MEASURING CONCEPTUAL UNDERSTANDING: THE CASE OF FRACTIONS

Ian Jones, Matthew Inglis, Camilla Gilmore — Loughborough University
Jeremy Hodgen — King's College London

Developing measures of the quality of understanding of a given mathematical concept has traditionally been a difficult and resource-intensive process. We tested an alternative approach, called Comparative Judgement (CJ), that is based not on psychometric instruments or clinical interviews but collective expertise. Eight mathematics education experts used CJ to assess 25 student responses to a test designed to probe conceptual understanding of fractions. Analysis revealed the CJ assessment process yielded high internal consistency, inter-rater reliability and validity. We discuss the implications of the results for using CJ to measure mathematical understanding in a variety of domains and contexts.

CONCEPTUAL UNDERSTANDING OF MATHEMATICS

Many scholars distinguish between conceptual and procedural understanding in mathematics education research (e.g., Hiebert & Lefevre, 1986; Shneider and Stern, 2010; Skemp, 1976). Conceptual understanding is commonly associated with deep, flexible knowledge of underlying abstract principles and procedural understanding is commonly associated with operational knowledge required for stepwise problem solving (Star, 2005). In the research reported here our interest is in the measurement of conceptual understanding.

Traditionally there have been two main approaches to measuring conceptual understanding. The first is to develop and psychometrically validate a bespoke instrument to probe students' understanding of a particular content domain such as calculus (Epstein, 2007) or a particular concept such as equivalence relations (Rittle-Johnson, Matthews, Taylor & McEldoon, 2011). However this has the disadvantage of being a painstaking, resource-intensive process that must be repeated for every concept of interest. The second approach to measuring conceptual understanding, which is sometimes combined with the first, is to record one-to-one clinical interviews and analyse or rate the quality of each participant's understanding (e.g., Knuth, Stephens, McNeil, & Alibali, 2006; Piaget, 1952). However clinical interviews have the disadvantage of requiring skill and consistency on the part of the interviewers and raters, and do not always lead to trustworthy results (Posner & Gertzog, 1982).

The expense, lengthiness and difficulty of measuring conceptual understanding is a barrier to progress in mathematics education. Without valid, reliable and efficient measures it is challenging to evaluate the effectiveness of educational interventions, or to resolve debates in the literature such as whether learning via abstract or concrete

2013. In xxxx (Eds.). *Proceedings of the 37th Conference of the International Group for the Psychology of Mathematics Education*, Vol. 1, pp. XXX-YYY. Kiel, Germany: PME.

representations better aids knowledge transfer (de Bock, Deprez, van Dooren, Roelens & Verschaffel, 2011; Kaminski, Sloutsky & Heckler, 2008). In this paper we report a study designed to test a novel method to measuring conceptual understanding, called Comparative Judgement, that offers promise for overcoming the drawbacks of traditional methods.

COMPARATIVE JUDGEMENT (CJ)

The CJ approach to measuring mathematical understanding involves two stages. First participants complete a test question designed to probe their understanding of a particular concept. The test question is likely to be open-ended and allow a wide variety of types of responses from participants. In the study reported here we used a question designed by a teacher for diagnosing teenage students' understanding of fractions, shown in Figure 1.

The second stage of the CJ approach requires mathematics education experts to make pairwise judgements of the quality of the test responses. Each expert is presented with two responses, such as those shown in Figure 2, and asked to decide which is 'better' in terms of a given construct (ties are not allowed), in our case 'conceptual understanding of fractions'. There are no detailed assessment criteria or scoring rubrics and instead each decision is holistic and based solely on the expert's judgement. Many such pairings are presented to several experts and the decisions are statistically modelled (see Methods section) to produce a scaled rank order of test responses from 'worst' to 'best'.

