

# The problem of assessing problem solving: can comparative judgement help?

Ian Jones<sup>1</sup> · Matthew Inglis<sup>1</sup>

Published online: 29 May 2015

© Springer Science+Business Media Dordrecht 2015

**Abstract** School mathematics examination papers are typically dominated by short, structured items that fail to assess sustained reasoning or problem solving. A contributory factor to this situation is the need for student work to be marked reliably by a large number of markers of varied experience and competence. We report a study that tested an alternative approach to assessment, called comparative judgement, which may represent a superior method for assessing open-ended questions that encourage a range of unpredictable responses. An innovative problem solving examination paper was specially designed by examiners, evaluated by mathematics teachers, and administered to 750 secondary school students of varied mathematical achievement. The students' work was then assessed by mathematics education experts using comparative judgement as well as a specially designed, resource-intensive marking procedure. We report two main findings from the research. First, the examination paper writers, when freed from the traditional constraint of producing a mark scheme, designed questions that were less structured and more problem-based than is typical in current school mathematics examination papers. Second, the comparative judgement approach to assessing the student work proved successful by our measures of inter-rater reliability and validity. These findings open new avenues for how school mathematics, and indeed other areas of the curriculum, might be assessed in the future.

**Keywords** Assessment · Problem solving · Validity · Comparative judgement

Typical mathematics examination papers are not fit for the purpose of assessing students' mathematical knowledge and skills. Analyses of the content and style of examination papers support the conjecture that mathematics examination papers comprise mainly short items that assess the rote learning of isolated facts and procedures (Berube, 2004; NCETM, 2009; Noyes,

---

**Electronic supplementary material** The online version of this article (doi:10.1007/s10649-015-9607-1) contains supplementary material, which is available to authorized users.

✉ Ian Jones  
i.jones@lboro.ac.uk

<sup>1</sup> Mathematics Education Centre, Loughborough University, Loughborough LE11 3TU, UK

Wake, Drake, & Murphy, 2011). An example question from a recent General Certificate for Secondary Education (GCSE) examination paper, a national qualification in England taken by most school leavers, illustrates the problem, as shown in Fig. 1.

At first glance, the question looks promising for assessing students' mathematical knowledge and skills. It makes use of a calendar context thereby appealing to the everyday relevance of mathematics. It also builds on students' experience of a counting system grouped in 7s to introduce an interesting generality that wherever the 2 by 2 square is positioned the provided algorithm will always give 7. However, to achieve full marks, all a student needs to do is compute the provided algorithm using the provided inputs. No explanation or proof of why the result is always 7 is required or rewarded. An efficient examination taker can achieve full marks without noticing there is a mathematically interesting generality at all.

The question might be improved by asking students to compute the algorithm for a few 2 by 2 squares of their own choosing, and then asking them to explain what they notice. Such an adapted version of the question might better test those attributes reported to be valued by

Here is a calendar for May 2010.

Su	Mo	Tu	We	Th	Fr	Sa
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

This 2 by 2 square is take from the calendar.

3	4
10	11

Multiply the diagonal numbers together.       $4 \times 10 = 40$   
     $3 \times 11 = 33$

Then find the difference.                               $40 - 33 = 7$

Difference = 7

Do the same for this 2 by 2 square taken from the calendar.

5	6
12	13

Show your working. (3 points)

**Fig. 1** Question from a recent mathematics GCSE examination paper (reproduction of a question that appeared in AQA (2010, p. 6))

stakeholders of high schooling systems, such as problem solving and sustained reasoning (NCTM Research Committee, 2013; Suto, 2013; Vorderman, Porkess, Budd, Dunne, & Rahman-Hart, 2011).

This fragmented presentation of mathematics is at odds with the stated aims of mathematics curricula (Noyes et al., 2011; Ofsted, 2008), and fails to test what is valued most by educators and employers (e.g. ACME, 2011; McLester & McIntire, 2006; Walport et al., 2010). So why do examination papers contain so many short, closed questions? There are several constraints that shape examination papers (Burkhardt, 2009), but our focus here is on a specific and, we argue, addressable constraint: the need for examination papers to be marked reliably and how this might impact on validity.

To assist international readers, we now explain the examination situation and terminology in England. At the end of compulsory schooling most students aged 16 sit examination papers for the General Certificate for Secondary Education (GCSE) in mathematics. The GCSE is intended as a 2-year course and, following a recent policy change, candidates sit two examination papers as the conclusion of the course. The examination papers are designed by question writers working for one of three competing examination boards. For each examination paper a *mark scheme*, or scoring rubric, is also produced detailing how responses to each question should be *marked* (or scored). The examination boards outsource the marking, typically to teachers who undertake the work during school holidays. In this paper, we refer to those who mark student work as *examiners*. Once the marking is complete a committee decides on grade boundaries using statistical methods and human judgement. The final outcome is a letter grade for each student.

In the remainder of the paper, we report a study designed to explore how high-school mathematics might be assessed in the absence of traditional marking procedures. First, we discuss the need for valid and reliable assessments, and how the need for high reliability can constrain the types of questions used in examination papers. We then describe an alternative approach to assessing mathematics, called comparative judgement, which requires no marking and no mark schemes. For the empirical study, a special examination paper was designed by experienced question writers, administered to 750 students, and assessed using comparative judgement and a specially designed, resource-intensive marking procedure. The study design was intended to address the research question, “What is the potential of comparative judgement for improving the validity while retaining the reliability of high-stakes examinations?”

## 1 Validity

A common, if sometimes contested, definition of validity is that a valid assessment measures what it purports to measure (Koretz, 2008), often referred to as *construct validity* (Messick, 1980). In school examination papers, the construct of interest tends to be broad, such as mathematical knowledge and skills. Two common approaches exist for investigating whether an assessment has construct validity (Newton & Shaw, 2014). First, *content validity* can be evaluated directly through the analysis of examination questions by relevant experts such as teachers. Second, empirical evidence can be obtained by correlating assessment outcomes with independent outcomes that are believed to measure the same, or a similar, construct. The resulting correlation coefficient can be considered a measure of the *criterion validity* of an assessment. In the research reported here, we investigated both content and criterion validity in order to evaluate the performance of the assessment.

