Measuring Conceptual Understanding Using Comparative Judgement

Marie-Josée Bisson, Camilla Gilmore, Matthew Inglis and Ian Jones

Mathematics Education Centre, Loughborough University

Correspondence concerning this article should be addressed to

Marie-Josée Bisson, Mathematics Education Centre, Loughborough University,

Loughborough, United Kingdom, LE11 3TU.

E-mail: M.Bisson@lboro.ac.uk; Phone: +44 (0) 1509 228247;

Fax: +44 (0)1509 228211

Abstract

The importance of improving students' understanding of core concepts in mathematics is well established. However, assessing the impact of different teaching interventions designed to improve students' conceptual understanding requires the validation of adequate measures. Here we propose a novel method of measuring conceptual understanding based on comparative judgement (CJ). Contrary to traditional instruments, the CJ approach allows test questions for any topic to be developed rapidly. In addition, CJ does not require a detailed rubric to represent conceptual understanding of a topic, as it is instead based on the collective knowledge of experts. In the current studies, we compared CJ to already established instruments to measure three topics in mathematics: understanding the use of p-values in statistics, understanding derivatives in calculus, and understanding the use of letters in algebra. The results showed that CJ was valid as compared to established instruments, and achieved high reliability. We conclude that CJ is a quick and efficient alternative method of measuring conceptual understanding in mathematics and could therefore be particularly useful in intervention studies.

Keywords: Conceptual understanding of mathematics; Comparative judgement; Measure; Validity; Reliability

Measuring Conceptual Understanding Using Comparative Judgement

Mathematics education researchers commonly distinguish between two broad types of mathematical knowledge (Hiebert & Lefevre, 1986; Skemp, 1976). One, commonly referred to as *procedural* knowledge, relates to fluency at applying algorithms step-wise to solve problems or transform notation (Byrnes & Wasik, 1991). The other, commonly referred to as *conceptual* knowledge, relates to understanding concepts, relationships and principles (Hiebert & Lefevre, 1986). Conceptual understanding is often defined as a network of relationships between pieces of information in a domain (Byrnes, 1992; Hiebert & Lefevre, 1986) as well as an understanding of the principles that govern a domain (Rittle-Johnson, Siegler & Alibali, 2001). However, there are a variety of definitions and assumed uses of conceptual understanding in the literature, and these have changed over time (Crooks & Alibali, 2014). For example, traditionally the *relationships* between pieces of knowledge were considered definitional but are increasingly seen as one feature of conceptual understanding, which grow and strengthen with increasing expertise (Baroody, Feil, & Johnson, 2007). Rittle-Johnson & Schneider (2014) argued that a less constrained definition of conceptual understanding as *knowledge of concepts* better reflects the contemporary research literature.

Improving students' understanding of concepts in mathematics, that is, improving students' conceptual understanding, has become of interest in recent years (e.g. NCTM, 2000; Ofsted, 2008). Increasingly, improved conceptual understanding is a key goal of educational interventions and policy changes. To achieve increased conceptual understanding in classrooms there need to be valid

and reliable measures of conceptual understanding. However, measuring knowledge of a given concept with acceptable validity and reliability is a major challenge for mathematics education researchers (e.g., Code, Piccolo, Kohler & MacLean, 2014; Crooks & Alibali, 2014). There have traditionally been two approaches to measurement. The first is to develop and psychometrically validate a bespoke instrument to measure knowledge of a particular concept (e.g. *The Calculus Concept Inventory - CCI,* Epstein, 2007, 2013; *Concepts in Secondary Mathematics and Science,* Brown, Hart, & Küchemann, 1984; Hart, Brown, Küchemann, Kerslake, Ruddock, & McCartney, 1981; Küchemann, 1978; *Reasoning about p–values and Statistical Significance scale - RPASS-7*; Lane-Getaz, 2013). However this has the disadvantage of being a long and resource-intensive process that must be repeated for every concept of interest. Moreover, on close analysis, measures do not always correspond to proposed definitions of the understanding of a given concept (Crooks & Alibali, 2014). Furthermore, confounds such as students' different reading levels can impact on outcomes and so threaten the validity of measures (Thurber, Shinn & Smolkowski, 2002), and we return to this issue later. The second approach to measuring conceptual understanding is to record one-to-one clinical interviews and develop a scoring rubric to rate the quality of each participant's understanding. However this has the disadvantage of requiring skill and consistency on the part of the interviewers and raters, and does not always lead to trustworthy results (Posner & Gertzog, 1982).

Here we report on a series of experiments using a novel method of measuring conceptual understanding based on comparative judgement (CJ). A

major advantage of this method is that it only requires one open-ended question about the topic of interest, and therefore only takes a few minutes to design.