CJ is a well established technique in psychometrics. It derives from the psychological principle that humans are better at comparing two objects against one another than they are at comparing one object against specified criteria (Thurstone, 1927). For example, people are more reliable at stating which of two objects is the heavier than they are at stating how many kilograms a single object weighs. A traditional drawback of CJ is that large numbers of judgements were required to produce a scaled rank order, limiting much past research to the ranking of six or fewer objects. Recent developments in information technology have helped overcome this drawback, enabling the rapid delivery of (virtual) objects for judging by remotely located experts, and making use of algorithms and statistical techniques to reduce the number of judgement decisions required (Pollitt, 2012). Consequently CJ can now be used routinely for educational research (e.g., Kimbell, 2012; Jones & Alcock, 2012) and practice (e.g., Bramley, 2007).

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

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

Figure 1: Test question for assessing understanding of fractions.

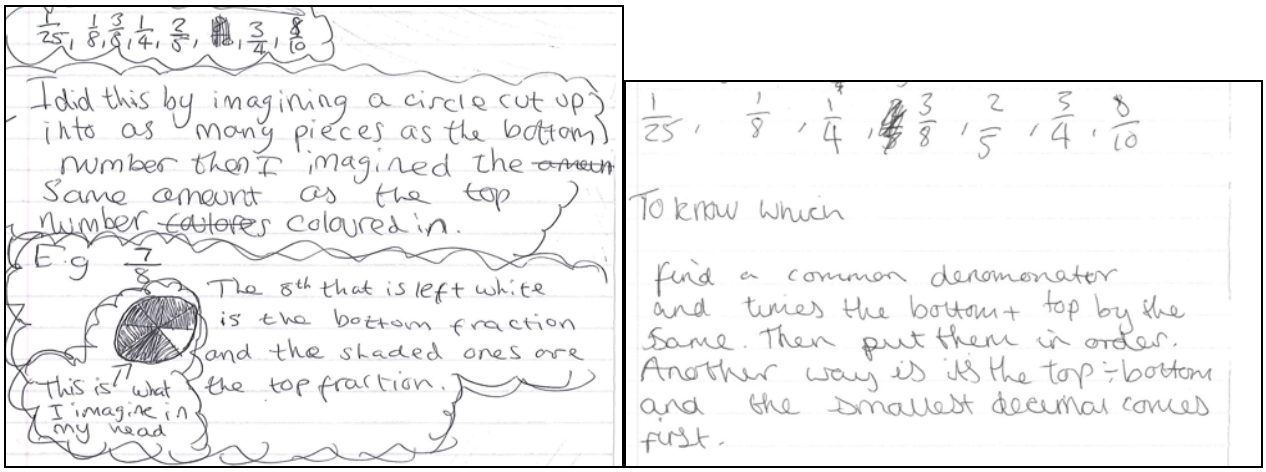

Figure 2: Two example test responses used in the study.

The theoretical strength of the CJ approach is its reliance on collective expertise in the absence of assessment criteria and scoring rubrics. In this sense validity can be thought of as defined in terms of what the experts collectively consider the construct to be in practice. This makes CJ particularly promising for assessing constructs that are recognised and considered important by education experts, such as problem solving and conceptual understanding, but difficult if not impossible to specify comprehensively in rubrics (Laming, 2004). Moreover, constructs such as conceptual understanding lend themselves to open-ended test questions (e.g. Figure 1) that evoke a wide variety of responses (e.g. Figure 2). This variety of responses is difficult to anticipate in rubrics, but is well suited to holistic pairwise judgement by experts. Readers may wish to try judging which of the two responses in Figure 2 they consider to be 'better' in terms of conceptual understanding of fractions.

The practical motivation for studying the use of CJ for assessing conceptual understanding is its potential efficiency. Unlike painstakingly developed psychometric instruments, CJ can be rapidly applied to any target concept with little effort beyond recruiting judges with the requisite expertise. Unlike clinical interviews CJ exploits the long-established psychological principle of pairwise comparisons and yields high validity and reliability with minimal training (e.g., Jones, Swan & Pollitt, submitted).