In recent decades, some theorists have argued that the purpose and impact of an assessment should be considered as central to evaluating its validity (Cronbach, 1988; Messick, 1989; Shepard, 1997). The underlying motivation for the present study rests with concern about the *consequential validity* of many high-stakes assessments in mathematics; specifically, that the prevalence of short, closed examination questions such as that exemplified above results in the piecemeal learning of fragmented mathematics in classrooms (ACME, 2005; Black et al., 2012; Duncan, 2010; NCETM, 2009). The challenge, then, is to better align the content of mathematics examination papers to what is sought by stakeholders of education systems. In the present time, there is a broad consensus that learning mathematics should involve sustained problem-solving activities (NCTM Research Committee, 2013; Swan, 2014; Vorderman et al. 2011). Evaluating the consequential validity of an assessment is a long-term and difficult process, which some have argued is not possible (e.g., Borsboom, Mellenbergh, & van Heerden, 2004), and is beyond the scope of this paper. However, marking and comparative judgement offer distinctive ways of thinking about the validity of an assessment, and exploring approaches that might improve the educational consequences of high-stakes assessment was a key motivation for the research.

## 2 Reliability

An assessment cannot be said to be valid unless it is also *reliable* (William, 2001). Reliability relates to the consistency of outcomes of an assessment procedure and our focus here is on *inter-rater reliability*, which refers to the level of agreement between different examiners when assessing students' work. The lower the inter-rater reliability, the more dependent a given candidate's outcome is on the idiosyncrasies of whoever happened to mark the work, and so the less fair the assessment. Inter-rater reliability is usually investigated by recruiting different examiners to mark the same students' work and comparing the outcomes, typically using the Pearson product-moment correlation coefficient. Willmott and Nuttall (1975) undertook a study of the inter-rater reliability of terminal UK school examinations sat in 1969 and 1970 across numerous subjects and awarding bodies. They reported inter-rater reliabilities across different subjects ranging from 0.54 to 0.95 with most achieving >0.80. Similarly, Murphy (1982) conducted a study in which Chief Examiners remarked student work that had originally been marked by teams of examiners working under their remit, and reported inter-rater reliabilities ranging from 0.80 to 1.00.

Such studies typically compare inter-rater reliabilities across different subject disciplines. Mathematics examinations often prove the most reliable, closely followed by science, and the least reliable examinations are in languages (James, 1974; McVey, 1976; Newton, 1996). It would seem, then, that disciplines commonly associated with precision and accuracy more naturally lend themselves to assessments with high marking reliability. However, Murphy (1982) undertook a detailed scrutiny of his data and discovered that differences were more dependent on the design of the examination than subject domain. This was evident in the inter-rater reliabilities of the three independent examination papers that made up some examinations in biology (0.98, 0.98 and 0.61, respectively), French (0.98, 0.99 and 0.81, respectively) and English (0.73, 0.85 and 0.76, respectively). Unsurprisingly, the lower reliability examination papers were essay-based and the higher reliability examination papers were "made up of highly structured, analytically marked, questions" (p. 62). Similar variations across assessment formats within

subject disciplines have been reported elsewhere (e.g., van Aalst & Chan, 2007; Willmott & Nuttall, 1975).

In light of the literature on marking reliability, we conjectured that school examination papers do not assess sustained mathematical problem solving due in part to the drive to achieve high inter-rater reliability through detailed and objective mark schemes. As Swan and Burkhardt (2012) put it: “Mathematics examiners have long been proud of their ‘reliability’—the consistency of marks when independent examiners using the same mark scheme assess the same collection of responses” (p. 32). The need for reliable marking leads to examination paper writers favouring short, structured items to ensure a limited pool of predictable responses from candidates.

### 3 Comparative judgement (CJ)

In this paper, we explore the potential of an approach to assessment, called comparative judgement (CJ), that offers an alternative to traditional marking. The underlying theoretical basis is a well-established psychological principle that people are more reliable when comparing one sense impression against another than they are at judging an impression in isolation (Laming, 1984; Thurstone, 1927). For example, it is easier to decide which of two weights is the heavier than to estimate (to the nearest gram, say) a single weight in isolation.

The basic mechanics of CJ are simple. Experts are presented with pairs of students’ work and asked to decide which is “better” in terms of some global construct such as “mathematical ability”. The experts’ decisions are fitted to a statistical model to produce a standardised parameter estimate (z-score) for each student (Pollitt, 2012a). The parameter estimates are then used to construct a scaled rank order of student work from “worst” to “best” and the usual assessment arrangements, such as allocating grades, can be applied to the rank order (see Jones & Alcock, 2014; McMahon & Jones, 2014).

CJ has been used in a range of educational research and practice (e.g., Bramley, 2007; Bramley, Bell, & Pollitt, 1998; Heldsinger & Humphry, 2010; Seery, Canty, & Phelan, 2012). Thurstone’s (1927) underlying principle of comparative judgement suggests that we should obtain reliable assessment outcomes even though the process is based on “subjective” judgements. In previous studies, we have found this to be the case when used to assess traditional GCSE mathematics examination papers (Jones, Swan, & Pollitt, 2014), conceptual tests of children’s understanding of fractions (Jones, Inglis, Gilmore, & Hodgen, 2013) and undergraduates’ understanding of calculus (Jones & Alcock, 2014).

One potential contribution that CJ might offer education is its suitability for assessing nebulous constructs that are deemed important but which are difficult to specify comprehensively in mark schemes (Pollitt, 2012b). Key to this potential are contrasting assumptions about how construct validity is achieved when using CJ compared to marking. Mark schemes attempt to capture the construct of interest using explicit, precise and detailed assessment criteria. CJ instead relies on the collective understanding of the construct by a relevant community of experts. It might be countered that this is an opaque view of assessment validity, and that a given construct might vary from one expert to another. Moreover, CJ is suited to handling a construct that is specified only at the most global level, but we argue that this is a key strength of the approach. There is no pretence of a universally defined construct that all examiners must interpret in the same way, and examination paper writers are not imposing through mark schemes their own view of the construct on

markers. Rather, CJ assimilates the varied ways in which a given a community of experts understands a construct in practice.

## 4 The study

In this article, we report a study that evaluated the potential of CJ for the assessment of high school mathematics. The motivation was our contention that marking is a major reason why current mathematics examination papers, at least in England, require mainly short, precise responses from candidates, which may make valid assessment of sustained mathematical problem solving difficult.

There were two main parts to the research study, a design phase and an assessment phase, as summarised in Table 1. In the design phase, we explored whether examination paper writers, when freed from the constraints of mark schemes, design an examination paper that is more open and less structured than is currently common. In the assessment phase, we investigated the outcomes of using CJ to assess student responses to the specially designed examination paper.

