Measuring conceptual understanding in randomised controlled trials: can comparative judgement help?

Ian Jones*, Marie Bisson+, Camilla Gilmore* and Matthew Inglis*

* Mathematics Education Centre, Loughborough University
+ School of Applied Social Sciences, De Montford University
Abstract

An impediment to conducting high-quality quantitative research studies in education is the paucity of valid measures of learning gains. Studies often seek to investigate students’ deep, conceptual understanding yet many measures assess only surface, procedural understanding. One reason is that the development of validated measures of conceptual understanding is resource intensive, time consuming, and success is not guaranteed. We evaluated a novel and efficient technique, based on comparative judgement, for assessing conceptual understanding. We applied the technique to a randomised controlled trial in which students were taught simple algebra based on either the Grid Algebra or the MiGen software package. The participants were Year 5 students ($N = 188$) drawn from four primary schools who had not encountered algebra previously. An instrument from the literature (Concepts in Secondary Mathematics and Science: Algebra Scale), and a novel comparative judgement assessment were administered following the intervention. Students in the Grid Algebra condition outperformed those in the MiGen condition on both post-test measures. The comparative judgement technique performed similarly to the standard instrument but was far more efficient to design and implement. The technique can, in principle, be quickly applied to any target concept of interest. We conclude that comparative judgement is a valid, reliable and practical tool that could help to increase both the quantity and quality of quantitative research in education.

Keywords: Assessment, Comparative Judgement, Quantitative Research Methods, Mathematics Education
Background

The past two decades of educational research in the United Kingdom has seen an increase in the use of quantitative research methods. Evidence for this increase comes from regular government accountability exercises for evaluating the quality of research (e.g. Higher Education Funding Council for England, 2008, 2015), and some authors have argued that these exercises are also the drivers for the increase (Gilroy & McNamara, 2009; Marques et al., 2017). Similar exercises are increasingly common around the world (Hicks, 2012), some of which are influenced by the United Kingdom model (e.g. Karlsson, 2017). It is therefore reasonable to expect increased take up of quantitative methods in educational research around the world over the coming years.

However, reports of the outcomes of research evaluation exercises note that quantitative research studies are not always of the highest quality (e.g. Higher Education Funding Council for England, 2015). In this paper we address a key barrier to producing high-quality quantitative evidence in education research: the paucity of valid and reliable assessments of students’ conceptual understanding of learning content (Bisson et al., 2016; Laming 1990; Pollitt, 2012). The National Mathematics Advisory Panel (2008) cited the quality of assessments, along with randomisation and low attrition, as one of the three hallmarks of high-quality evidence. Without appropriate measures quantitative research studies, no matter how carefully designed and rigorously conducted, cannot provide robust evidence.

Education research is often concerned with conceptual rather than (or in addition to) procedural understanding (e.g. National Research Council, 2012; The Royal Society, 2014). It follows that researchers need to assess conceptual learning reliably and validly. Procedural
understanding involves the recall and application of facts, the execution of actions to solve problems, and tends not to be generalisable to novel contexts (Rittle-Johnson, Siegler & Alibali, 2001). It is therefore relatively straightforward to define and test. For example, understanding of equation solving procedures can be assessed by scoring how many equations students could correctly solve.

Conceptual understanding, in contrast, relates to fundamental ideas within a coherent network of relationships, and tends to be generalisable (Star, 2005). Unlike procedural understanding, conceptual understanding is difficult to define and assess (Crooks & Alibali, 2014; Lobato, 2008). Two methods are commonly deployed, both of which are resource intensive and not guaranteed to yield valid and reliable results. One approach is to record one-to-one clinical interviews and rate the quality of each participant’s understanding (Ginsburg 1997). A rating rubric must be established which requires consensus amongst experts regarding conceptual understanding in a domain (Zazkis & Hazzan 1998). In addition, every interview must be recorded and rated by a researcher with a subset rated by a second, independent researcher. For the case of large-scale studies this might result in a lag between the first and last interview. Interviews also require skill and consistency, and might not lead to high reliability (Posner & Gertzog 1982).