Comparative Judgement

CJ is based on a long-standing psychological principle that humans are better at comparing two objects against one another than they are at comparing one object against specified criteria (Thurstone, 1994). When applied to educational assessment, CJ offers an alternative to traditional educational testing based on scoring rubrics (Pollitt, 2012). The method is simple. First researchers collect students' answers to a short open-ended question about the concept of interest. The student's responses are then presented in pairs to several experts, and for each pair of responses, a decision as to which one shows the better conceptual understanding of the topic is reached. This decision is generally reached quickly (a few minutes at the most) and once all the answers have been judged several times, a ranked order (from "worst" answer to "best" answer) is constructed using statistical modeling to calculate a standardized parameter estimate (z-score) representing the quality of each student's answer. CJ uses no detailed assessment criteria or scoring rubrics and the final rank order is instead grounded in the collective expertise of the judges. Previous work has shown that CJ performs validly and reliably in a variety of contexts, for example, to assess traditional mathematics examination scripts and mathematical problem solving tasks (Jones, Swan, & Pollitt, 2014), to conduct peer-assessments (Jones & Alcock, 2013) and to evaluate conceptual understanding of fractions (Jones, Inglis, Gilmore & Hodgen, 2013).

The theoretical motivation for using CJ is that conceptual understanding is an important but nebulous construct which experts can recognise examples of, but which is difficult to specify comprehensively and accurately in scoring rubrics. ("Experts" in this context refers to researchers in a discipline for which the concept of interest is important, such as mathematicians for the case of variable and derivative, and psychologists for the case of p values.) The shift from rubrics to a reliance on collective expertise for measuring understanding can be an uncomfortable notion, and is sometimes viewed by those used to traditional measurement methods as opaque and under-defined. However we argue that this is a key strength: a given concept is defined by how it is perceived, understood and used by the relevant community of expert practitioners. In contrast, rubrics attempt to capture the letter of a concept but risk losing the spirit. The perceived transparency and objectivity of rubrics can result a narrow and rigid definition that fails to capture the full meaning and usage that exists in practice. A related issue is that conceptual understanding is best assessed using open-ended and relatively unstructured tasks (e.g. explanation of concepts tasks, see Crooks & Alibali, 2014), which result in a wide variety of student responses that are difficult to anticipate in rubrics. CJ bypasses this shortcoming through relying on direct expert judgement of student responses that typically vary widely and unpredictably.

Another possible concern is that CJ is subjective because it relies on expert perception without reference to detailed, declarative criteria. This is an important issue and it is essential that CJ is conducted by many experts so that biases are cancelled out. For example, one mathematician might privilege responses that contain formal notation, and another privilege responses that

6

contain diagrams. We see such variety of expert perceptions as central to capturing the meaning and usage of a given concept. Whereas traditional instrument design attempts to capture variety through consultation and review, CJ attempts to capture it directly through collating experts' direct judgements of student work.

There is also a practical motivation for using CJ. Because the validity and reliability of the approach derives from the expertise of the judges, not from the psychometric properties of a particular instrument, CJ has the potential to be rapidly applied to any target concept with little effort on behalf of the researcher (beyond the recruitment of judges with sufficient expertise). In contrast, existing approaches to assessing conceptual understanding are resource intensive or very difficult, as discussed in the opening section. CJ therefore has the potential to effectively and efficiently evaluate a variety of educational interventions in a wide range of contexts.

The series of experiments reported in this article aimed to evaluate the suitability of CJ to assess conceptual understanding of different topics (statistics, calculus and early algebra) in mathematics. In addition, we investigated the suitability of this method for use with two different populations, namely undergraduate students and children. Finally, although CJ requires no scoring rubric, we investigated the usefulness of providing experts with guidance notes in Study 2.

In order to be useful to assess the impact of educational interventions, CJ must have acceptable validity and reliability. Validity can be assessed by establishing construct validity, i.e., evaluating whether CJ measures conceptual understanding of each mathematical topic. In our prior research, this was

achieved by comparing results on the CJ assessment to examination grades (Jones et al., 2014), module results (Jones & Alcock, 2013) or teachers' assessment of general mathematics achievement (Jones et al., 2013). This method is often called "criterion validity". In the current studies, in addition to using achievement results we assess validity of CJ by comparing it with existing validated instruments. In effect, our choice of topics for the current studies (statistics, calculus and algebra) was driven by the existence of validated instruments to measure conceptual understanding of those topics. Another condition of a useful measure is that it is reliable in the sense that a student's outcome is independent of whoever happened to assess their work. Reliability can be measured by recruiting raters to assess the same scripts independently and by comparing the outcomes (inter-rater reliability), and this is the approach we use throughout.

Study 1: Undergraduates' understanding of *p*-value

Method

Participants

Participants ($N = 20$) were all students at a university in England enrolled in an Applied Statistics undergraduate module, and they took part in the study during one of their one-hour weekly lectures. Participation in the study was voluntary, and students were given the option of not taking part at all by leaving the room before the start of the administration of the instruments, or of completing the instruments but having their data subsequently removed from the study. They were told that their answers would remain anonymous, but that their lecturer would have access to the anonymous scripts in order to address general

misconceptions in future lectures. In addition, general feedback about the answer scripts was provided to the lecturer.

