

ASSESSING MATHEMATICAL PROBLEM SOLVING USING COMPARATIVE JUDGEMENT

Received: 25 August 2012; Accepted: 26 November 2013

ABSTRACT. There is an increasing demand from employers and universities for school leavers to be able to apply their mathematical knowledge to problem solving in varied and unfamiliar contexts. These aspects are however neglected in most examinations of mathematics and, consequentially, in classroom teaching. One barrier to the inclusion of mathematical problem solving in assessment is that the skills involved are difficult to define and assess objectively. We present two studies that test a method called comparative judgement (CJ) that might be well suited to assessing mathematical problem solving. CJ is an alternative to traditional scoring that is based on collective expert judgements of students' work rather than item-by-item scoring schemes. In study 1, we used CJ to assess traditional mathematics tests and found it performed validly and reliably. In study 2, we used CJ to assess mathematical problem-solving tasks and again found it performed validly and reliably. We discuss the implications of the results for further research and the implications of CJ for the design of mathematical problem-solving tasks.

KEY WORDS: assessment, comparative judgement, examinations, mathematical problem solving, reliability, validity

INTRODUCTION

Mathematical problem-solving skills are increasingly valued and sought after by employers and higher education institutions (CBI, 2006; NCETM, 2009; Vordermann, Porkess, Budd, Dunne & Rahman-Hart, 2011; Walport, Goodfellow, McLoughlin, Post, Sjøvoll, Taylor & Waboso, 2010). A key driver for this demand is that many students leave school unable to apply mathematics to real-world, work-based and advanced study contexts (ACT, 2006; Ofsted, 2008; Toner, 2011). It seems that the mathematical knowledge and skills children spend many years learning in classrooms do not readily transfer to studying further abstract mathematics or using mathematics in the workplace. Even those students who are most successful in terms of performing strongly in school mathematics assessments struggle to apply their learning to novel problem-solving situations (Treilibs, 1979). A widespread contention is that school mathematics teaches students how to pass specific tests and

examinations rather than nurturing a flexible and conceptual understanding of mathematics (Ofsted, 2012). As a consequence, many countries are prioritising the development of problem solving in mathematics curricula and pedagogy (e.g. NGA & CCSSO, 2010; OECD, 2009a; QCA, 2008; Rocard, 2007; Soh, 2008).

One challenge to promoting mathematical problem solving is that it is more difficult to define and assess than the recall of facts and performance of standard procedures that constitute most assessment instruments in mathematics (Black, Burkhardt, Daro, Jones, Lappan, Pead & Stephens, 2012). Past attempts to assess problem solving have often produced fragmentary schemes that assess components rather than holistic performances and complete chains of reasoning. This often results in assessment tasks that purport to assess problem solving, but that in reality only assess how well students can follow a series of structured prompts (e.g. DfE, 2011). In contrast, truly valid assessments of mathematical problem solving require students to carry out varied processes, such as modelling and interpreting, using holistic tasks that are relatively unstructured. A widely regarded approach to designing such tasks is provided by the PISA Assessment Framework (OECD, 2009b). The focus is on “mathematical literacy” (p. 84), which involves using mathematical knowledge flexibly and meaningfully in order to tackle a range of problem types in both abstract and real-world contexts. However, such approaches render the development of precise and objective scoring schemes very difficult (Laming, 1990; Pollitt, 2012; Swan & Burkhardt, 2012).

In this article, we present two studies designed to test an approach to assessing mathematical problem solving that offers an alternative to scoring. The approach, called comparative judgement (CJ), is based on expert judgement of the relative merits of students’ mathematical work. We describe CJ in detail in the first section, before going on to contrast traditional assessment materials with those designed to assess mathematical problem solving. In study 1, we conducted an extreme test of CJ, using it to assess current mathematics exams that do not lend themselves to holistic judging, and comparing the outcome to traditional scoring. In study 2, we apply CJ to the case of innovative assessment tasks to evaluate its feasibility for assessing mathematical problem solving.

COMPARATIVE JUDGEMENT

CJ offers an alternative to traditional scoring for assessing students’ work. A key strength of the approach, in terms of assessing difficult-to-specify

constructs such as mathematical problem solving, is that outcomes are based on the collective expertise of examiners. In other words, its validity is grounded in what is valued by the community of practice within a given discipline.

The basic method is straightforward. Examiners are presented with successive pairs of students' work and asked to decide, for each pair, which student has displayed the greatest proficiency in the domain of interest. Ties are not permitted and the examiners must choose one student's work in preference to the other's. The outcomes of many such pairings presented to several examiners are then used to construct a scaled rank order of students from "least" to "most" proficient, as detailed later. The scaled rank order can then be used to assign grades or for other assessment purposes in the usual manner.

The underlying rationale of using CJ for assessing mathematics derives from Thurstone's (1927) discovery that people are very unreliable when making absolute judgments of physical properties such as weight, temperature and pitch, but highly reliable when comparing one physical property with another, such as determining which of two weights is the heavier (see also Laming, 1984). Thurstone went on to apply CJ to psychological phenomena that have no measureable physical correlates, such as attitudes and social values (Thurstone, 1954). Later, Pollitt and Murray (1996) used CJ to investigate how examiners assess spoken performances, and Bramley, Bell & Pollitt (1998) used it to investigate changes in mathematics and English standards over time. CJ lends itself to this because, unlike scoring, it enables the direct rank ordering of scripts from different but equivalent exam papers. More recently, CJ has been used in a variety of assessment studies including Design and Technology ePortfolios (Kimbell, 2012), scientific enquiry skills (Davies, Collier & Howe, 2012) and narrative writing (Heldsinger & Humphry, 2010). However, no previous work has investigated the potential of CJ for directly assessing mathematics or mathematical problem solving.

MATHEMATICS ASSESSMENT TASKS

In this section, we contrast traditional tasks, typical of current mathematics examinations, with tasks specifically designed to assess problem-solving processes. To do this, we consider the traditional and problem-based assessments used in the two studies reported below.

The traditional assessments used in study 1 were examinations used to assess the General Certificate of Secondary Education (GCSE), a qualification taken by almost all students at age 16 in England, Wales and Northern Ireland. Mathematics GCSE exam papers have been criticised for containing

mostly short items that test only recall and rote application of routine procedures (Noyes, Wake, Drake & Murphy, 2011). An analysis of the six papers used in study 1 here supports this criticism. Each paper contained between 18 and 26 questions with up to nine parts per question and a total of 100 points available. The majority of available points (84 %) were contained within question parts worth just 1, 2 or 3 points. Across the six exams, the average number of points per question part varied from 1.34 to 1.91, with the largest question part worth 6 points. The number of points per question part provides a rough sense of the “reasoning lengths” required of candidates. “Reasoning length” is the average time that students are expected to allocate to reading a question, interpreting it and answering it. This may be calculated by taking 72 s as the time allowed per point ($2 \text{ h to sit each exam} \div 100 \text{ points per exam}$). In a typical paper, reasoning length is somewhere between 70 and 220 s.