The second approach to assessing conceptual understanding requires the development and validation of a bespoke instrument. Examples include the Force Concept Inventory (Hestenes, Wells & Swackhamer 1992) and the Calculus Concept Inventory (Epstein, 2007). Such instruments comprise multiple-choice or short-response items that can be scored objectively. Their development requires consulting with experts, administering tests, analysing responses, and refining questions. These activities are undertaken cyclically until the instrument performs reliably and experts agree that it validly measures understanding of
the target domain. As with interviews, the development of concept inventories is resource-intensive and must be repeated for every concept of interest. However, despite painstaking developmental work, instruments do not necessarily perform satisfactorily due to inconsistencies between how conceptual understanding is defined and how it is operationalised (Crooks & Alibali, 2014). Moreover, instruments cannot be assumed to transfer beyond the context in which they were developed. For example, Bisson et al. (2016) found that a sample of items from the US-developed Calculus Concept Inventory (Epstein, 2007) was not internally consistent when administered in England.

In sum, existing methods for measuring students’ conceptual understanding are inefficient and do not guarantee robust data. Next we outline an innovative method, known as comparative judgement, which may be able to overcome these limitations.

**Comparative judgement**

Comparative judgement (CJ) exploits the finding that people are better at comparing two objects against each other than at evaluating one object against criteria (Thurstone 1927). For example, people are more consistent when judging whether one weight is heavier than another, than when judging the weight of a single object in grams. Thurstone (1927) applied this phenomenon to construct scales of physical sensations, such as weight and temperature, based on participants’ pairwise judgements of which object had the greater amount of the property of interest. Subsequently he applied CJ techniques to construct scales of subjective phenomena such as social attitudes (Thurstone 1954).

CJ methods have since been applied to educational measurement in a variety of contexts, including primary school students’ essays (Heldsinger & Humphry, 2013), secondary school students’ laboratory reports (McMahon & Jones, 2015), undergraduate
mathematicians’ understanding of calculus (Jones & Alcock, 2014), and secondary school students’ design ePortfolios (Kimbell, 2012). The commonality across such studies is using CJ to assess student outputs for which traditional marking procedures are unreliable (Pollitt, 2012). Specifically, marks are typically inconsistent across different assessors for relatively unstructured tasks such as essays and practical work (Newton 1996; Murphy 1982). Conversely, CJ has produced reliable outcomes for unstructured tasks across a variety of disciplines (e.g., Heldsinger & Humphry, 2013; Jones, Swan & Pollitt, 2014; Kimbell, 2012; McMahon & Jones, 2015) by harnessing direct pairwise comparisons of students’ work instead of absolute judgements based on rubrics.

**Research focus**

Despite the promise of the approach, the details of educational studies that use CJ-based assessment have been criticised. Bramley (2015) found that reliability is consistently reported as high but is not always defined. In general, CJ studies use a measure of internal consistency adapted from Rasch modelling that Bramley refers to as Scale Separation Reliability (SSR). SSR is more analogous to Cronbach’s alpha (Pollitt, 2012) than to inter-rater reliability, which involves two or more independent raters marking the same student work, yet some authors have implicitly contrasted SSR results with the latter (Bramley & Vitello, 2018). Moreover SSR can be artificially inflated (Bramley, 2015; Bramley & Vitello, 2018) by using adaptive algorithms (Pollitt, 2012) to select pairs of student work that are presented to assessors. This presents a serious challenge given the widespread argument that CJ enables high reliability where traditional methods fail. To ensure reliability is estimated objectively and transparently, and unlike most CJ studies in the literature (Bramley, 2015), we report both SSR and inter-rater estimates of reliability, and did not employ adaptive algorithms for selecting pairs of student work.
Another problem with the design of CJ-based education studies is that the validity of outcomes is often not considered systematically. Although face validity is commonly addressed through the use of expert review of tasks and responses, and interviews and surveys with assessors and students (e.g. Davies, Collier & Howe, 2012; Hunter & Jones, 2018), investigation of criterion and divergent validity using independent measures is rarer in CJ studies, and where it does occur can be limited to a single large-grain measure (e.g. Jones & Sirl, 2017; Seery, Canty & Phelan, 201). Given that assessors are likely to attend to construct irrelevant features (Crisp, 2013) when assessing the kinds of open-ended tasks that are common to traditional methods but particularly prevalent in CJ-based assessments, a systematic exploration of validity is essential. In the present study we administered a battery of tests and collected existing student achievement data in order to conduct a detailed validity analysis.