## 5 Design phase

In the design phase of the study, four experienced examination paper writers were commissioned to produce a mathematics examination paper that would require no mark schemes and no marking. Our motivation for this approach was to explore the extent to which marking impacts on content validity. In particular, we were interested to know whether an examination paper designed free of marking considerations would contain tasks that are qualitatively distinct to those in typical contemporary GCSE examination papers. In order to evaluate

**Table 1** Summary of steps undertaken during the design and assessment phases of the research

Design phase	
Design workshop 1	4 GCSE question writers drafted 7 initial examination questions.
Question trialing 1	Draft questions trialed with 57 high-school students from 3 schools.
Question refining 1	Question writers redrafted 6 questions in light of student responses.
Design workshop 2	Question writers compiled draft examination paper.
Question trialing 2	Draft examination paper trialed with 43 high-school students from 1 school.
Question refining 2	Question writers revised examination paper in light of student responses.
Mark scheme writing	1 examination writer designed post-hoc mark scheme using sample of student responses to final examination paper.
Teacher survey	106 mathematics teachers responded to an online survey about the examination paper (94 responses used in the analysis).
Assessment phase	
Examination paper administered	750 students aged 14 and 15 from 2 schools sat the examination paper.
Marking	4 highly experienced mathematics teachers marked the student work.
CJ	20 experts (mathematics education researchers and research students; teachers) comparatively judged the student work.
Judge survey	13 judges responded to an online survey about the judging procedure.

content validity, the examination paper was scrutinised by mathematics teachers who then completed an online survey.

### 5.1 Designing the examination paper

Four GCSE question writers, who were available and willing to undertake the work, were commissioned for a total of three days each. The question writers had been involved in a previous study that investigated the feasibility of using CJ to assess existing GCSE examination papers (Jones et al., 2014). It was necessary for the question writers to be familiar with CJ as we were interested in seeing how knowledge of the assessment method would impact on the work of experienced question writers.

The question writers attended a design workshop (see Table 1) where they were first briefed on the requirements for the examination paper. The overall goal was to produce a GCSE-like examination paper that could in principle be used for large-scale summative assessment using CJ rather than marking. They were asked to keep CJ in mind at all times, and to put mark schemes and marking out of mind when drafting questions (for research purposes the student work was marked, see below, but the question writers were not told this until the examination paper had been written). They were also told that the examination paper needed to be accessible to students of all abilities, from those expected to achieve the highest grade through to those predicted to achieve the lowest grade. Finally, the question writers were informed that the examination paper should take students up to 50 min to complete so that it could be administered by teachers in a single lesson. A further practical constraint arose that the examination paper would be administered to students near the start of their 2-year course of study for a GCSE qualification in mathematics. Therefore, the question writers could not assume that a great deal of specific content would have been covered by the potential candidates. It was emphasised verbally and in writing that beyond these requirements the researchers would leave the logistics and details of the examination paper's development to the discretion of the question writers.

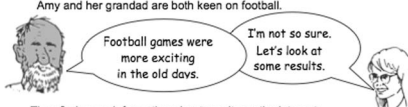
During the remainder of the workshop, the question writers worked together to write examination questions. Over the following 2 months, the questions were trialled and redrafted following the steps shown in Table 1 until a final examination paper was produced. The examination paper, entitled "Maths Problems", contained six "tasks" spread over a total of 11 pages, including a "resource sheet" that contained information required for completing some of the tasks. Tasks were identified by names (e.g. "Nines", "Money money!") rather than numbers, and the number of marks per question was not shown because this would have been meaningless in the absence of a mark scheme. An example question focussed on applied statistics can be seen in Fig. 2. The final examination paper is provided as Electronic Supplementary Material 1.

### 5.2 Designing the mark scheme

The examination paper designers had been briefed that the examination paper need not be marked to ensure that mark schemes and marking were put out of mind, as described above. It was therefore necessary to commission an examination writer to develop the mark scheme only after the final examination paper had been produced, which goes against the recommendation and practice that a mark scheme should be designed at the same time as an examination paper (Taggart, Phifer, Nixon, & Wood, 1998). The outcome was an unusually complicated



**Good old days?**  
Amy and her grandad are both keen on football.



Football games were more exciting in the old days.

I'm not so sure. Let's look at some results.

They find some information about results on the internet.


1911 – Saturday April 22 <sup>nd</sup>		
Aston Villa	4 – 2	Manchester United
Blackburn Rovers	3 – 0	Tottenham Hotspur
Everton	1 – 1	The Wednesday
Manchester City	1 – 2	Bristol City
Oldham Athletic	0 – 0	Bury
Sunderland	1 – 1	Notts. County
Woolwich Arsenal	2 – 0	Preston North End


2011 – Saturday April 23 <sup>rd</sup>		
Aston Villa	1 – 1	Stoke City
Blackpool	1 – 1	Newcastle United
Chelsea	3 – 0	West Ham United
Liverpool	5 – 0	Birmingham City
Manchester United	1 – 0	Everton
Sunderland	4 – 2	Wigan Athletic
Tottenham Hotspur	2 – 2	West Bromwich Albion
Wolverhampton Wanderers	1 – 1	Fulham

These results are for Saturday April 22<sup>nd</sup> 1911 and Saturday April 23<sup>rd</sup> 2011. They are both for the top division.  
For example, the first table shows that when Aston Villa played Manchester United in April 22<sup>nd</sup> 1911, Aston Villa scored 4 goals and Manchester United scored 2 goals.

Use the information in the tables to answer these questions.  
You must support your answers with numbers or calculations.

(a)  If two teams score the same number of goals in a game then it is a draw. Draws were more likely a hundred years ago than they are now.

Do you agree with Amy?

(b)  Games were more exciting a hundred years ago.

Do you agree with Amy's Grandad?

**Fig. 2** Example question from the final examination paper

mark scheme that ran to 16 pages. The cover page is shown in Fig. 3 and the full mark scheme is provided as Electronic Supplementary Material 2.

### 5.3 Content validity

To evaluate the examination paper in comparison to typical GCSE examination papers, we asked a sample of mathematics teachers to scrutinise the examination paper and then complete an online survey. The participants were self-selecting teachers who responded to calls made via online teacher forums (the Times Educational Supplement staffroom and the National Centre for Excellence in the Teaching of Mathematics web portal), teacher email lists held by three universities in England, and the educational charity Mathematics Education and Industry.