The problem-solving mathematics assessments used in study 2 were taken from teacher-assessment materials that form part of a professional development package sent to all secondary schools in England and Wales by the Bowland Charitable Trust (Swan & Pead, 2008). These materials contain a collection of tasks that provide teachers with a way to assess problem-solving skills, referred to as “Key Processes” in the UK (QCA, 2008). These processes are summarised in the Bowland professional development materials and this summary is reproduced in the left hand column of Table 1.

The Bowland tasks are each designed to occupy about 20 min. Each task has just one or two prompts, allowing much longer reasoning lengths than traditional tasks. The questions are less structured than the GCSE tasks, allowing multiple solution pathways.

The Bowland assessment task in Fig. 1a, for example, invites students to design a sports bag. Clearly a single task cannot assess all the processes, but some aspects of each process are usually evident. The relationship between this task and the Key Processes is shown in the right hand column of Table 1. There are further qualitative differences between this task and more traditional exam tasks. The task has a context and the goal is reasonably authentic, requiring the construction and transformation of a net with labels and explanatory text. Students must interpret the requirements and constraints, such as noting the additional material needed for seams, identify the required mathematics, including the need for π and the conversion of metric units, and optimise the positioning of the pieces in order to minimise the cloth required to make the bag. An example of a high attaining student’s response is shown in Fig. 1b.

For comparison, we looked for analogous items in the GCSE exam papers used in the studies. We found three items across the exam papers

TABLE 1

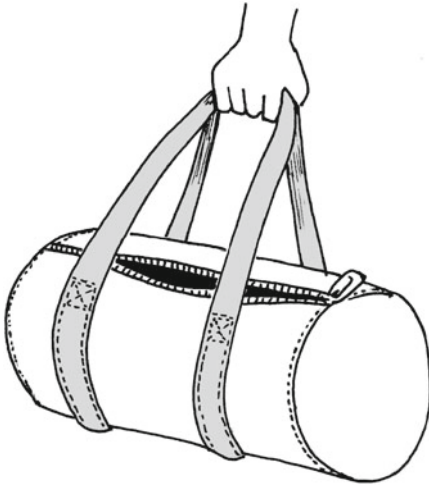
The “Key Processes” as described in Bowland Maths (Swan & Pead, 2008) and their relation to the Bowland problem solving task: *Design a Sports Bag*

<i>Key processes</i>	<i>The Bowland task: design a sports bag</i>
Simplify and represent: <ul style="list-style-type: none"> • Identify the maths. • Simplify and represent the problem. • Select information, methods and tools. 	Students simplify the problem by identifying the pieces from which the bag is constructed. Students interpret the constraints and make a sketch. They choose appropriate methods to calculate missing dimensions.
Analyse and solve: <ul style="list-style-type: none"> • Make connections with what is known. • Visualise, draw diagrams. • Systematically change variables. • Look for patterns and relationships. • Make calculations and keep records. • Make conjectures and generalisations. • Use logical, deductive reasoning. 	Students make connections between dimensions. For example, they recognise that the material for the body of the bag needs to be the same length as the circumference of the bag; they vary the positions of the pieces to determine the best way of using the cloth to minimise waste.
Interpret and evaluate: <ul style="list-style-type: none"> • Form conclusions, arguments and generalisations. • Consider appropriateness and accuracy. • Relate back to the original situation: is the solution good enough? 	Students consider assumptions and constraints in making the bag. For example, they take note of the additional material that will be needed for a seam.
Communicate and reflect: <ul style="list-style-type: none"> • Communicate and discuss findings effectively. • Consider alternative solutions. • Consider elegance, efficiency and equivalence. • Look for connections to other problems. 	Students describe their method and solution effectively and accurately, using words and sketches.

that were relatively long (worth 4 points or more) and, in common with the Bowland task described above, required the use of π . An example of one of these three items is shown in Fig. 2. It has a context but the goal is of questionable authenticity and requires only a single number as an answer. The item is semi-structured and students must identify the need to calculate the volume of a sphere, the volume of a cylinder and the need to subtract one from the other. The formulae for calculating these volumes are provided on a separate sheet within the paper. The examiner has

allocated 4 points to the question thereby anticipating it will require about five minutes to complete.

In sum, the traditional examination items assess mathematical problem solving less well than the Bowland tasks on two counts. Firstly, most traditional tasks are so short that there is no scope for problem solving. Secondly, where they are slightly longer, the problem-solving activities



You have been asked to design a sports bag.

- The length of the bag will be 60 cm.
 - The bag will have circular ends of diameter 25 cm.
 - The main body of the bag will be made from 3 pieces of material; a piece for the curved body, and the two circular end pieces.
 - Each piece will need to have an extra 2 cm all around it for a seam, so that the pieces may be stitched together.
1. Make a sketch of the pieces you will need to cut out for the body of the bag. Your sketch does not have to be to scale. On your sketch, show all the measurements you will need.
 2. You are going to make one of these bags from a roll of cloth 1 metre wide. What is the shortest length that you need to cut from the roll for the bag? Describe, using words and sketches, how you arrive at your answer.

Figure 1a. Reproduction of the “Sports Bag” task from the Bowland assessment materials