THE STUDY

In the remainder of the paper we report a feasibility study into using CJ to measure conceptual understanding of mathematics. Eight mathematics education experts comparatively judged the responses of 25 secondary students to the question shown in Figure 1. The experts' decisions were used to construct a scaled rank order of students' responses from 'weakest' to 'strongest' conceptual understanding. Our research goal was to evaluate the method's internal consistency, inter-rater reliability and validity. We conclude the paper by discussing the implications of the findings for measuring conceptual understanding.

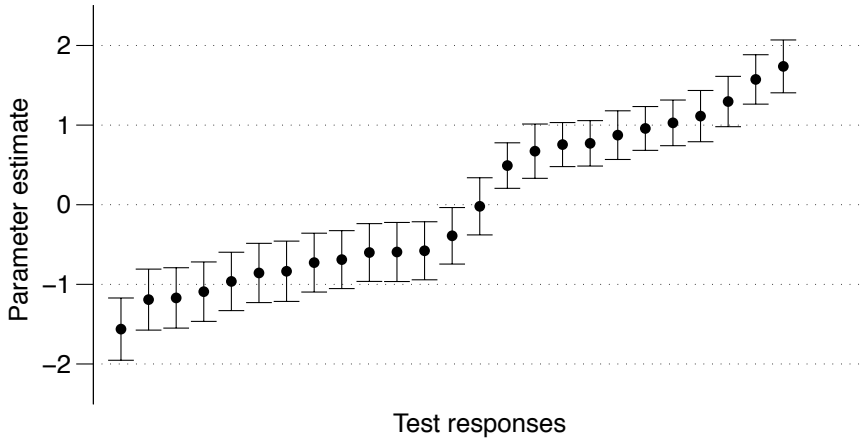

Figure 3: Scaled rank order of test responses from 'worst' (left-most) to 'best' (right-most), with standard errors for the parameter estimation for each response.

METHOD

Materials. The test question used for the CJ exercise is shown in Figure 1. The question was designed by a mathematics teacher in a school in central England who used it with her classes for diagnostic purposes. For each class the teacher wrote the question on the board and allowed them around ten minutes to complete it. We obtained 25 responses from children aged 12 to 15 years for the study, including the two examples shown in Figure 2.

Participants. Eight mathematics education experts (four teachers, two examiners, two research students who were both former teachers) were recruited for the CJ exercise. All had experience of using CJ to assess mathematics from their involvement in previous projects and none required training.

System. Pairs of scripts were delivered to the participants online by TAG Development's *e-scape* system (Derrick, 2012), which uses an adaptive algorithm (Pollitt, 2012) to select scripts in order to minimise the number of judgements required to construct a stable rank order.

Procedure. Each participant completed 20 practice judgements then 50 live judgements online within a two-week window. Only the 400 live judgements (8 participants × 50 judgements) were used in the analysis. For each judgement the *e-scape* system recorded which participant made the judgement, which two scripts were presented, and which script the judge preferred.

ANALYSIS

Rank order. The 400 judgement decisions made by the expert participants were fitted to the Bradley-Terry model using a maximum likelihood estimation procedure (Firth, 2005). This produced a *z*-score parameter and standard error for each test response, shown in Figure 3. We then assessed these parameters in terms of internal consistency, inter-rater reliability and validity.

Internal consistency. We conducted three checks on the internal consistency of the rank order (Pollitt, 2012). First we calculated the Rasch Separation Reliability Coefficient, often considered directly analogous to Cronbach's alpha for traditional test items (Bond & Fox, 2008), and found it was acceptably high (.88). Next we calculated and standardised judge 'misfit' figures, which give an estimate of the consistency of a given judge's decisions with the final rank order. A common guideline is to consider those judges whose misfit figures are less than two standard deviations above the mean (i.e. $z < 2$) to be performing consistently. We found that the judges' misfit figures were all well within two standard deviations of the mean suggesting the judges performed consistently. Similarly we calculated test response misfit figures to provide an estimate of how consistently each response was judged relative to the final rank order. The scripts' misfit figures were all well within two standard deviations of the mean, bar one response that was marginally above the threshold ($z = 2.07$). Taken together the Rasch Separation Reliability Coefficient, judge misfit figures and response misfit figures indicate that the final scaled rank order was internally consistent.

