

desirable and can itself be a valuable activity both for learning subject content and for learning how to assess (Gielen et al., 2011). In future research we aim to explore how training might improve the reliability and validity of peer assessment outcomes. A possible way to implement training would be to allow students to practice and become familiar with comparatively judging one another's work as part of routine learning and teaching activities.

Varied goals of peer assessment. We have addressed only one aspect of peer assessment, namely as an assessment tool (Gielen et al., 2011). Consequently our focus has been on the psychometric properties of the outcomes. Other common goals include engaging students in assessment processes, improving learning gains and generating personalised feedback. Future work is needed to investigate the potential and drawbacks of comparative judgement towards these and other goals. This will require a substantially different approach to validating comparatively judged peer assessment arrangements (Kane, 2011).

One possible fruitful direction for comparative judgement is peer assessment as a learning tool, engaging students in discussion and consideration of what constitutes a high quality response to an open-ended or ambiguous assessment task. There is cautious optimism for believing that comparative approaches to learning may be more beneficial than absolute approaches (Pachur & Olsson, 2012). However, one possible limitation to this approach is that comparative judgement does not necessarily provide descriptive feedback to students. Comparative judgement studies have provided peer assessors with the opportunity to provide a text-based comment when making each judgement (Seery et al., 2012). However, in our experience that slows the judgement process and produces comments that lack insight. 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.

CONCLUSION

Comparative judgement offers an approach for teachers and researchers wishing to implement peer assessment in contexts in which inter-rater reliability and validity

matter. One such context is when peer assessment outcomes are used for summative purposes. With the advent of Massive Open Online Courses (MOOCs) and other technology-enabled developments in education, it is likely that scenarios in which the student-to-teacher ratio is very large will be increasingly common. Automated assessment is likely to have an important role to play, as might peer assessment if consistently reliable and valid outcomes can be supported. Moreover, the increasing use of technology for learning and teaching has led to interest in the possibility of new modes of assessment, notably multimedia outcomes such as videos and ePortfolios. Although the present study used a paper test, albeit a relatively open-ended test question within the context of mathematics assessment in the UK, comparative judgement has been implemented using digital media rather than paper-based scripts (Davies, Collier & Howe, 2012; Kimbell, 2012). Going forward, a potentially exciting use of comparative judgement is for summative peer assessment of digital artefacts produced by large cohorts of online students. However, further research is required to explore the viability of peers assessing such rich and multifaceted evidence of student achievement.

Acknowledgements

The research reported in this article was funded by the Royal Society. We are grateful to Lara Alcock for her insightful feedback and comments on an early draft of the manuscript.

References

Anderson, R. (2014). *Careers 2020: Making Education Work.* London: Pearson.

Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement, 2*, 451–462.

Black, B., & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education, 23*, 357‑373.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika, 39*, 324–345.

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.

Burnett, W., & Cavaye, G. (1980). Peer assessment by fifth year students of surgery. *Assessment in Higher Education, 5*, 273–278.

Butcher, A. C., Stefani, L. A. J., & Tariq, V. N. (1995). Analysis of peer-, self- and staff-assessment in group project work. *Assessment in Education: Principles, Policy & Practice, 2*, 165‑185.

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

Davies, D., Collier, C., & Howe, A. (2012). Assessing scientific and technological enquiry skills at age 11 using the *e-scape* system. *International Journal of Technology and Design Education, 22*, 247–263.

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

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*, 395–430.

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.

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

Goffin, R. D., & Olson, J. M. (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science, 6*, 48–60.

Hunter, D., & Russ, M. (1996). Peer assessment in performance studies *British Journal of Music Education,* 13, 67–78.

Heldsinger, S. A., & Humphry, S. M. (2013). Using calibrated exemplars in the teacher-assessment of writing: an empirical study. *Educational Research, 55*, 219–235.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*, 1774-1787.

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, Germany: IGPME.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1-73

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

Laming, D. (2004a). *Human Judgment: The Eye of the Beholder.* London: Thompson Learning.

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

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*, 707–710.

Lejk, M., & Wyvill, M. (2001). Peer assessment of contributions to a group project: A comparison of holistic and category-based approaches. *Assessment & Evaluation in Higher Education, 26*, 61–72.

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

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

Pachur, T., & Olsson, H. (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.

Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*, 1–31.

Seery, N., Canty, D., & Phelan, P. (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.

Suto, I. (2013). 21st Century skills: Ancient, ubiquitous, enigmatic? *Research Matters, 15*, 2–8.

Swan, M., & Burkhardt, H. (2012). Designing assessment of performance in mathematics. *Educational Designer, 2*, 1–41.

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

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

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education, 25*, 149–169.

Topping, K (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy, & E, Cascallar (Eds.), Optimising *New Modes of Assessment: In Search of Qualities and Standards* (Vol. 1, pp. 55–87). Dordrecht: Kluwer Academic Publishers.

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*, 339–343.

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

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

Walport, M., Goodfellow, J., McLoughlin, F., Post, M., Sjøvoll, J., Taylor, M., & Waboso, D. (2010). *Science and Mathematics Secondary Education for the 21st Century: Report of the Science and Learning Expert Group.* London: Department for Business, Industry and Skills.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis Rasch Measurement.* MESA Press, Chicago, IL.