ASSESSING PROBLEM SOLVING

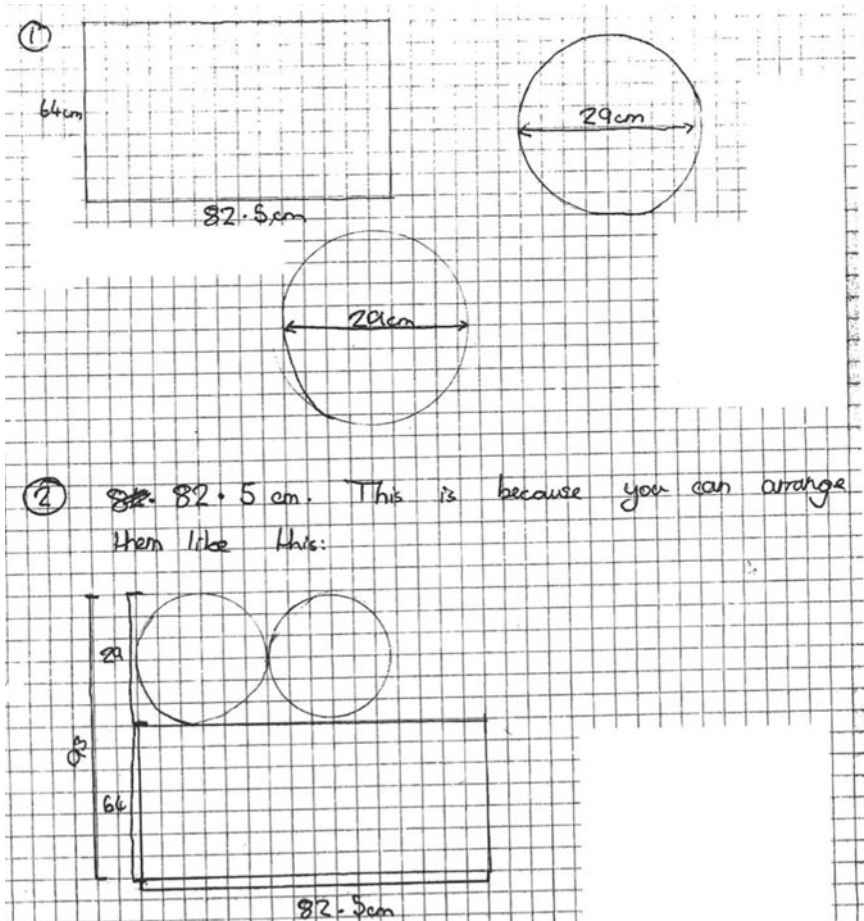
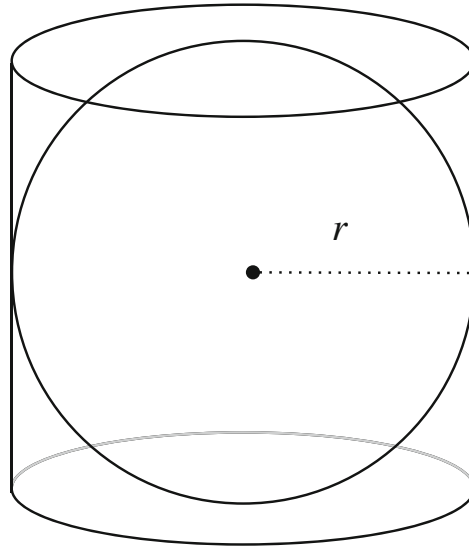


Figure 1b. Example of a high attaining student's response to the task. (The white spaces are where scoring comments have been removed.)

required of the students are limited and semi-structured and allow few opportunities for students to use the Key Processes shown in Table 1.

However, the use of many short items has some advantages over fewer problem-solving tasks, one of which is they allow precise scoring schemes and so achieve a high inter-rater reliability. Thus, we try to ensure that the score allocated to each student is not subjective. Student responses to longer, less structured tasks are more varied and less predictable and it is more difficult to achieve inter-rater reliability. This means scoring decisions are likely to be more subjective. On this basis, Laming (1990) has argued that questions should be split into as small components as possible to maximise the likelihood of each student being awarded the correct grade.

A tennis ball of radius r is packaged in a cylindrical box.
The ball touches the sides, top and base of the box.



What fraction of the volume of the box is empty space?
You **must** show all your working.

Figure 2. An example of an atypically long question from a GCSE exam paper (reproduction of a question that appeared in AQA (2010, p. 19))

In this article, we explore whether CJ, which does not require scoring schemes, might offer an approach to assessing mathematical problem solving while retaining a high inter-rater reliability.

RESEARCH FOCUS

We present two studies that were undertaken to establish the feasibility of CJ for assessing mathematics. In study 1, we tested CJ as an assessment method for the case of existing traditional exams (GCSEs). This was an extreme test case given the fragmentary nature of GCSE exam papers to establish whether CJ can replicate traditional scoring. In study 2, we tested CJ for the case of mathematical problem-solving tasks (Bowland). This was a feasibility test to explore the potential of CJ for assessing tasks that do not readily lend themselves to traditional scoring.

In both studies, two groups of mathematics education experts judged pairs of students' scripts and their decisions were used to construct scaled rank orders of scripts from "worst" to "best". The rank orders produced by the experts were compared to establish inter-rater reliability. The

outcome was also correlated with scripts' grades to obtain a measure of the validity of CJ as referenced against scoring.

STUDY 1: TRADITIONAL ASSESSMENTS

GCSE exams are frequently criticised for being highly structured and for encouraging recall and rote application of routine procedures, as discussed above. They therefore might be expected to be particularly unsuitable for rank ordering using CJ and so offer an extreme test case. Moreover, several equivalent forms of every GCSE exam are published and this enabled us to ensure an even more extreme test case for CJ by using scripts sourced from different forms of the exam. Study 1 enabled us to compare the rank order produced by CJ against that produced by scoring for GCSE scripts from several different but equivalent exam papers.

Assessment Materials

The assessment materials for study 1 were 18 scripts taken from six different but comparable GCSE exam papers (three awarding bodies \times two levels of difficulty).¹ The scripts came from the terminal exam papers for GCSE mathematics sat by the candidates on 7 and 11 June 2010. Six candidates' scripts were requested from each of three awarding bodies corresponding to one each of grades A*, A, B and C at higher tier (the more difficult forms of the exam) and grades C and D at foundation tier (the less difficult forms of the exam). Grades across different awarding bodies are standardised and we requested scripts that lay well within grade boundaries. Candidates' names, grades, scores and examiners' comments were removed for anonymity and to avoid influencing judges' decisions (Murphy, 1979).

Participants

The participants were 23 mathematics education professionals made up of ten GCSE examiners,² one non-GCSE examiner, seven mathematics education lecturers, two researchers, one research student and two advisors. The participants were allocated into two groups of 12 and 11 depending on which of two sessions they were able to attend.

Procedure

The implementation of CJ used for this study was TAG Development's *e-scape* system (Derrick, 2012), which presents pairs of scripts to experts online via an

internet browser. The *e-scape* system in fact supports Adaptive Comparative Judgement (ACJ) using an algorithm similar to that used in traditional adaptive testing to reduce the required number of judgements (Pollitt, 2012). However, reducing the number of judgements was not an issue for the small number of scripts used in the studies reported here and is not discussed in detail.

Copies of the exam papers were sent to participants 2 weeks prior to the study. They were requested to familiarise themselves with the questions but not the scoring schemes. The study was conducted in a single room with participants in each group working in parallel at laptop computers. The participants were first trained how to use the *e-scape* software and told to decide for each pair of scripts “which candidate is the more able mathematician”.

Discussion amongst participants was permitted during and encouraged after the judging sessions, which was audio recorded and subsequently transcribed. The independence of judgements was assured by instructing participants not to discuss individual decisions, and a member of the research team remained present at all times to monitor this. A short semi-structured feedback form was emailed to participants following the session.

Analysis

Group 1 completed 151 judgements during a session lasting about 105 min, and group 2 completed 150 judgements during a session lasting about 100 min. For each group, we fitted the judgement decisions to the logistic form of the Rasch model using FACETs (Bond & Fox, 2007), resulting in a parameter (standardised *z* score) and standard error being assigned to every script. The parameters were then used to construct a scaled rank order of scripts for “worst” to “best”. A detailed description of modelling comparative judgement data using the Rasch model can be found in Pollitt (2012).