Inter-rater reliability. Inter-rater reliability measures the extent to which the same rank order would have been produced by a different group of expert judges drawn from the same population. To measure inter-rater reliability we split the eight judges into two groups of four and used their judgements (200 per group) to construct two new separate scaled rank orders. We then calculated Pearson's product-moment correlation coefficient between the two sets of estimated parameters. We repeated this process 36 times, once for every possible unique split of the eight judges into two groups of four, to produce 36 estimates of inter-rater reliability. We found that the correlation coefficients ranged from $r = .79$ to .95 and the mean was $r = .87$, suggesting an acceptably high inter-rater reliability.

Validity. We explored the validity of the CJ assessment process in terms of the correlation of the outcomes with measures of students' general mathematical achievement, and their performance on the fractions task measured in purely procedural terms (the extent to which they correctly ordered the fractions).

General mathematical achievement was measured by teacher estimates of ability. For the fifteen oldest (13-15 years old) students in the study predicted grades for the terminal mathematics examination in England (GCSE) were available. These ranged from A* (highest) to F (lowest). For the ten youngest students (12-13 years old) we obtained a dichotomous (high/low) assessment of their ability from their class teacher.

Procedural performance on the task was assessed by calculating the difference between the correct ordering and that given by each student using the Levenshtein distance metric. This is a calculation of the number of steps required to correct a sequence (Levenshtein, 1966). This produced a score for each student's ordered

fractions that ranged from 0 (fractions correctly ordered) to 7 (fractions very out of order).

If our CJ approach was measuring something beyond procedural understanding of the fractions task, then we would expect the teachers' assessments of students' mathematical achievement to be better predictors of the CJ parameters than the procedural accuracy scores. To investigate this we conducted multiple regression analyses predicting CJ parameters with general mathematical achievement and procedural accuracy scores. In view of the differing measures of general mathematical achievement, this analysis was conducted separately for the older and younger students.

For the 15 oldest students we found that the two predictors explained 53% of the variance in the parameter estimates, $R^2 = .53$, $F(2, 12) = 6.69$, $p = .011$. Mathematical achievement (predicted grade A* to F) significantly predicted parameter estimates, $\beta = .40$, $t(12) = 2.64$, $p = .022$, but Levenshtein distance was not a significant predictor, $\beta = -.07$, $t(12) = -.519$, $p = .613$. Similarly, for the ten youngest students we found that the two predictors explained 68% of the variance in the parameter estimates, $R^2 = .68$, $F(2, 7) = 7.33$, $p = .019$. Mathematical achievement (high or low) significantly predicted parameter estimates, $\beta = 1.38$, $t(7) = 3.83$, $p = .006$, but as with the older children Levenshtein distance was not a significant predictor, $\beta = -.23$, $t(7) = -1.84$, $p = .108$.

In sum, for both groups of students we found that teachers' assessments of mathematical achievement were better predictors of students' CJ parameters than was a measure of the procedural accuracy on the same task. This provides some evidence to indicate that the CJ method was measuring something other than procedural understanding, and the relationship with general mathematical achievement is consistent with the suggestion that it was measuring conceptual understanding.

DISCUSSION

We tested an approach to measuring conceptual understanding based on the collated holistic judgement of experts. Traditionally the subjectivity of holistic judgement leads to poor psychometric properties compared to methods based on objective scoring rubrics (Laming, 2004). However the CJ approach reported here yielded high internal consistency (Rasch Separation Reliability Coefficient = .88), high inter-rater reliability ($r = .87$) and high validity in terms of independent student achievement data ($r = .72$). We believe these acceptable psychometrics arising from subjective assessment decisions were due to the underlying principle that people are much more reliable at comparing one object against another than they are at rating an object isolation.