Closely related to validity is the issue of the sensitivity of CJ-based assessments. The use of large-grain measures to evaluate criterion validity provides some reassurance, but does not provide evidence that CJ-based assessments are sensitive enough to detect effects of interest to researchers. We addressed this in the present study by applying a CJ technique to a randomised control trial (RCT). The purpose here was to investigate whether the technique produced outcomes that were sensitive enough to detect group differences. To this end we compared two contrasting teaching interventions designed to introduce algebra to primary school children. This provided a context that was of interest to the authors, and for which an existing, validated instrument could be applied. We refer to this instrument as the *algebra scale* and describe it in detail later on. The administration of both a CJ-based assessment and the existing algebra scale enabled us to investigate group differences using the two measures, and so compare their relative performance.
The research questions were primarily methodological, and focused on the appropriateness of the dependent measures used to assess understanding.

1. Is there a difference in conceptual understanding outcomes between the two interventions?
2. Can comparative judgement detect an effect of similar size to that detected by a standardised instrument?

The teaching interventions

We experimentally compared two teaching interventions based on software developed by mathematics education researchers and readily available to teachers in the UK, namely Grid Algebra (Hewitt, 2016) and MiGen (Mavrikis et al., 2013). Both software packages are designed to build on students’ arithmetic knowledge when introducing algebra. The MiGen software presents pictorial and symbolic relationships between variables that students interact with (Mavrikis et al., 2013). An online demonstration video can be viewed at www.youtube.com/watch?v=flHMZNxGa0T4. The Grid Algebra software presents interactive numerals and letters on a grid (Hewitt, 2016) and mathematical expressions are constructed by dragging numerals and letters around the grid according to operational rules. A demonstration video can be viewed at www.youtube.com/watch?v=LaH3S_0igak&t=404s.

Method

Participants

218 students (ages 9 and 10 years) from four primary schools in England participated in the study. The sample size was based on a power analysis assuming dropout of 20%, a medium effect size of $d = 0.5$, 90% power, and no clustering. Participants’ prior mathematics achievement, based on a national scoring scheme, ranged from 2a to 5c (students at level 2
were working at the level expected of a 7 year old, whereas students at level 5 were exceeding
the expectations of an 11 year old). 30 participants (13.8%) were excluded because they were
absent for one or more sessions ($N = 29$) or did not complete all items in the pre-lesson
activities ($N = 1$). The study therefore falls within the National Mathematics Advisory Panel’s
(2008) recommended attrition rate (<20%) for high-quality RCTs. The final sample was 188
students (86 females, mean age = 9.76 years), and within each school students were randomly
assigned to either the Grid Algebra group ($N = 98$) or the MiGen group ($N = 90$). Bias is
minimised if attrition rates are similar between randomised groups (D. Torgerson and C.
Torgerson, 2003), and we found no significant difference in attrition between the Grid
Algebra group (12 participants removed, 10.9%) and the MiGen group (18 participants
removed, 20.0%), $\chi^2 (1, N = 218) = 1.08, p = .300$.

**Ethics and open-access**

Approval was obtained from Loughborough University’s Ethics Approvals (Human
Participants) Sub-Committee prior to conducting the research. Schools were provided with
parental information sheets and opt-out forms, and signed consent was obtained from head
teachers. Anonymised data and analysis scripts can be downloaded from
https://lboro.figshare.com/s/2d4cfa142ebf1ff24841.

**Interventions**

The teaching interventions consisted of three one-hour lessons designed to introduce the use
of letters in algebra. Within each school, one group of children was taught using Grid Algebra
and the other group used MiGen. An educational consultant, with substantial experience of
teaching mathematics, planned and delivered the lessons in collaboration with the research
team. Both software designers were consulted on the design of the lessons. Example lesson
plans provided with the software packages were the basis for our own lesson plans (see
Appendix). Prior to the main interventions the lessons were piloted with students ($N = 39$) aged 9 and 10 years by the educational consultant, and then amended. The children involved in the piloting were not involved in the main study.

The teaching interventions were conducted in each of the participating schools in their computer labs or a classroom. Three one-hour lessons were scheduled for each teaching group in each school, one per day, during the same week. The order of the teaching groups (MiGen or Grid Algebra) was counterbalanced across schools.