There were three parts to the analysis. First we checked the internal consistency of the scaled rank order produced by CJ in three ways, as follows: (1) We calculated the Rasch sample separation reliability of each rank order, which is a measure considered by some to be analogous to Cronbach’s α (e.g. Wright & Masters, 1982). (2) We scrutinised the judges’ information mean square, or “misfit” figures, which provide a measure of the consistency of each judge’s performance as compared to all the other judges (Bond & Fox, 2007). To interpret the misfit figures, we applied the convention of considering any judge with a misfit figure greater than 2 standard deviations above the mean as performing spuriously. (3) We scrutinised the scripts’ misfit figures, which provide a measure of how consistently each script was assessed by all the judges. We again considered as spurious those scripts with a misfit figure greater than 2 standard deviations above the mean.

The second part of the analysis was to measure inter-rater reliability by calculating the Pearson product-moment correlation coefficient between scripts' parameters across the two groups. Third, we investigated validity by calculating the Spearman rank order correlation coefficient between scripts' parameters and grades (GCSE) or scores (Bowland). (We elaborate below as to how the Bowland scripts were scored.)

Results

The parameters, standard errors and misfit figures for each script produced by the two groups are shown in the [Appendix](#). The Rasch sample separation reliabilities were 0.80 and 0.93 for groups 1 and 2, respectively, suggesting an acceptably high internal consistency for both rank orders. One judge in group 1 had a misfit figure (2.02) marginally greater than 2 standard deviations above the mean ($\bar{x} + 2sx = 1.93$), and all judges' misfit figures in group 2 were lower than 2 standard deviations above the mean ($\bar{x} + 2sx = 1.98$). This suggests that the participants judged the scripts with acceptable mutual consistency. One script in group 1 had a misfit figure (1.38) marginally greater than 2 standard deviations above the mean ($\bar{x} + 2sx = 1.37$), and all scripts' misfit figures in group 2 were lower than 2 standard deviations above the mean ($\bar{x} + 2sx = 1.52$). This suggests that the scripts were each judged with acceptable consistency by the participants. Taken together, the Rasch sample separation reliabilities, judge misfit figures and script misfit figures suggest both scaled rank orders were of acceptable overall internal consistency.

The Pearson product-moment correlation coefficient of the scripts parameters across the two groups was 0.87, suggesting a high inter-rater reliability. The correlation of the scripts' parameters is shown in Fig. 3.

To generate a measure of validity, we first calculated a mean of the parameters produced by the two groups for each script. The Spearman rank order correlation coefficient between the mean parameters and grades was 0.91, suggesting the CJ method can validly rank students by mathematical achievement for the case of traditional exam scripts. The relationship between the scripts' parameters and grades is illustrated in Fig. 4.

Discussion

The two groups produced scaled rank orders that were internally consistent and yielded a high inter-rater reliability. The mean of the parameters produced by the two groups for each script correlated strongly with grades supporting the validity of the approach. Therefore, study 1 demonstrated that the judges successfully used CJ to assess the GCSE

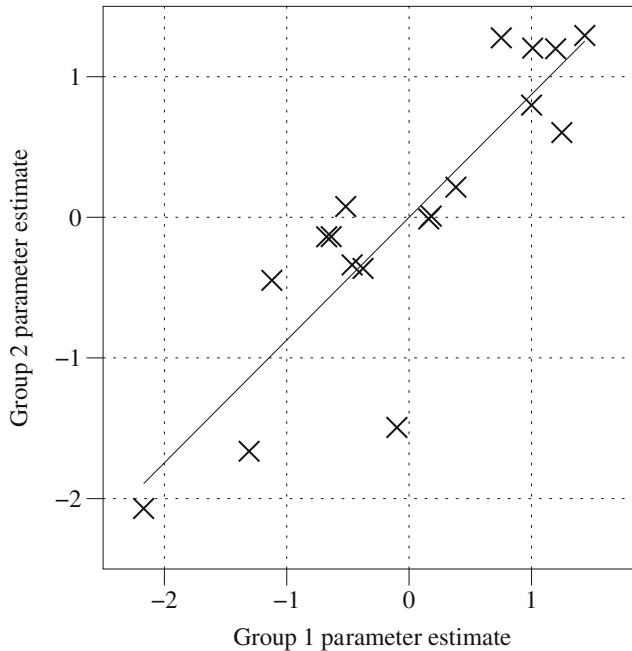


Figure 3. Correlation of the two groups' parameters for the GCSE scripts

scripts. The study provided an extreme test case of using CJ because of the unsuitability of the scripts for making pairwise judgements about mathematical ability. Moreover, the assessment materials comprised scripts from six different exam papers, thereby demonstrating the robustness of CJ for coping with equivalent forms of an assessment.

Transcripts of the judges' open discussion during and immediately following the workshops, and their completed feedback forms sent within the following week, revealed that many found making judgements difficult and at times stressful. Three main problems were cited. First, the scripts were far too long (up to 47 pages) for forming relative global judgements of pairs of candidates' mathematical abilities. Some judges felt that they were behaving unprofessionally by skimming work and not carefully examining every response to every question. Second, many of the items were too short and objective to contribute to an overall judgement of a candidate's mathematical ability. As one participant said in the open discussion immediately following the study: "There were a lot of questions in that paper that didn't tell you anything."

Third, many of the judgements involved comparing scripts from different exams. This is inevitably more challenging and time consuming than comparing like for like. These difficulties led some examiners to complain during the workshop that their decisions felt arbitrary and that

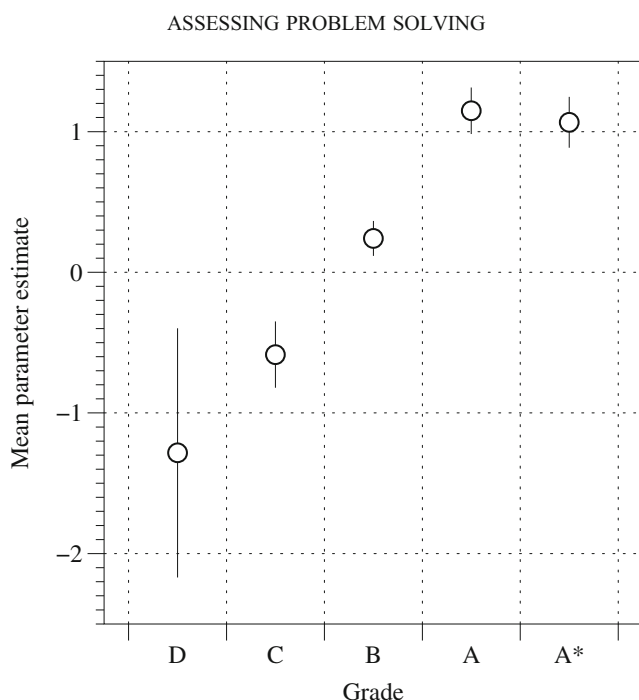


Figure 4. Correlation between the GCSE scripts' mean parameters and grades. Error bars show the standard deviation of the scripts' parameters at each grade

the process was a waste of time. However, the results reported above showed these same examiners were in fact making decisions consistent with one another and consistent with the scripts' grades.