The strong association found between the CJ outcomes and teachers' assessment of students' general mathematical achievement suggests that the experts assessed *mathematics* as opposed to surface features such as neatness or length of response. However, can we be confident that the method measured *conceptual* rather than

procedural understanding of mathematics? For insight we turn to how accurately the students ordered the fractions in their responses as scored by the Levenshtein distance. The Levenshtein distance can be considered a measure of the procedural knowledge required to complete this task accurately. We found that teachers' assessments of students' mathematical achievement were a predictor of CJ assessment outcomes but the Levenshtein distances were not. Therefore it seems the CJ method produced measures more closely related to general mathematical ability than to specific performance on the fraction ordering test.

Nevertheless, to claim that the CJ process reported here measured conceptual understanding would be subject to two criticisms. First, the distinctiveness of conceptual and procedural knowledge, and the realism of operationalising them independently, has been questioned (e.g., Schneider & Stern, 2010; Star, 2005). For example, the procedures used by students to order the fractions are likely to be strongly influenced by how (and how well) the students conceived the underlying abstract principles of fractions. Second, the student achievement data used to establish the validity of the CJ method provide a general measure of mathematical achievement because they are based on the entire school mathematics curriculum, not just fractions. However teachers commonly use evidence to generate student achievement data that is subject to criticism by the mathematics education community and august bodies for being highly procedural, such as past papers from GCSE exams (e.g. Ofsted, 2008).

Therefore some caution must be exercised in claiming we have demonstrated the measurement of conceptual understanding of fractions. Further work is required to establish the extent to which CJ may offer a method that can be used routinely in mathematics education research and practice. A next step will be to validate CJ for the case of domains and concepts for which psychometrically validated instruments and methods for measuring conceptual understanding already exist.

References

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Abingdon: Routledge.

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

de Bock, D., Deprez, J., van Dooren, W., Roelens, M., & Verschaffel, L. (2011). Abstract or concrete examples in learning mathematics? A replication and elaboration of Kaminski, Sloutsky, and Heckler's study. *Journal for Research in Mathematics Education, 42*, 109–126.

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

Epstein, J. (2007). Development and validation of the Calculus Concept Inventory. In A. R. D.K. Pugalee & A. Schinck (Eds.), *Proceedings of the Ninth International Conference on Mathematics Education in a Global Community* (Vol. 9, pp. 165–170). Charlotte, NC.

Firth, D. (2005). Bradley–Terry models in R. *Journal of Statistical Software, 12*, 1–12.

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and Procedural Knowledge: The Case of Mathematics* (pp. 1–27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jones, I., & Alcock, L. (2012). Summative peer assessment of undergraduate calculus using Adaptive Comparative Judgement. In P. Iannone & A. Simpson (Eds.), *Mapping University Mathematics Assessment Practices*. Norwich: University of East Anglia.

Jones, I., Swan, M., & Pollitt, A. (submitted). Assessing mathematical problem solving using Comparative Judgement.

Kaminski, J. A., Sloutsky, V. M., & Heckler, A. F. (2008). The advantage of abstract examples in learning math. *Science, 320*, 454–455.

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

Knuth, E. J., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. *Journal for Research in Mathematics Education, 36*, 297–312.

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

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

Ofsted. (2008). *Mathematics: Understanding the Score*. London: Office for Standards in Education.

Piaget, J. (1952). *The Child's Conception of Number*. London: Routledge & Kegan Paul Ltd.

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

Posner, G. J., & Gertzog, W. A. (1982). The clinical interview and the measurement of conceptual change. *Science Education, 66*, 195–209.

Rittle-Johnson, B., Matthews, P., Taylor, R. S., & McEldoon, K. L. (2011). Assessing knowledge of mathematical equivalence: A construct-modeling approach. *Journal of Educational Psychology, 103*, 85–104.

Schneider, M., & Stern, E. (2010). The developmental relations between conceptual and procedural knowledge: A multimethod approach. *Developmental Psychology, 46*, 178–192.

Skemp, R. R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching, 77*, 88–95.

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