**Material and procedure**

*Pre-lesson activities*

A few weeks before the teaching interventions, a researcher visited each school to complete activity booklets with the students. These were completed in small groups and took approximately one hour. The booklets included four activities, each assessing one of the following: writing skills, mathematics anxiety, calculation skills and non-verbal general achievement. All children completed the activities in the same order. There were three purposes to the pre-lesson activities. First, to assess whether randomisation produced any systematic group differences across a variety of measures. Second to allow us to control for individual differences in prior achievement. Third, to identify whether the CJ measure and the algebra scale were influenced by different factors, suggesting that they may have been measuring slightly different constructs

*Writing skills.* The open-ended CJ measure required participants to communicate their understanding, so a standardised writing measure was administered to control for differences in writing skill. We administered the Written Expression subtest of the Welchslr Individual Achievement Test Second UK Edition (WIAT-IIUK; Welchslr, 2005). Children were
presented with the beginning of a sentence (“On a rainy day, I like…”) and asked to continue writing a paragraph for 10 minutes. Paragraphs were marked according to the WIAT-II rubric (maximum 24 points), as follows: spelling and punctuation (9 points); sentence structure, linking expressions and development (10 points); vocabulary (5 points). A subset of all responses (20%) was double marked by a second researcher and reliability was high, $r = .833$.

Mathematics anxiety. We were interested in whether either the CJ measure or the standardised algebra scales might have been influenced by students’ mathematical anxiety. We administered 4 items adapted from the Child Maths Anxiety Questionnaire (CMAQ; Ramirez et al., 2013; items 2, 3, 5 and 7) and 6 items from the Scale for Early Mathematics Anxiety (SEMA; Wu et al., 2012; items 1, 2, 4, 5, 7 and 8). We used questions that concerned “doing maths” rather than social pressures or situational anxiety. Each question was read aloud, and children responded using a scale which ranged from “not nervous at all” to “very, very nervous”, scoring 1 to 5 points. Therefore, children could score from 5 to 50.

Calculation skills. The CJ measure and standardised algebra scale were both designed to assess conceptual understanding, but we expected both to also be related to procedural understanding given the connection between the two (Rittle-Johnson, Siegler & Alibali, 2001). Calculation skills were therefore assessed using the Numerical Operations subtest from the Wechsler Individual Achievement Test II UK Edition (WIAT-IIUK; Wechsler, 2005). This test presents written calculation problems of increasing difficulty including subtraction, addition, multiplication and division. We administered items 8 to 44, these being appropriate for the age of our participants, and children completed as many questions as they could in 15 minutes. One point was allocated for each correct answer, maximum 37 points.
Non-verbal general achievement. To investigate any group differences we administered Raven’s Educational UK Edition Standard Progressive Matrices Plus Version (SPM+; Raven, 2008), which requires participants to find the missing piece of visual puzzles of increasing difficulty. We administered matrices from Sets A, B and C, and children had 15 minutes to solve 36 puzzles, each worth one point.

Post-lesson activities

The post-lesson activities consisted of two worksheets. The first worksheet was an open-ended prompt to use for the CJ, as follows:

“Explain how letters are used in algebra to someone who has never seen them before. You can use examples and writing to help you give the best explanation that you can”.

Children were given 10 minutes to write their explanation.

The second worksheet included 11 items from the Concepts in Secondary Mathematics and Science: Algebra Scale (items 2.1, 2.2, 4.1, 4.2, 4.3, 5.1, 5.2, 5.3, 6.1, 11.1, 11.2; see Brown, Hart & Küchemann, 1984; Hart, 1981). The original full scale was validated for assessing the broad construct of understanding of algebra, and items selected for the present study were intended to assess the more focused construct of understanding of letters in algebra. All the selected items were included and analysed in Küchemann (1978). In addition the selected items were subjected to face validity checks by mathematics education experts, specifically the advisory group which provided feedback to the research team. A principle components analysis was conducted to confirm that the 11-item scale can be considered unidimensional. Children were also given 10 minutes to complete this worksheet. Each correct response scored 1 point. Children completed both worksheets on the afternoon of their final intervention lesson, in the same groups as their teaching intervention. They worked on their own in silence, and completed the worksheets in the same order in all schools.
**Procedure for CJ scoring**

Student responses to the open-ended question were scanned and uploaded to an online CJ platform (www.nomoremarking.com). Ten mathematics PhD students, who were blind to the purpose of the study, were recruited to comparatively judge the student responses. The judges were studying at one university but had internationally diverse cultural and educational backgrounds. Judges were presented with a sequence of pairs of student responses and asked to decide, for each pairing, which student demonstrated the better conceptual understanding of the use of letters in algebra. Each judge completed 189 judgements. The presented pairings of responses were pseudo-randomly selected from all possible pairings, with each response being presented between 19 and 21 times. Each judge viewed all or most of the responses (range: 180 to 189 responses seen by each judge) at least once during their 189 judgements. One participant was excluded after the judging was complete when it was discovered the pre-lesson activities had not been completed leaving 188 responses for analysis.