STUDY 2: PROBLEM-SOLVING ASSESSMENTS

Study 1 demonstrated that CJ works reliably with scripts sourced from different forms of a traditional mathematics exam. However, our purpose is not to suggest CJ replace scoring for current mathematics exams, and in a second study we explored whether CJ might better enable some aspects of school mathematics to be assessed using problem-solving tasks.

Assessment Materials

The assessment materials for study 2 were 18 scripts based on a set of three tasks from the Bowland teacher-assessment materials described earlier in the article. Scripts were obtained from previous trials of three Bowland tasks. Students had been allowed up to 1 hour to complete the three tasks in a single sitting. The scripts had been scored using a traditional 10-point scoring rubric

with a total of 30 points available per script (10 points × 3 items). An example scoring rubric for the “Sports Bag” task is shown in Fig. 5. For the purposes of study 2, we selected 18 scripts that varied between 5 and 28 points (see Appendix for details). Unlike study 1, where the scripts were taken from different exam papers, the Bowland scripts were all taken from the same assessment tasks and were short, varying between three and four pages.

Participants

The judges were the same 23 participants from study 1. Some of the participants were familiar with the Bowland tasks used in study 2 but not their scoring schemes. One participant was substantially involved in the development, scoring and grading of the tasks as part of a previous, unpublished study.

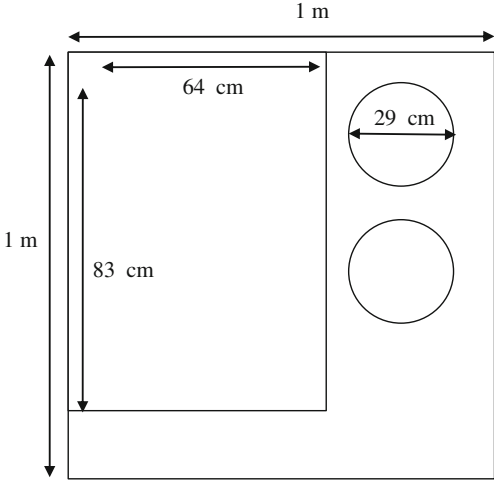
		Marks	Section marks
1.	<p>Circular ends $C = \pi d = \pi \times 25 = 78.5 \text{ cm}$</p> <p>Main body is a rectangle measurements $60 + 4 \text{ by } 78.5 + 4 = 64 \text{ by } 82.5 \text{ cm}$</p> <p>Two circular ends have diameter 29 cm</p>	2 2 1	 5
2.	<p>Draws sketch showing that 1 metre of cloth will make the bag.</p> 	5	5
	Total		10

Figure 5. Scoring rubric for the “Sports Bag” task

Procedure

The procedure was the same as that used for study 1.

Analysis

Group 1 completed 173 judgements in a session lasting about 55 min. Group 2 completed 177 judgements in a session lasting about 50 min. The analytical procedure was the same as for study 1.

Results

Details of the scaled rank orders produced by the two groups can be found in the [Appendix](#). The Rasch sample separation reliabilities were 0.85 and 0.93 for group 1 and 2, respectively, suggesting an acceptably high internal consistency for both rank orders. All the judges' misfit figures across both groups were lower than 2 standard deviations above the mean, suggesting they performed with acceptable mutual consistency. The scripts' misfit figures across both groups were lower than 2 standard deviations above the mean, suggesting each script was judged consistently by the participants. This suggested the two scaled rank orders were of acceptable internal consistency.

The Pearson product–moment correlation coefficient for the scripts' parameters across the two groups was 0.84, suggesting a high inter-rater reliability. The correlation of the scripts' parameters is illustrated in Fig. 6.

The Pearson product–moment correlation coefficient between the scripts' mean parameters and scores was 0.88, suggesting the CJ method can validly rank students in terms of mathematical achievement for the case of problem-solving tasks. The relationship between the scripts' parameters and standardized scores is illustrated in Fig. 7.

Discussion

Overall, the findings from study 2 demonstrated the suitability of CJ for assessing mathematical problem solving. The two groups produced internally consistent scaled rank orders that yielded a high inter-rater reliability. The mean parameter for each script correlated strongly with scores, supporting the validity of the assessment procedure. This demonstrated that the judges were capable of using CJ to rank order assessments containing longer, more problem-based tasks than are typically found in many mathematics exams.

Feedback revealed that most judges were more comfortable judging the Bowland materials than the traditional materials used in study 1. We note that the GCSE scripts contained up to 47 pages whereas the Bowland scripts were at most 4 pages, and this difference in volume no doubted accounted for much of the difference in judges' comfort. Nevertheless,

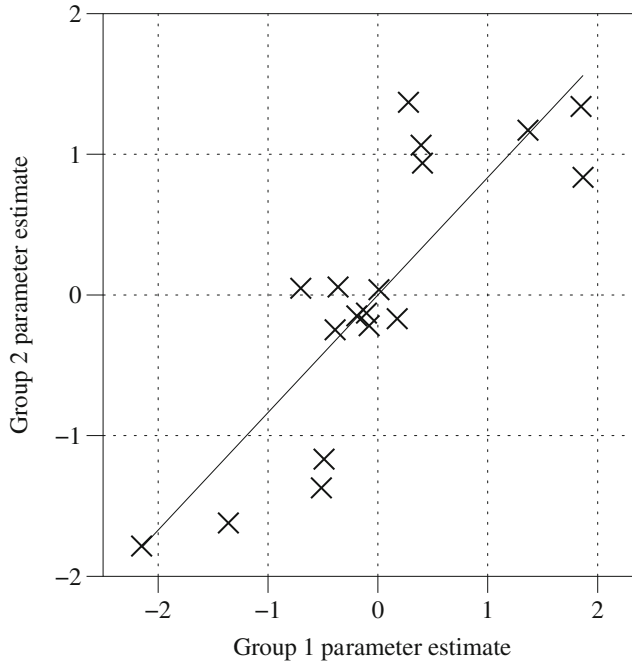


Figure 6. Correlation of the two groups' parameters for the Bowland scripts

some participant feedback suggests it was the nature of the Bowland tasks that eased the judging process too, for example:

Once I had got the hang of it the Bowland assessments were perfectly manageable although they required a lot of concentration. All the students had done the same three tasks, and as these were fairly open-ended it was not normally too difficult to decide which of each pair had done better.