The survey was designed on the basis of our own scrutiny of the final examination paper, and included the following four closed questions:

1. How well do you think the paper assesses mathematical problem solving?
2. How well do you think the paper assesses mathematical content?
3. How well do you think the paper assesses the Key Stage 4 Process Skills in mathematics?
4. How well do you think your students would perform on this paper?

Question 3 might be loosely considered a rewording of Question 1 using language familiar to teachers in England (“Key Stage 4” is the stage of schooling that most students are taught GCSE mathematics, and “Process Skills” are defined as the individual stages of problem solving, specified in the National Curriculum for England as: representing; analysing; interpreting and evaluating; communicating and reflecting; QCA, 2007).

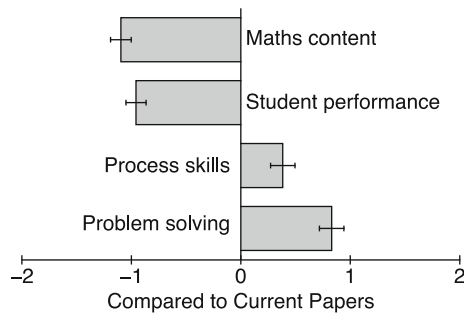


<p><b>Mark scheme – Notes</b></p> <p><b>Examples</b> Examples in the last column are shown in italics. An example on the right relates to the description of a type of answer on the left. They are <b>only examples</b>. Other possibilities are credit-worthy if they (more or less) fit the description. But if a response doesn't fit the mark scheme, <i>use your judgement</i>.</p> <p><b>Structures</b> The mark schemes for different questions have different structures.</p> <p><b>Type 1: Simple</b> <i>Factory (a) and (b); Cooking (a)</i> These are very straightforward. The answer is either right or wrong. There are not many questions like this!</p> <p><b>Type 2: Levels of response</b> <i>Nines; Factory (c), (d) and (e)</i> Different types of response to these questions are worth different numbers of marks. Try to match the student's response to <b>one</b> of the descriptions on the left, using the examples as a guide. But if a response doesn't fit the mark scheme, use your judgement.</p> <p><b>Type 3: Points</b> <i>Pool; Good old days (a) and (b); Money, Money! (a) and (b); Cooking (b)</i> There is a list of different 'points' that a student might make in the left hand column, with descriptions of responses that are worth different numbers of marks. Marks may be awarded for each point that the student makes. So in Pool, for example, a student might, possibly, discuss all four points – <i>Accuracy</i>, the <i>Social context</i>, the <i>Physical context</i> and <i>Measurement</i>, and get two or three marks for each giving a maximum total possible of 8 marks. In reality, though, most students make just one or two points, so the marking on <i>Pool</i> is much lower than this. Here again, if a response doesn't fit the mark scheme, <i>use your judgement</i>.</p> <p><b>Type 4: Steps</b> <i>Money, money (c)</i> Part (c) of Money, money has two 'points': the calculations made, and the degree to which the student actually related to the context of the problem. Within the first point there are three methods described, and two of these (using <i>Volumes</i> and using <i>Layers</i>) include a number of possible steps each of which is worth one mark. Do not agonise too long over responses to this question or you may lose the will to live. Here especially, if a response doesn't fit the mark scheme <i>use your judgement</i>.</p> <p><b>Mark record</b> For the more complex, multi-mark questions where students often pick up marks for making different 'points' (<i>Good old days (b)</i>, <i>Money, money (a), (b) and (c)</i>, and <i>Cooking (b)</i>) I found it helpful to keep a record of the number of marks awarded for each 'point' and then add them up for the whole question. I used a <i>Mark record</i> sheet which I have pasted in at the end of this mark scheme.</p> <p><b>Most important point</b> If a response doesn't fit the mark scheme... <i>use your judgement!</i></p>
---

**Fig. 3** Cover page of the retrospectively designed mark scheme

Each question was answered on a five-point Likert scale from -2 ("A lot worse than a typical current GCSE paper") to +2 ("A lot better than a typical current GCSE paper"). An optional open-text question was also provided with the prompt "Please add any comments you might have on the paper". There were a total of 106 online survey responses. Twelve respondents had incomplete data and were removed, leaving a total of 94 respondents included in the analysis. Sixty-eight of these left open-text comments.

A summary of the teachers' ratings is shown in Fig. 4. A mean rating of 0, as shown by the vertical line, indicates the level of a typical current GCSE examination paper. We investigated whether mean ratings were significantly different from 0 using non-parametric one-sample Wilcoxon signed-rank tests. This revealed each question was significantly different from the value representing a typical GCSE paper at the  $p < .002$  level (Bonferroni adjusted). The teachers' mean ratings for how well the examination paper would assess problem solving (0.83) and Key Stage 4 Process Skills (0.38) were higher than for a typical GCSE examination paper, and their mean ratings for how well the examination paper would assess mathematical



**Fig. 4** Teachers' responses ( $N=94$ ) to the survey evaluation of the examination paper.  $-2$  represents "a lot worse than a typical current GCSE paper" and  $+2$  represents "a lot better than a typical current GCSE paper". The vertical line at  $0$  indicates "about the same as a typical current GCSE paper"

content ( $-1.10$ ) and how well their students would perform ( $-0.96$ ) were lower than for a typical GCSE examination paper. These results were in line with expectations following our own scrutiny of the examination paper. In addition, teachers considered the examination paper to be significantly better at assessing problem solving than Key Stage 4 Process Skills,  $t(93)=4.37$ ,  $p<.001$ . This is interesting as it indicates problem solving is not synonymous with Process Skills, although exploration of this issue is beyond the scope of the present article.

To further investigate content validity, we turned to the 68 open-text responses. Many of the comments related directly to three of the questions (problem solving, mathematical content and student performance) and we consider these in turn. In addition, some teachers commented, without direct prompting, on the difficulty of marking the examination paper, and given the relevance of marking to the larger study we also report these comments.

Four respondents left particularly enthusiastic comments about the problem solving focus, for example:

Love the paper and the focus on functional Mathematics. Students initially will be disadvantaged as I am unsure to what extent functional mathematics is embedded within schools. This style would 'force' the adoption of developing what is the most neglected element of the Mathematics curriculum.

Other respondents were more reserved, and stated that the focus on problem solving would be beneficial for some students but disadvantageous to others.