**Results**

**Pre-lesson activities**

There were no significant group differences on any of the pre-lesson measures.

**Post-lesson activities**

Because of the hierarchical structure of the data (participants were randomly allocated to teaching groups within each school), it was appropriate to use a multi-level linear model with school as a level 2 variable to investigate whether the teaching interventions impacted the results of post-lesson activities differently in each condition (Grid Algebra vs. MiGen). Condition was coded in the analyses as “0” for the Grid Algebra groups and “1” for the MiGen groups.
The decisions from all the judges were submitted to the Bradley-Terry model (Firth, 2005) to calculate parameter estimates for each student response, as is standard for CJ decision data (Pollitt, 2012). The parameter estimates were z-transformed before further analysis. An example of the highest scoring test response for each condition is shown in Figure 1.

FIGURE 1 ABOUT HERE

We assessed the internal consistency of the judging using the Scale Separation Reliability (SSR) coefficient, a measure for scaled rank orders that is considered analogous to Cronbach’s alpha (Pollitt, 2012). Internal consistency was high, SSR = .857. Inter-rater reliability was checked using a split-halves technique (whereby the 10 judges were randomly split into two groups of 5 and the parameter estimates remodelled for each group of judges before being correlated; we repeated this process 20 times). It ranged from $r = .623$ to $r = .760$ (Median = .695). We then calculated the mean parameter estimate for each condition ($M_{\text{Grid Algebra}} = 0.26$, $SD_{\text{Grid Algebra}} = 0.93$; $M_{\text{MiGen}} = -0.27$, $SD_{\text{MiGen}} = 1.01$) before building a multi-level model.

We first verified whether allowing the intercepts to vary across schools improved a basic model (intercepts only). This was the case, $\chi^2(1) = 24.95$, $p < .001$, so we used a random effect for the intercepts for subsequent analyses. We then added the effect of condition in model 2, before adding our pre-lesson measures as fixed-effects predictors in model 3. Both models 2 and 3 fitted the data significantly better than the previous model, $\chi^2(1) = 16.65$, $p < .001$ and $\chi^2(4) = 52.73$, $p < .001$ respectively. Finally, in model 4, we verified whether allowing the slopes for the effect of condition to vary across schools further improved the fit.
and found that it did, $\chi^2(2) = 7.62, p = .022$. Model 4 (see Table 1) therefore includes both random intercepts and random slopes for the effect of condition across schools.

The key result was that, based on the CJ method of measuring conceptual understanding, there was a significant effect of condition. Participants taught using the Grid Algebra software were judged to have performed significantly more strongly on the open-ended post-test than participants taught using the MiGen software, as indicated by the $b = -0.47$ in model 4 of Table 1. The random intercepts and slopes in Model 4 reflect that achievement and the effect of condition varied by school, but overall the effect was significant in favour of the Grid Algebra condition.

**TABLE 1 ABOUT HERE**

*Algebra scale*
We calculated the total number of correct answers on the algebra scale and averaged this across participants for each condition ($M_{\text{Grid Algebra}} = 5.05, SD_{\text{Grid Algebra}} = 2.04; M_{\text{MiGen}} = 4.54, SD_{\text{MiGen}} = 2.03$). We found that the internal consistency of the algebra scale was acceptable to low, $\alpha = .638$, suggesting that the items from the standardised instrument did not perform particularly well. Nevertheless, participants’ results on the algebra scale correlated significantly with their parameter estimates from CJ, $r = .463, p < .001$.

The algebra scale scores were analysed using the same method as for the CJ outcomes. In the first instance we allowed the intercepts to vary across schools, and this significantly improved the basic model, $\chi^2(1) = 22.78, p < .001$. In model 2, we added the effect of condition, and this significantly improved the fit compared to the previous model, $\chi^2(1) = 4.21, p = .040$, as did adding our pre-lesson measures as fixed-effects predictors in model 3,
$\chi^2(4) = 78.71, p < .001$. Finally, in model 4 we investigated whether allowing the slopes to vary for the effect of condition across schools further improved the model, and found that it did not, $\chi^2(2) = 0.28, p = .871$.