GENERAL DISCUSSION

We conducted two studies to investigate whether CJ could be used to assess written mathematics exams. In both studies, two groups of judges used CJ to produce internally consistent scaled rank orders. The scripts' parameters across the two groups correlated strongly, demonstrating high inter-rater reliability. The mean of the scripts' parameters correlated strongly with scripts' grades or scores, supporting the validity of the method in both cases. These findings suggest CJ is feasible as a method for assessing students' mathematical proficiency as evidenced by traditional and problem-solving assessments.

ASSESSING PROBLEM SOLVING

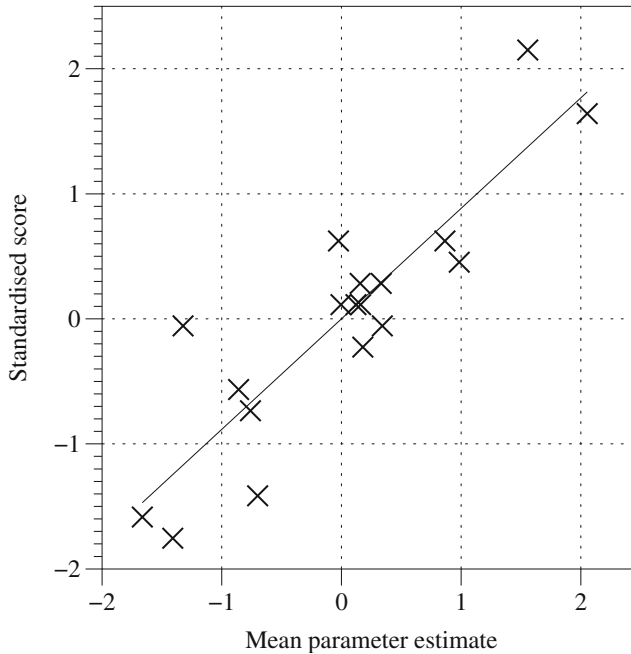


Figure 7. Correlation of mean parameters and standardized scores for the Bowland scripts

In the remainder of the article, we consider issues arising from the studies in terms of interpreting the results reported, areas for future research and development, and implications of using CJ for assessment at the regional and national level.

Interpreting Correlations

To obtain a measure of the inter-rater reliability, we repeated the procedure with a different group of examiners and correlated the two sets of parameters. However, we did not have access to score-rescore data for the GCSE or Bowland scripts and so could not obtain inter-rater reliabilities for scoring to compare with those for CJ. For GCSE scripts, we might expect scoring to achieve higher inter-rater reliabilities because of the exams being made up mostly from short, objective items (Murphy, 1982; Newton, 1996; Willmott & Nuttall, 1975). For the Bowland scripts, we might expect CJ to achieve inter-rater reliabilities equaling or even exceeding those achieved by scoring because of the relatively unstructured nature of the tasks. To our surprise, however, we found that the inter-rater reliability of the GCSE scripts (.87) was higher than that for the Bowland scripts (.84), although the difference was not significant ($p = 0.76$, two-tailed). This was despite the GCSE scripts being based on

different but compatible forms of an exam and the Bowland scripts being based on a single test.

We are unsure as to why the GCSE inter-rater reliability was marginally higher than the Bowland inter-rater reliability. One possibility is that the substantially longer GCSE papers provided more information to inform participants' judgements. The items were numerous and short but provided a broad sampling of a 2-year mathematics course and may have offered judges a more rounded view of each student's mathematical understanding. By contrast, the Bowland scripts comprised only three or four pages of students' responses to particular, if unstructured, problem-solving tasks and so did not offer a broad picture of mathematical understanding.

Another possible reason the GCSEs achieved marginally higher inter-rater reliability is that many participants were already familiar with scoring GCSEs, but few had experience scoring Bowland tasks. It may be that some experienced examiners mentally graded scripts, perhaps without consciously intending to do so, and compared grades rather than the scripts themselves when making decisions. To establish whether this was the case, we calculated parameter estimates from the judgements of participants with experience marking GCSE mathematics ($N=10$), and parameter estimates from the judgements of participants with no experience marking GCSE mathematics ($N=13$). The Pearson product-moment correlation coefficient between the two sets of parameter estimates was 0.84, which is not significantly different to the inter-rater reliability reported for study 1 ($p=0.76$, two-tailed), indicating high inter-rater reliability between the examiner and non-examiner groups. This suggests that even if some judges reverted to mental scoring techniques their performance was consistent with those judges for whom mental scoring was not an option. We return to the issue of judging processes later in the discussion.

To obtain a measure of the validity of CJ, we correlated the rank orders produced by the judges with those produced by grades (GCSE) or scores (Bowland). However, this measure of validity requires careful interpretation because CJ and scoring are not necessarily assessing the same construct. The strength of the correlations between CJ- and scoring-derived rank orders suggests that both measure "mathematics" in some sense. Given our contention that CJ might better assess problem solving and scoring can better assess factual recall and application, we should not expect the correlations to be too high, particular for the case of the Bowland tasks. However, caution must be taken because imperfect correlations can arise due to other reasons than methods measuring related but different constructs. For example, an imperfect correlation intended to measure validity may arise due to imperfect inter-rater reliabilities. It may

also arise due to poor construct validity of either or both the GCSE and Bowland tests (and the validity of GCSEs has come in for severe criticism, as discussed earlier).

These issues around reliability and validity need to be addressed in future work. In order to compare the inter-rater reliability of CJ with scoring, it will be helpful to obtain scripts for which score–rescore data is available or can be generated. In order to more confidently interpret validity, it will be helpful to obtain independent student data based on teacher assessments as well as standardised measures of mathematical understanding and problem solving.

Judging Processes

A key strength of CJ for assessing high-order thinking is its reliance on what is valued by the community of experts within a given discipline, rather than specification documents. As such an area of interest for further study is expert judging processes when undertaking pairwise comparisons. Some indicators for future research arose from participant feedback during and following the two studies.

In study 1, the difficulty of judging pairs of scripts comprising many short items over many pages led to participants using time saving strategies. For example, some used a sampling approach in which they looked at just a handful of items. These were sometimes the final two or three questions, which tended to be worth more points than most of the preceding questions. (This was not an item for item comparison because the scripts were sourced from different exam papers.) Interestingly, the use of time-saving strategies may have contributed to some judges feeling that they were operating unreliably. Some reported that they had not given students a fair hearing because they had not inspected every page of every script. The study was designed such that every script was viewed 18 times rather than once, as is typical with scoring, yet some examiners still felt uncomfortable. For example:

I did not feel ‘secure’ in my own judgement when I had skipped whole pages that a [student] had written. As an examiner I take my task seriously.