Thirty-four respondents provided open-text responses about mathematical content and in the main were negative, expressing concern that the relatively low amount of mathematical content would not prepare students for later study. Some teachers expressed this strongly, for example: "Where is the assessment of mathematical rigour? This obsession with functionality ignores the need for study of algebraic manipulation as training for further study." We feel some sympathy towards this view, and note that survey respondents were not informed of the constraint imposed on question writers that the examination paper would be administered to students near the beginning of GCSE study, and therefore could be expected to contain less mathematical content than an examination paper designed to be taken at the end of GCSE study.

Most of the teachers expected their students would perform worse on the examination paper than on a typical GCSE examination paper. This was an interesting finding in light of the teachers' overall consensus that the examination paper was relatively light on mathematical

content, and appeared to be due to concern about the literacy demand of the examination paper. In total, 23 teachers expressed concern about the presumed literacy required of students, for example: “The literacy needs are quite high. There is [sic] a lot of questions that require a strong level of literacy. The literacy level is above the mathematical level.” Seven of the teachers who commented on the literacy skills required by the examination paper were concerned about weakly performing students, or students for whom English is a second language.

Twelve respondents commented on how difficult the examination paper would be to mark, for example: “Marking would also be difficult due to the range of possible answers—there couldn’t be a standardised answer for many of the questions”. It was notable that marking arose in the open-text comments because the online survey did not inform participants about the underlying rationale for the examination paper (that it lacked a mark scheme).

In summary, the teachers who scrutinised the examination paper online responded that it was better at assessing problem solving and worse at assessing mathematical content than a typical GCSE examination paper. They also expected that their students would perform less well than on a typical GCSE examination paper, and this seems to be due to concerns about the level of literacy required to access the questions. In addition, some respondents commented on the difficulty of marking such an examination paper.

## 6 Assessment phase

In the assessment phase of the study, the examination paper was administered to high-school students and then marked and comparatively judged, and the outcomes compared. Following this, the experts who undertook the CJ process were surveyed to obtain feedback about their experiences of assessing the students’ work to provide insights on construct validity. The steps taken to assess the student work are shown in Table 1.

### 6.1 Examination paper administration

The examination paper was administered to 750 school students aged 14 or 15 from across two high schools. Both schools were large and located in medium-sized towns in the midlands of England, and were sourced using the authors’ existing contacts. The overall socio-economic background of students was above the national average and the number of students from ethnic minorities was below the national average. The GCSE results for mathematics were at the national average for one school and above the national average for the other school. Teachers were requested to administer the examination paper to all Year 10 pupils who were present on a particular day to ensure a spread of prior mathematical achievement. Most of the students ( $N=745$ ) were candidates at the start of the 2-years GCSE course with predicted grades ranging from F (lowest possible) to A\* (highest possible), and the remaining five were not studying GCSE mathematics due to poor achievement.

The examination paper was administered to each class of students in a regular mathematics lesson by their usual teacher. Students were allowed 50 minutes to complete the examination paper and were allowed the use of calculators. Following this, the students’ work was anonymised and scanned for assessment.

## 6.2 Marking

The marking procedure used for the study was designed as a research tool to estimate content validity. As such the procedure differed from typical marking procedures used for routine educational assessment in several ways. The mark scheme assumed a high level of experience and competence, often instructing the marker to “use your judgement” (see Fig. 3). As such three teachers with at least 10-years’ classroom experience each were commissioned to mark the students’ work. The markers were requested to spend two hours familiarising themselves with the examination paper and mark scheme, and to mark a sample of 19 students’ work before undertaking the work. Anonymised and unmarked hardcopies of the students’ work were provided, and marks were recorded for each question on provided marking sheets. To obtain an estimate of inter-rater reliability for the marking, we commissioned a fourth highly experienced teacher to mark a randomly selected sample of 250 students’ work.

The range of the 750 marks was 0 to 50. The distribution was approximately normal and the internal consistency was acceptably high (Cronbach’s  $\alpha = .720$ ). This distribution of the sample of 249 marks (this should have been 250 but one student’s work was accidentally skipped by the marker) was again approximately normal, and the internal consistency acceptably high (Cronbach’s  $\alpha = .729$ ). Inter-rater reliability was measured by calculating the Pearson product-moment correlation coefficient for the subset of 249 students across the two groups of markers, and was found to be high ( $r = .91$ ). Criterion validity of the marking was estimated by correlating the marks with students’ predicted GCSE grades and was found to be high for both the full set of 750 students ( $r = .72$ ), and for the remarked subset of 249 students ( $r = .73$ ).

These findings provide reassurance that the unconventional marking procedures adopted exceeded the outcomes that would be expected from traditional marking for this style of examination paper, and resulted in a reliable research tool that could be used as a basis to evaluate the criterion validity of CJ outcomes.

## 6.3 Comparative judgement

The implementation of CJ used for the study was supported by TAG Development’s *e-scape* system (Derrick, 2012), which presents pairs of students’ work online via an internet browser and the examiner selects either the left or right student’s work by clicking a button. The 750 pieces of student work were scanned and uploaded to the *e-scape* system. All the student work was presented to examiners unmarked and anonymised to avoid examiner bias (Murphy, 1979).

Twenty-three mathematics education professionals, referred to here as judges, were recruited to assess the students’ work using CJ. Unlike the markers, who were all highly experienced teachers, the judges had varied backgrounds and years of experience, and varied from first-year PhD students with one year’s experience in the classroom, through to teachers with ten or more years’ experience in the classroom. The judges were selected from the authors’ existing contacts, and their variation of skills and experience was sought to reflect the typical variation of a large group of markers. Three of the judges withdrew from the study before completion and subsequently their judgements are not included in the analysis, leaving a total of 20 judges.

To prepare for the CJ procedure, the judges were first sent a copy of the examination paper and asked to complete all the questions themselves. They were not provided with a copy of the mark scheme to ensure pairwise judgement decisions were not based on aggregated marks for

question parts. Eleven judges attended a 30-min training workshop where a researcher presented the rationale of CJ and demonstrated how to make pairwise judgements online. The judges were told to decide for each presented pairing which student they considered the “most mathematically able” based on the evidence in front of them. For the remainder of the training workshop the judges then practiced making judgements. The nine judges who were unable to attend the workshop received one-to-one training either face-to-face or remotely via videoconferencing software.

Fifteen of the judges were each assigned between 250 and 300 judgements, totalling 3,607 judgements. The judges were paid an hourly rate that assumed an average for 50 judgements per hour per judge. They were informed that in order to complete their judgements within the allocated timeframe they needed to develop sampling or other time-saving strategies when judging pairs of students’ work in order. Possible strategies were discussed during training such as focusing on particular questions, focusing on aspects of several questions, and taking into account how many questions students had attempted.