The results in Table 2 show that the key finding based on the CJ outcomes was replicated for the algebra scale outcomes. There was again an overall significant effect of condition in favour of the Grid Algebra teaching group, as indicated by the $b = -0.49$ in model 3. Notably, this coefficient is almost identical to the equivalent coefficient found when using the CJ outcomes as the dependent measure.

In sum, for both of our dependent measures we found that the participants who were taught using Grid Algebra outperformed those who were taught using Migen.

**TABLE 2 ABOUT HERE**

**Discussion**

Educational research can only generate robust evidence about learning if student understanding is validly assessed. The development of valid assessments has been a particular challenge where the focus is change in conceptual understanding. For the case of mathematics education, assessing participants’ procedural understanding of a topic is relatively straightforward. However, assessing conceptual understanding of a topic is resource-intensive, and reliable and valid outcomes are not guaranteed.
We explored the use of a novel method, based on CJ, for assessing participants’ understanding of a mathematical concept within the context of an RCT. We choose a target domain, simple algebra, for which an instrument designed to assess conceptual understanding already exists. This instrument provided a yardstick for evaluating the validity of the CJ method. We selected items from the instrument that targeted a specific concept, the use of letters in algebra, and contrasted two approaches to introducing primary students to algebra.

The CJ approach to assessment was successful in this particular case where the focus was on introducing algebra to older primary students. Specifically, both the CJ and traditional approaches to assessing conceptual understanding had sufficient measurement sensitivity to detect a significant group difference at post-test in favour of the Grid Algebra teaching intervention. Importantly, the CJ approach was relatively efficient to develop, administer and evaluate. It matched the performance of the items from the standardised instrument, which were considerably more time-consuming to develop, requiring a year’s iterative empirical work (see Küchemann, 1978). Moreover, we found that the items from the standardised instrument performed rather poorly, $\alpha = .638$, suggesting that the CJ method had better internal consistency than the traditional method. The methodological conclusion from this is that the CJ approach matched the standardised instrument in terms of evaluating the learning gains between conditions within our particular context. Therefore CJ seems to have the potential to offer an efficient and valid approach to assessing student understanding, at least where there is no reliable instrument for the construct of interest.

A key consideration is the generality of the results beyond the context of the study reported here. There is published evidence that specific mathematical concepts can be reliably and validly assessed using CJ techniques across a range of topics and levels including primary school, secondary school and university (Bisson et al. 2016; Hunter & Jones, 2018; Jones et
al. 2013; Jones & Karadeniz, 2016). In addition, CJ assessments have been successfully applied to student artefacts in range of subjects including design and technology ePortfolios (Kimbell, 2012), geography essays (Whitehouse & Pollitt, 2012), chemistry experiment reports (McMahon & Jones, 2015), digitised visual arts portfolios (Newhouse, 2014), history essays (Holmes, Black & Morin, 2018), narrative texts (Heldsinger & Humphry, 2013), and engineering ePortfolios (Williams, 2012). We add to these findings by showing that the CJ approach can be sensitive enough to detect changes in conceptual understanding arising from RCTs, and note that it could plausibly be applied to detecting changes in understanding in other subjects.

Implications

A key motivation for using CJ approaches to measure learning differences when conducting RCTs is their efficiency relative to developing a traditional instrument. Nevertheless, deploying CJ requires non-trivial work, particularly once participants have completed the open-ended test. Here we describe what is involved in using CJ to assess conceptual understanding.

Initially a test prompt is developed. Explanations are central to conceptual understanding (e.g., Friedman, 1974; Wilkenfeld, 2014), and therefore it is unsurprising that humans are more willing to attribute ‘understanding’ of a given concept to a third party (as opposed to mere ‘knowledge’ of it) when they are able to offer a high-quality explanation (Wilkenfeld, Plunkett & Lombrozo, 2016). Given this, we favour questions of the general form “Explain concept X” as used here. Example test prompts covering a range of mathematical concepts are given by Jones and Karadeniz (2016), Hunter and Jones (2018) and Bisson et al. (2016). It is advisable to pilot test prompts prior to deployment, particularly where the prompt is novel, or where participants are drawn from a population for which the
prompt has not previously been used. This is particularly the case for subjects other than mathematics where an “Explain concept X” format may be novel. In the present case a similar test prompt had been used previously with secondary school students (Bisson et al., 2016). We therefore piloted the prompt with a sample of primary school students \((N = 39)\) before conducting the main study. The responses from the pilot were scrutinised by the research team and no subsequent amendments were made to the test prompt.