Some judges reported prioritising particular mathematical content areas over others in order to complete their judgements efficiently. For example, one examiner and former schoolteacher said during the open discussion immediately following study 1:

I posed my own judgement as a mathematics teacher what I would have wanted them to be able to do, so I looked at key areas. I am not bothered if they can do rotations, I am not bothered if they can do probability, I am not bothered if they can do the easy things. I want to see if they can do algebra and geometry.

The issue of developing a single measure for a construct such as mathematics that is made up of several distinct domains applies to any assessment method. CJ can bring the issue to the surface because examiners may feel their judgements are more prone to bias than when using a precise and detailed scoring scheme. Bias is known to play a role in scoring (Husbands, 1976; Suto & Greatorex, 2008), and research is needed to explore the role of bias in CJ.

Redesigning Assessments

Our motivation for exploring CJ is not to replace scoring for existing assessments, but to free assessment designers from restricting tasks to those that may be easily scored by conventional means. This would then allow them to introduce components into the assessment process that more adequately reflect its intended purpose. After the workshop, we asked those participants actively involved in writing GCSE exams: “If [CJ] was used nationally to grade papers would it affect how you write questions? Please elaborate.” Six of the participants across the two workshops were GCSE exam writers at the time of the study. Four answered positively, with varying degrees of enthusiasm. For example:

YES!!! It would be possible to have much, much better and richer questions and tasks ... I have at least one draft question that was rejected [by an awarding body] as being too open and unmarkable.

Others were more cautiously positive, seeing the potential of CJ as allowing more variation in forms of assessment than is used at present. For example:

I would write a couple of pages of short questions to judge basic mathematical knowledge and then a couple of longer, open-ended questions to allow for reasoning and process skills.

Two of the examiners answered negatively, highlighting the benefits of current exam papers. One of them expressed concern that a move to CJ might result in the loss of perceived strengths of existing papers:

Not necessarily. Papers are written taking into account many aspects, for example, accessibility, variety, novelty, standard algorithms open and closed questions, structured or multistep and I would hope this would continue.

Whether examiners would and could, in practice, produce substantially more open-ended assessment tasks if the need for scoring schemes were removed remains an open question for future research.

Further Issues of CJ for Assessing Mathematical Problem Solving

To conclude, we consider three further issues around using CJ rather than scoring for assessment purposes: sampling error, lack of feedback and scaling up.

Sampling Error. One drawback of open-ended problem-solving assessments, which may account for the higher inter-rater reliability in study 1 compared to study 2, is that such tasks cannot sample the content domain as broadly as an assessment comprising numerous short items. Narrow sampling threatens test–retest reliability because the material assessed in any given form of an assessment is likely to suit some students better than others. However, sampling error would only increase if the assessment was based solely on sustained problem-solving tasks. This need not be the case and it has long been acknowledged that a diverse raft of assessment formats is preferable to homogeneity (Black & Wiliam, 2007). For the case of assessing mathematics, it has recently been commented that:

Many really important aspects and ideas in mathematics cannot be assessed by exam only, for example sampling and data collection in statistics, mathematical modelling, numerical analysis, extended problem solving and appropriate use of computer software. Although these are written into the learning outcomes of many qualifications ... the present regulations ensure that they are assessed ... superficially. (Vordermann et al., 2011, p.97)

A well-known UK example of diverse assessment, including practical work, oral exams, written exams and project work, was Nuffield A-level physics (Black, 2008). We envisage that CJ would enable the reliable assessment of an unstructured component within a diverse range of assessment formats carefully chosen to appropriately assess the full range of valued competencies from technical fluency and conceptual understanding, through to problem-solving processes.

Lack of Feedback. Examiners often annotate scripts when scoring them and in some assessment systems their comments are fed back to candidates. This is not the case for CJ where the output is a scaled rank order of scripts and accompanying statistics about reliability, judges' performance and so on. The *e-scape* system does in fact allow judges to enter comments and explain their reasoning when making a judgement, but this has been found to slow the process down considerably and we requested the judges not to do so in the workshops. For systems in which written feedback is expected, this is likely to be a barrier to adopting CJ.

Scaling Up. If CJ were to be used as part of a regional or national mathematics assessment system, then it must be scalable. The first barrier

to scaling up is the time and costs required compared with traditional scoring. If this cannot be demonstrated to be comparable then CJ is unlikely to adopt for any large-scale assessment procedure. We have no data to inform scalability from the present studies although estimations based on other uses of the technology suggest the costs required would indeed be comparable (e.g. Kimbell, 2012). Scalability is now feasible due to a modified form of CJ known as ACJ in which an algorithm, similar to those used in adaptive computer tests, is used to select which pairs of scripts to present to judges. Where many scripts are involved (around $N > 100$), ACJ has been shown to require substantially fewer judgements than traditional CJ to construct stable rank orders (Pollitt, 2012). As mentioned earlier, the studies reported here were based on ACJ rather than traditional CJ. This was for reasons of convenience, and in practice for only 18 scripts there is little difference between ACJ and CJ in terms of the number of judgements required. Nevertheless, scalability is a pressing research area for developing CJ as a feasible assessment method at the regional or national level.

CONCLUSION

High-stakes assessment is a powerful influence on the inhibition or promotion of educational innovation (Jürges, Schneider, Senkbeil & Carstensen, 2012; Looney, 2009). It is therefore vital that mathematics exams assess what is valued and expected of a modern mathematics education. At present, sustained problem solving is valued and expected, but this stands in tension with current exams, which depend on short, precise items to achieve acceptable scoring reliability. We have reported here results from an alternative approach that derives its validity directly from what is valued and expected by mathematics experts, rather than what can be precisely captured in scoring rubrics. We believe our findings offer one way forward to improving the quality of summative mathematics exams.

ACKNOWLEDGMENTS

This work was supported by grants from the Nuffield Foundation and Royal Society. The authors are grateful to AQA, OCR, WJEC and the Bowland Charitable Trust for kindly supplying the materials used in the studies.

ASSESSING PROBLEM SOLVING

GCSE judges

ID	Grade	GCSE scripts				GCSE judges			
		Group 1		Group 2		Group 1		Group 2	
		Parameter	SE	Misfit	Parameter	SE	Misfit	ID	Misfit
1	A*	1.0097	0.5006	1.07	1.2020	0.1772	1.36	1	0.91
2	A*	0.7531	0.4329	0.75	1.2761	0.1587	1.09	2	0.4
3	A*	1.4348	0.5417	0.95	1.2921	0.2210	0.67	3	1.75
4	A	1.0008	0.3873	1.05	0.7974	0.1821	0.86	4	1.07
5	A	1.2474	0.3527	0.85	0.6019	0.1804	1.23	5	0.55
6	A	1.1963	0.3998	1.15	1.1984	0.2253	0.67	6	1.20
7	B	0.1598	0.4504	0.82	-0.0123	0.1519	1.25	7	0.96
8	B	0.3826	0.3844	1.38	0.2140	0.1600	1.14	8	1.28
9	B	1.802	0.3764	0.99	0.0088	0.1897	1.21	9	0.86
10	C	-0.6729	0.3968	0.83	-0.1391	0.1524	0.89	10	1.43
11	C	-0.0994	0.3633	1.07	-1.4945	0.2955	0.96	11	0.99
12	C	-0.5177	0.3965	1.24	0.0760	0.1840	0.90	12	1.53
13	C	-1.1206	0.5569	0.80	-0.4481	0.2657	1.07	Mean	2.02
14	C	-0.4663	0.3344	1.09	-0.3383	0.2078	0.83	S.D.	1.01
15	C	-0.6359	0.3823	1.13	-0.1371	0.1669	0.60		0.46
16	D	-0.3760	0.4361	1.15	-0.3639	0.2018	1.17		
17	D	-2.1686	0.7285	1.13	-2.0700	0.6563	0.18		
18	D	-1.3074	0.4504	0.72	-1.6634	0.2719	0.85		
Mean				1.01			0.94		
S.D.				0.18			0.29		