To obtain an estimate of the inter-rater reliability for the CJ procedure, the remaining five mathematics education professionals were recruited to carry out CJ on a randomly selected sample of 250 students. The students were the same 250 used to estimate inter-rater reliability for marking. The judges completed 250 judgements each, totalling 1,250 judgements across all five judges.

## 6.4 Outcome of the CJ procedure

The 3,607 judgements of all 750 students’ work were fitted to the Bradley–Terry model using a maximum likelihood estimation procedure (Turner & Firth, 2005). This produced an estimated parameter ( $z$ -score) for every student enabling the construction of a scaled rank order of students’ work from “best” to “worst”. This procedure was repeated for the independently judged subset of 250 students’ work.

To obtain an estimate of the inter-rater reliability of the CJ procedure, we calculated the Pearson product-moment correlation coefficient for the subset of 250 students’ work in the two scaled rank orders, which was high ( $r = .86$ ). This correlation is similar to those reported in the literature, which typically range from about .80 to .99 (Murphy, 1982; Newton, 1996; Willmott & Nuttall, 1975), suggesting the CJ procedure produced consistent outcomes across independent groups of judges.

## 6.5 Criterion validity

To evaluate the criterion validity of the CJ process, we correlated the parameter estimates with marks and students’ predicted GCSE grades. The correlation between the parameter estimates and marks was high for all 750 students ( $r = .86$ ). To establish the replicability of the validity measure, we also calculated the correlation between the parameter estimates and marks for the reassessed sample of 249 students, which was also high ( $r = .89$ ). To further investigate criterion validity, we correlated the parameter estimates with students’ predicted GCSE grades. The correlation was high for both the full set of 750 students ( $r = .71$ ) and for the rejudged subset of 250 students ( $r = .76$ ). Taken together these measurement results suggest that the CJ produced an assessment outcome that was reliable, and provide evidence in support of validity.

## 6.6 Judging processes

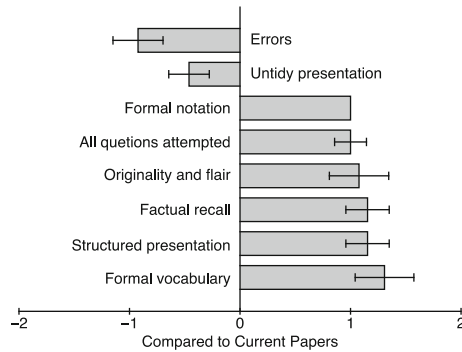
We were interested in unpacking construct validity by considering the strategies used by the judges when deciding one student's work was better than the other. We therefore requested them to complete an online survey after the judging was complete. The survey asked judges to compare two students' work presented online to stimulate recall of the judging experience. Two pieces of work were chosen that were close together in the final rank order, around the 75th percentile, to ensure they were of similar quality and that the simulated judgement was effortful without an obvious "correct" answer. The judges were then presented with the following eight "features" of students' work:

1. Student displays originality and flair;
2. Presence of errors;
3. Use of formal notation;
4. Untidy presentation;
5. Structuredness of presentation;
6. All questions attempted;
7. Student displays good factual recall;
8. Use of formal mathematical vocabulary.

The list of features was derived from our reading of the literature on examiner processes (Crisp, 2008; Pollitt & Murray, 1996; Suto & Greateorex, 2008; Suto & Nadas, 2009) as well as our scrutiny of the students' work. The judges were asked to "indicate the influence of the listed features when judging" using a five-point Likert scale from  $-2$  ("Strong negative influence") to  $+2$  ("Strong positive influence"). This was followed by three open-text prompts that read: "Please state any other features you think may have influenced you when judging pairs of [students' work]"; "Please comment on the quality and suitability for judging of the examination paper"; "Please comment on your overall experience and feelings about the judging process".

Thirteen of the 20 judges involved in the study completed the survey and the results are shown in Fig. 5. A mean rating of 0, as shown by the vertical line, indicates no influence. As for the teacher survey reported earlier, a one-sample Wilcoxon signed-rank tests revealed that the mean rating for all but one of the features was significantly different from the rating representing no influence at the  $p < .005$  level (Bonferroni adjusted). The exception was "untidy presentation" which was rated as influential at the  $p = .020$  level. This was in line with our expectation that judges would rate the six positive features as positively influencing their decisions (rating means ranged from 1.00 to 1.31), and rate the two negative features as negatively influencing their decisions (means  $-0.92$  and  $-0.46$ ).

A normal distribution of ratings could not be assumed and so a non-parametric test was chosen to investigate differences in ratings across the six items. A Kruskal-Wallis one-way analysis of variance revealed no significant difference between the items rated as positively influential on their judgement decisions,  $p = .645$ . This suggests the three explicitly mathematical features (formal notation, factual recall, formal vocabulary) were no more influential than the three generic features that are arguably not mathematical (originality and flair, structured presentation, all questions attempted). That is, it seems the judges were as impressed by positive non-mathematical features, such as presentation, as they were by positive



**Fig. 5** Mean judge ratings ( $N=13$ ) for the influence of eight “features” of students’ work on their judging decisions (*error bars* represent  $\pm 1$  standard error of the mean; all responses to “formal notation” were precisely 1).  $-2$  represents “strong negative influence” and  $+2$  represents “strong positive influence”. The vertical line at 0 indicates “no influence”

mathematical features of the work. However, we acknowledge that given the small sample of respondents ( $N=13$ ) no detailed conclusions regarding construct validity can be drawn.

We turned to the open-text responses for further influences suggested by the judges, although only found one suggestion not covered by the eight closed items. This was the presence of irrelevant comments by students, illustrated by the following judge comment: “If they made a rude comment about the question (‘this is such a silly question’) or drew a silly picture then I found it hard not to be negative towards them!” However, we were surprised that the judges suggested only one further influence in the open-text feedback. It is unlikely that our survey covered all possible influences and perhaps stimulated recall is not a thorough method for establishing how judges make their decisions.

Similarly, we were surprised that only one judge described a sampling strategy for making a quick decision when comparing a pair of 11-page examination papers: “I looked at the first 3 questions first and then backwards from the last question.” One other judge commented that it was difficult to sample “because I felt I wanted to read the whole paper.” The issue of the length of the assessments and the short time judges were given to make their decisions is discussed further below.