Once completed, the test responses are scanned and uploaded to the CJ website. To ensure anonymity participants write their name on a cover sheet that can then be removed, and a unique identifier is used for the filenames. Expert judges, in this case mathematics PhD students, are recruited and assigned judgements via the website. Researchers working in subjects other than mathematics need to consider carefully which population to draw judges from. For the case of mathematics we have previously found that educationalists, research students and research-active mathematicians all produce valid and reliable scores when judging students’ work (Jones & Alcock, 2014; Jones & Inglis, 2015; Jones, Swan & Pollitt, 2014). However, an interesting case arose when CJ was applied to assess undergraduates’ understanding of \(p\) values in the context of null-hypothesis significance testing (Bisson et al. 2016). Given the widely documented conceptual difficulties regarding the meaning of \(p\) values (Greenland et al., 2016), we recruited expert psychologists rather than mathematicians on the assumption that the former would have greater expertise regarding \(p\) values.

We have found that, in the context of mathematics assessment, judges do not require training beyond short written instructions (Bisson et al., 2016), and the judging can be undertaken remotely. A possible expense for researchers is the payment of judges where volunteers cannot be recruited. In the present study we spent approximately £300 (about $380) to judge 189 test responses. The total time taken by the judges was approximately 13
hours at an average rate of about two judgements per minute. We offer this as a description of the costs of the present study and not as a design principle for future studies. The resources required to fully assess test responses will depend on the form of the test and the concept being assessed.

In the current study we complemented the post-test with instruments that allowed us to control for individual differences in prior achievement and explore how the outcome measures related to key skills. One variable of interest was written expression, which we thought might impact on the perceived quality of the test responses, particularly for the case of primary school children. As shown in Tables 1 and 2, written expression was found to predict CJ outcomes but not algebra scale outcomes, and the correlation with CJ outcomes was moderate, $r = 0.448$. Therefore, in situations such as here, where the effects of writing skills need to be controlled, researchers should consider using an existing record of writing skills or, if this is not available, administering a writing skills test. There are likely to be situations in which writing skills should be considered integral to an outcome measure, such as an RCT to investigate the impact of an intervention on essay writing. Even in areas such as mathematics, written communication is sometimes inherent to the domain being assessed. For example, clear communication in fully formed sentences is central to learning university level mathematics (Alcock 2012; Houston 2009; Yoon 2017). In such scenarios obtaining or constructing a scale of writing skills for the participants is likely to be unnecessary. In other subjects and contexts researchers need to consider carefully possible confounds and how they might be controlled for.

**Validity**
There are many ways in which validity can be defined and interpreted (Newton & Shaw 2014). Kane (2016) argued that the validation of assessments should focus on the nature and purpose of the test under discussion. Here we consider those dimensions of validity against which we could evaluate our CJ measure, as well as those likely to be relevant to researchers using CJ approaches to evaluate quantitative research studies, such as RCTs.

**Face validity.** A feature of CJ for assessing learning is that open-ended test prompts can be used (Jones, Swan & Pollitt 2014). Here we used a question that explicitly asked participants to explain, in their own words and diagrams, the meaning of a concept. A project advisory panel reviewed the test prompt, and the research team scrutinised participant responses from the pilot study. These processes provided assurance that the test yielded reasonable face validity; that is, the test responses appeared to reflect participants’ understanding of the role of letters in algebra.

Ensuring the face validity of an assessment may come more readily to subjects other than mathematics. For example, the production of a physical artefact is an inherent aspect of creative writing or design technology. Nevertheless, researchers need to take care that the objects being judged transparently reflect the specific target concept of interest when considering using CJ techniques (Pollitt, 2012).

**Internal validity.** We used two methods to evaluate the internal validity of the CJ outcomes. The first was to calculate Scale Separation Reliability, which is commonly reported in studies that use CJ techniques. A high SSR value (> 0.7) suggests that judges were consistent with one another when making pairwise decisions about the test responses. If a test prompt produced participant responses that were relatively homogenous then we would not expect
judges to be able to discriminate reliably between pairs of responses, and this would be reflected in a low SSR statistic.