NOTES

¹ Awarding bodies are competing private institutions that publish equivalent exam papers according to government regulations.

² When using the term “examiner”, we include anyone involved in the writing, reviewing or scoring of mathematics exams.

REFERENCES

- ACT (2006). *Ready for college and ready for work: Same or different?* Iowa: American College Tests, INC.
- AQA (2010). *GCSE higher tier mathematics paper 1 (Specification A). Monday 7 June 2010*. Manchester: Assessment and Qualifications Alliance.
- Black, P. (2008). Strategic decisions: Ambitions, feasibility and context. *Educational Designer*, 1(1).
- Black, P. & Wiliam, D. (2007). Large-scale assessment systems: Design principles drawn from international comparisons. *Measurement: Interdisciplinary Research and Perspectives*, 5, 1–53.
- Black, P., Burkhardt, H., Daro, P., Jones, I., Lappan, G., Pead, D. & Stephens, M. (2012). High-stakes examinations to support policy. *Educational Designer*, 2(5).
- Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Abingdon: Routledge.
- Bramley, T., Bell, J. & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives*, 25, 1–24.
- CBI (2006). *Working with the three Rs: Employers' priorities for functional skills in Mathematics and English*. London: DfES.
- 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.
- Derrick, K. (2012). Developing the *e-scape* software system. *International Journal of Technology and Design Education*, 22, 171–185.
- DfE. (2011). *Independent evaluation of the pilot of the linked pair of GCSEs in Mathematics—First interim report (No. DFE-RR181)*. London: Department for Education.
- Heldsinger, S. & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, 37, 1–19.
- Husbands, C. T. (1976). Ideological bias in the marking of examinations: A method of testing for its presence and its implications. *Research in Education*, 15, 17–38.
- Jürges, H., Schneider, K., Senkbeil, M. & Carstensen, C. H. (2012). Assessment drives learning: The effect of central exit exams on curricular knowledge and mathematical literacy. *Economics of Education Review*, 31, 56–65.
- Kimbell, R. (2012). Evolving project *e-scape* for national assessment. *International Journal of Technology and Design Education*, 22, 135–155.
- Laming, D. (1984). The relativity of “absolute” judgements. *British Journal of Mathematical and Statistical Psychology*, 37, 152–183.

- Laming, D. (1990). The reliability of a certain university exam compared with the precision of absolute judgements. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 42, 239–254.
- Looney, J. (2009). *Assessment and innovation in education*. Paris: OECD.
- Murphy, R. (1979). Removing the marks from exam scripts before re-marking them: Does it make any difference? *British Journal of Educational Psychology*, 49, 73–78.
- Murphy, R. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, 52, 58–63.
- NCETM (2009). *Mathematics matters: Final report*. London: National Centre for Excellence in the Teaching of Mathematics.
- Newton, P. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, 22, 405–420.
- NGA and CCSSO (2010). *Common core state standards for Mathematics*. Washington, DC: National Governors Association and Council of Chief State School Officers.
- Noyes, A., Wake, G., Drake, P. & Murphy, R. (2011). *Evaluating Mathematics pathways final report (Technical Report No. DFE-RR143)*. London: Department for Education.
- OECD (2009a). *PISA 2009 results: Learning trends: Changes in student performance since 2000 (Volume V)*. Paris: OECD.
- OECD (2009b). *PISA 2009 assessment framework: Key competencies in reading, Mathematics and Science*. Paris: OECD.
- Ofsted (2008). *Mathematics: Understanding the score*. London: Office for Standards in Education.
- Ofsted (2012). *Mathematics: Made to measure*. London: The Office for Standards in Education.
- Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19, 281–300.
- Pollitt, A. & Murray, N. (1996). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th language testing research colloquium*. Cambridge: Cambridge University Press.
- QCA (2008). *National Curriculum for England 2008*. London: Qualifications and Curriculum Authority.
- Rocard, M. (2007). *Science education now: A renewed pedagogy for the future of Europe. Brussels: European Commission* (Technical Report No. EUR22845). Retrieved from http://ec.europa.eu/research/science-society/document_library/pdf_06/report-rocard-on-science-education_en.pdf
- Soh, C. K. (2008). An overview of mathematics education in Singapore. In Z. Usiskin & E. Willmore (Eds.), *Mathematics curriculum in Pacific Rim countries* (pp. 23–36). Mississippi: Information Age.
- Suto, W. M. I. & Greateorex, J. (2008). What goes through an examiner’s mind? Using verbal protocols to gain insights into the GCSE—marking process. *British Educational Research Journal*, 34, 213–233.
- Swan, M. & Burkhardt, H. (2012). Designing assessment of performance in mathematics. *Educational Designer*, 2(5).
- Swan, M. & Pead, D. (2008). Bowland Maths Professional development resources. *Bowland Trust/Department for Children, Schools and Families*. Retrieved from www.bowlandmaths.org.uk
- 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.

ASSESSING PROBLEM SOLVING

- Toner, P. (2011). *Workforce Skills and Innovation (OECD Education Working Papers)*. Paris: Organisation for Economic Co-operation and Development.
- Treilibs, V. (1979). *Formulation processes in mathematical modelling*. Nottingham: Unpublished MPhil, University of Nottingham.
- Vordermann, C., Porkess, R., Budd, C., Dunne, R. & Rahman-Hart, P. (2011). *A world-class mathematics education for all our young people*. London: The Conservative Party.
- 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.
- Willmott, A. S. & Nuttall, D. L. (1975). *The reliability of examination at 16+*. London: Macmillan Education.
- Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA.

Ian Jones
Mathematics Education Centre, Schofield Building
Loughborough University
Loughborough, LE11 3TU, UK
E-mail: I.Jones@lboro.ac.uk

Malcolm Swan
School of Education
University of Nottingham
Nottingham, UK

Alastair Pollitt
CamExam
Cambridge, UK