## 7 Discussion

Overall, the CJ approach to assessing mathematical problem solving was successful. The CJ procedure yielded an assessment outcome that had high inter-rater reliability. A resource-intensive marking procedure was undertaken in order to help evaluate the CJ procedure. We found that the parameter estimates resulting from the CJ procedure correlated strongly with marks, suggesting good criterion validity. The rank order also correlated strongly with students’ predicted mathematics GCSE grades further supporting criterion validity. In addition, the CJ procedure yielded a high inter-rater reliability when a sample of students’ work was judged by an independent group of examiners. Taken together these findings suggest that the CJ approach to assessing problem solving is reliable, and has good content and criterion validity.



We found that GCSE examination paper writers, when briefed to put marking out of mind, produced an examination paper that contained more open-ended, less structured questions than is typical in current GCSE mathematics examination papers. Survey data from mathematics teachers suggested the examination paper better assessed problem solving but contained less mathematical content than is currently typical in GCSE mathematics examination papers. This may have arisen due to the design constraints imposed on the question writers, and in particular that the examination paper was administered to students near the start of the two year GCSE course. Lower mathematical content does not appear to be an inherent constraint of the CJ process, which has been successfully applied to traditional mathematics exams (Bramley et al., 1998; Jones et al., 2014), as well as undergraduate multivariate calculus (Jones & Alcock, 2014).

The shift from short to more sustained examination questions is consistent with current trends around the world towards assessments that better test deep understanding and problem solving (e.g., Duncan, 2010; Gewertz, 2012; Truss, 2012). A traditional barrier to better mathematics assessments has been the need for affordable and objective tests on a large scale (Berube, 2004; Black et al., 2012). The design phase of the study, along with the evaluation of the examination paper by mathematics teachers, suggests that removing the constraint for reliable marking can free up examination paper designers to produce more open and sustained examination questions. This finding offers a way forward to support the assessment of problem solving and contextualised approaches to mathematics assessment (e.g., MEI, 2012).

The use of CJ for assessing mathematics has implications for how examination questions and tasks are designed. We found that GCSE question writers, when freed from marking considerations, produced an examination paper that was problem-based and relatively unstructured. This finding has important implications for consequential validity. It has been argued that standardised high-stakes assessments stimulate “teaching to the test” practices (Popham, 2001). High-stakes examination papers that are more closely aligned to the stated intentions of curricula to promote problem solving, creativity and sustained mathematical reasoning might positively influence teaching practice.

However, the examination paper may not have been entirely appropriate for being assessed using CJ on three counts. First, the examination paper was 11 pages long, and therefore the evidence available at each judgement across both students’ work totalled 22 pages over several questions. Making a holistic decision about comparative mathematical performance on the basis of such lengthy evidence presents an onerous challenge. Although the results presented here demonstrate that in terms of our measures of inter-rater reliability and criterion validity the judges rose to that challenge successfully, we question whether lengthy exams are the most appropriate for the CJ approach. Moreover, judges had to make their judgement decisions relatively quickly, at a pay rate assuming an average of 50 pairwise judgements per hour, in order to complete the judging work within a reasonable timeframe. Shorter exams and adequate time to absorb all the evidence may be a preferable way forward.

A second design issue, related to the length of the examination paper, was that it contained several mathematical constructs or dimensions. For example, one question (“Good old days?”) was statistical, another (“Money, money!”) was geometrical, and so on. Multidimensionality is typical in examination papers and it is usual to summarise a student’s performance across all these mathematical areas with a single mark or grade. However, multidimensional examination papers may not be the optimal design for CJ where examiners are required to make binary comparisons of whole exams. Shorter tests that focus on a single mathematical construct may be more appropriate (e.g., Jones & Alcock, 2014; Jones et al., 2013).

Third, one promise of CJ is for assessing evidence of student achievement that cannot be marked reliably. In the present study, we used marking as a benchmark for evaluating criterion validity. Although we used a resource-intensive approach, drawing on experienced teachers and not following typical procedures, the examination paper was nevertheless marked reliably. This may be because despite the examination paper's focus on problem solving and unstructured questions relative to present GCSE examination papers, it still resembled a traditional school mathematics examination paper. CJ offers the promise for more open and less structured tasks such as asking candidates to say everything they understand about a specific mathematical idea, using words, diagrams and mathematical symbols (Jones et al., 2013).

## 8 Final remarks

The findings reported here open new possibilities for how school mathematics might be assessed in the future. In this study, we have demonstrated how CJ might impact the design and assessment of written examination papers. Our findings raise the possibility of designing assessments that elude being marked entirely. A richer diet of assessment than is presently used might include practical work, coursework, computer-based activities and oral examination (e.g. ACME, 2005; Black, 2008). CJ offers a possible avenue towards enabling the design and reliable use of such open and diverse assessment methods.

A possible application of CJ not addressed here is its potential as a teaching tool. CJ has successfully been applied to peer assessment, in which students judge one another's work (Jones & Alcock, 2014; McMahon & Jones, 2014), and the role of using example student work for developing problem solving skills is receiving increased interest (e.g., Silver, Ghouseini, Gosen, Charalambous, & Font Strawhun, 2005). A teacher might encourage discussion about what makes a good solution to an unstructured mathematical problem without reference to mark schemes, potentially leading to the kinds of mathematical learning that are currently valued and sought.

Finally, although our interest here has been specifically in the potential of CJ for contributing towards the transformation of mathematics assessments, we believe our findings generalise in principle to a wide range of disciplines and performance types. Indeed the development of CJ for educational assessment has involved a diversity of disciplines ranging from design and technology (Kimbell, 2012) to narrative writing (Heldsinger & Humphry, 2010). Therefore, CJ may offer the potential to enable the assessment of rich and authentic educational outcomes in a wide variety of subject areas and contexts.

**Acknowledgments** This work was supported by a Royal Society Shuttleworth Research Fellowship to IJ, a Royal Society Worshipful Company of Actuaries Research Fellowship to MI, and the Nuffield Foundation.

## References

- ACME. (2005). *Assessment in 14–19 Mathematics*. London: Advisory Committee on Mathematics Education.
- ACME. (2011). *Mathematical needs: Mathematics in the workplace and in higher education*. London: Advisory Committee on Mathematics Education.
- AQA. (2010). *GCSE Foundation Tier Mathematics Paper 1 (Specification A)*. Monday 7 June 2010. Manchester: Assessment and Qualifications Alliance.
- Berube, C.T. (2004). Are standards preventing good teaching? *Clearing House*, 77, 264–267.