The second method for evaluating internal validity was the calculation of an inter-rater reliability statistic following the split-halves technique described earlier. This method is costly because it requires the construction of two independent scales and then calculating the correlation coefficient between them. To do this the number of judges and the total number of judgements must be doubled from the minimum required. Based on our experience of using CJ across a range of contexts, we recommend 10 judgements per response to ensure a reasonable SSR, and therefore 20 judgements per response are required to evaluate inter-rater reliability.

The internal validity estimates were reassuring but it is important to consider what can and cannot be inferred from this. We can conclude that the results were reliable across different contemporary judges, but not that they would be reliable over time. This can be relevant where CJ is used to explore educational standards over time (e.g. Jones et al., 2014). In the present study the algebra scale items were developed over four decades ago and results from today can be directly compared with those from the 1970s, albeit taking into consideration curricular and pedagogic changes (Hodgen et al., 2009). We are unable to confirm whether PhD students from four decades ago, or in four decades time, would make the same pairwise judgements as those reported here. Changes of expert judgement over time may be even more of an issue for subjects other than mathematics.

**Criterion validity.** We chose a topic for the teaching interventions, algebra, for which a standardised instrument, the Concepts in Secondary Mathematics and Science: Algebra Scale,

---

1 We are grateful to a reviewer of an early version of the manuscript for raising this point.
already existed. The algebra scale provided a set of outcomes against which we could compare the CJ approach adopted here. As shown in Tables 1 and 2, the CJ outcomes and the algebra scale both revealed a similar and significant effect for condition in favour of the Grid Algebra group. To further investigate the criterion validity of the results we calculated the correlation coefficient between the CJ outcomes and algebra scale and found it was moderate, \( r = 0.463 \).

A standardised instrument need not always be deployed alongside a CJ-based measure. Indeed, a motivation for using CJ is the absence of such a measure for the concept of interest. However, in subjects where no CJ-based studies have yet been published a pilot study might be advisable. One option is to obtain a standardised measure and compare the scores with CJ outcomes as was the case here. Where no standardised measure is available then another option, albeit one that adds time and expense to a study, is to develop a rubric for the open-ended test and compare the marks with CJ outcomes for a sample of student responses. The rubric should be developed iteratively with reference to a sample of student responses, and the marking conducted by two independent experts on a second sample of responses (see Jones & Inglis, 2015). However, this option can only provide a comparison of two methods of assessing the same responses and so is a merely proxy for criterion validity.

**Divergent validity.** As discussed earlier, students’ writing skills predicted CJ outcomes but not algebra scale scores. The correlation between CJ outcomes and written expression, \( r = 0.448 \), was higher than the correlation between the algebra scale and written expression, \( r = 0.351 \), although this difference was not significant, \( z = 1.437, p = 0.151 \). Conversely, the correlation between algebra scale scores and numerical operations, \( r = 0.581 \), was higher than the correlation between CJ outcomes and numerical operations, \( r = 0.485 \), however again this difference was not significant, \( z = 1.589, p = 0.112 \).
Divergent validity can be explored using a variety of techniques depending on the data available. For example, Jones and Alcock (2014) compared the CJ outcomes of judges who were experts (mathematicians) with those of non-experts (arts specialists) for an undergraduate calculus test. Another option is to regress prior achievement data from different subjects onto CJ scores. For example, Jones and Karadeniz (2016) found that prior mathematics but not prior reading scores predicted CJ outcomes for a conceptual mathematics test, demonstrating both criterion and divergent validity. Analogous approaches can be applied in different contexts and for different subjects.

**Concluding comments**

The CJ approach to testing between-groups differences in an RCT was successful, and detected the same effect as a traditional instrument, in the context it was applied here. We propose that the method offers the potential to assess conceptual understanding in quantitative studies across a range of educational subjects. Importantly, implementing a CJ approach is likely to be more efficient than developing and validating a traditional instrument. This relative efficiency means that CJ has the potential to increase both the quantity and quality of quantitative educational research, and so improve the quantity and quality of data that is available to policy makers and practitioners.

**Acknowledgements**

This work was funded by a Nuffield Foundation grant, a Royal Society Shuttleworth Education Research Fellowship awarded to IJ, a Royal Society Dorothy Hodgkin Fellowship awarded to CG, and a Royal Society Worshipful Company of Actuaries Education Research Fellowship awarded to MI.
References