- Black, P. (2008). Strategic decisions: Ambitions, feasibility and context. *Educational Designer*, 1(1). Retrieved from <http://www.educationaldesigner.org/ed/volume1/issue1/article1/>
- Black, P. at al. (2012). High-stakes examinations to support policy. *Educational Designer*, 2(5). Retrieved from <http://www.educationaldesigner.org/ed/volume2/issue5/article16/>
- Borsboom, D., Mellenbergh, G.J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.
- 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: Qualifications and Curriculum Authority.
- Bramley, T., Bell, J., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. *Education Research and Perspectives*, 25, 1–24.
- Burkhardt, H. (2009). On strategic design. *Educational Designer*, 1(3). Retrieved from <http://www.educationaldesigner.org/ed/volume1/issue3/article9/>
- Crisp, V. (2008). Exploring the nature of examiner thinking during the process of examination marking. *Cambridge Journal of Education*, 38, 247–264.
- Cronbach, L.J. (1988). Five perspectives on the validity argument. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale: Lawrence Erlbaum Associates, Inc.
- Derrick, K. (2012). Developing the *e-scape* software system. *International Journal of Technology and Design Education*, 22, 171–185.
- Duncan, A. (2010). *Beyond the bubble tests: The next generation of assessments*. Alexandria, VA, Secretary Arne Duncan's Remarks to State Leaders at Achieve's American Diploma Project Leadership Team Meeting. Retrieved from <http://www.ed.gov/news/speeches/beyond-bubble-tests-next-generation-assessments-secretary-arne-duncans-remarks-state-l>
- Gewertz, C. (2012). Consortia provide preview of common assessments. *Education Week*, 32, 18–19.
- Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, 37, 1–19.
- James, C. (1974). The consistency of marking a physics examination. *Physics Education*, 9, 271–274.
- 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: PME.
- Jones, I., Swan, M., & Pollitt, A. (2014). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, 13, 151–177.
- Kimbell, R. (2012). Evolving project *e-scape* for national assessment. *International Journal of Technology and Design Education*, 22, 135–155.
- Koretz, D. (2008). *Measuring up: What educational testing really tells us*. Cambridge: Harvard University Press.
- Laming, D. (1984). The relativity of “absolute” judgements. *British Journal of Mathematical and Statistical Psychology*, 37, 152–183.
- McLester, S., & McIntire, T. (2006). The workforce readiness crisis: We're not turning out employable graduates nor maintaining our position as a global competitor—why? *Technology and Learning*, 27, 22–28.
- McMahon, S., & Jones, I. (2014). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles Policy and Practice*. doi:10.1080/0969594X.2014.978839
- McVey, P.J. (1976). The “paper error” of two examinations in electronic engineering. *Physics Education*, 11, 58–60.
- MEI. (2012). *Integrating mathematical problem solving: Applying Mathematics and Statistics across the curriculum at level 3. End of project report*. London: Mathematics in Education and Industry.
- Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, 35, 1012–1027.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18, 5–11.
- Murphy, R. (1979). Removing the marks from examination 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.
- Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: Sage.
- Noyes, A., Wake, G., Drake, P., & Murphy, R. (2011). *Evaluating Mathematics pathways final report. DfE Research Report 143*. London: Department for Education.
- Ofsted. (2008). *Mathematics: Understanding the score*. London: Office for Standards in Education.

- Pollitt, A. (2012a). The method of adaptive comparative judgement. *Assessment in Education: Principles Policy and Practice*, 19, 281–300.
- Pollitt, A. (2012b). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22, 157–170.
- 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* (pp. 74–91). Cambridge: Cambridge University Press.
- Popham, W.J. (2001). Teaching to the test? *Educational Leadership*, 58, 16–20.
- QCA (2007). *National curriculum 2007*. Coventry: Qualifications and curriculum authority.
- Research Committee, N.C.T.M. (2013). New assessments for new standards: The potential transformation of mathematics education and its research implications. *Journal for Research in Mathematics Education*, 44, 340–352.
- 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.
- Shepard, L.A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16, 5–24.
- Silver, E.A., Ghouseini, H., Gosen, D., Charalambous, C., & Font Strawhun, B.T. (2005). Moving from rhetoric to praxis: Issues faced by teachers in having students consider multiple solutions for problems in the mathematics classroom. *Journal of Mathematical Behavior*, 24, 287–301.
- Suto, I. (2013). 21st Century skills: Ancient, ubiquitous, enigmatic? *Research Matters: A Cambridge Assessment Publication*, 15, 2–8.
- Suto, 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.
- Suto, I., & Nadas, R. (2009). Why are some GCSE examination questions harder to mark accurately than others? Using Kelly's repertory grid technique to identify relevant question features. *Research Papers in Education*, 24, 335–377.
- Swan, M. (2014). Improving the alignment between values, principles and classroom realities. In Y. Li & G. Lappan (Eds.), *Mathematics curriculum in school education* (pp. 621–636). Dordrecht: Springer.
- Swan, M., & Burkhardt, H. (2012). Designing assessment of performance in mathematics. *Educational Designer*, 2(5). Retrieved from <http://www.educationaldesigner.org/ed/volume2/issue5/article19/>
- Taggart, G.L., Phifer, S.J., Nixon, J.A., & Wood, M. (1998). *Rubrics: A handbook for construction and use*. Lancaster: Technomic Publishing.
- Thurstone, L.L. (1927). A law of comparative judgement. *Psychological Review*, 34, 273–286.
- Truss, E. (2012). *Elizabeth Truss calls for a renaissance in maths*. Norfolk: Speech to the National Education Trust. Retrieved from <https://www.gov.uk/government/speeches/elizabeth-truss-calls-for-a-renaissance-in-maths>
- Turner, H., & Firth, D. (2005). Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, 12(1). Retrieved from <http://www.jstatsoft.org/v12/i01>
- van Aalst, J., & Chan, C.K.K. (2007). Student-directed assessment of knowledge building using electronic portfolios. *Journal of the Learning Sciences*, 16, 175–220.
- 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., et al. (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.
- Wiliam, D. (2001). Reliability, validity, and all that jazz. *Education 3–13: International Journal of Primary Elementary and Early Years Education*, 29, 17–21.
- Willmott, A.S., & Nuttall, D.L. (1975). *The reliability of examinations at 16+*. London: Macmillan Education.