Stimuli

Participants completed both an open-ended question (to be used for CJ) and a subset from the RPASS-7 scale (Lane-Getaz, 2013). The open-ended question was as follows:

> "Explain what a ***p*-value** is and how it is used to someone who hasn't encountered it before. You can use words, diagrams and examples to make sure you explain everything you know about *p*-values. Write between half a page and one page."

For the multiple-choice instrument, we selected 13 items from the RPASS-7 scale. All selected items had CITC (corrected item to total correlation, i.e, the correlation of the dichotomous item score with the total score of the scale minus the item's score) > .21. Items with cultural references that could not be reformulated for a UK context (item 6.6 about the ban of cameras in the US Supreme court, items 4a.1 and 4a.3 about the SAT prep course) were not selected. Items reported as problematic by Lane-Getaz (2013) were not selected (items 4b.5 and 6.3) despite having acceptable CITC. Therefore the final selection included items 1.3, 2.2, 2.4, 3a.1, 3a.2, 3b.1, 3b.2, 5.1, 5.2, 5.3, 6.1, 6.2, 6.4 from the original scale (see Lane-Getaz, 2013, p.44, for a description of conceptions and misconceptions assessed by each item). The items we used are presented in Appendix A.

Procedure

Participants were given 20 minutes to complete the open-ended question before working on the subset of items from the RPASS-7 for another 20 minutes. It was important that participants completed the open-ended question first, as the contents of the questions included in the RPASS-7 scale subset could have helped them draft their answer. Participants' written answers to the open-ended question were then scanned and uploaded onto the judging website (www.nomoremarking.com). Judges for comparative judgements were all PhD students in Psychology from two research-intensive universities in England. The 10 judges initially completed the subset of items from the RPASS-7 scale to ensure that they had adequate knowledge about p-values, and they had to reach a minimum of 9 correct answers on the scale to be included in the study ($M = 11$, $SD = 1.4$). All judges reached the eligibility criterion and were sent a link by email to access the judging website as well as a short instruction manual on how to use the website. They were instructed to complete one hour of judging, and they completed between 25 and 50 judgements each ($M = 45.3$, $SD = 8.7$), with each script being compared between 40 and 50 times ($M = 45.3$, $SD = 3.0$). The to-be-compared scripts were selected randomly for the first pair, and displayed on the screen simultaneously, side by side (see example Figure 1). For the subsequent comparisons, the script that appeared on the left hand side on the previous comparison now appeared on the right hand side, and a randomly selected script appeared on the left. Judges only had to click on the "left" or "right" button to select the script they considered to be the better answer.

Figure 1 about here

Results

CJ

We calculated a *z*-transformed parameter estimate for each participant's script

using the Bradley-Terry model (Firth, 2005), and ranked each scripts from worst

to best. In order to estimate the inter-reliability of the CJ method we used a split-

half technique. The 10 judges were randomly split into two groups of 5 and we

remodelled the parameter estimates for each group of judges before correlating

them. We repeated this process 20 times, and found that the inter-rater

reliability (Pearson's correlation coefficient) ranged from $r = .664$ and $r = .855$

($M = .745$, Median = .762, $SD = 0.06$). In addition, we calculated a second

measure of reliability, the Scale Separation Reliability (SSR, a measure of internal

consistency), and found that it was high, $SSR = .882$.

For the subset of items from the RPASS-7, percentages of correct answers

were calculated for each participant ($M = 64.6\%$, $SD = 18.2\%$, Cronbach's $\alpha =$

.539), and these were correlated with the parameter estimates from the Bradley-

Terry model to assess the validity of our CJ method. The results showed a

significant correlation between the two measures, $r = .457$, $p = .043$ (or $r = .721$

after correction for attenuation). Furthermore, both CJ parameter estimates and

RPASS-7 scores were significantly correlated with the students' Applied

Statistics module results, $r = .555$ $p = .021$ and $r = .553$ $p = .021$ respectively (see

Figure 2), and were not significantly different from each other, $t(19) = 0.01$, $p =$

.992.

Figure 2 about here

The results of Study 1 therefore showed that our CJ method of assessing conceptual understanding of p-values compared reasonably well with a validated instrument and yielded high inter-rater reliability and internal consistency.

Study 2: Undergraduates' understanding of derivative

The second study was conducted as a pilot for an intervention study investigating the effect of context on students' conceptual understanding. Therefore one of the aims was to compare the impact of providing a contextualised or decontextualized example prior to the open-ended question on the rating of responses. This aspect of the pilot study is not relevant to the CJ validation activities reported here and is therefore not included. There were a few other differences to the method of Study 2. One of the premises of CJ is that it requires no marking rubric of model answer because it is based on the collective knowledge of judges as to what represents conceptual understanding of a topic. We wanted to investigate the extent to which was the case by providing guidance notes to some of the judges and by evaluating the impact of this on the rating of responses during CJ. We therefore recruited three groups of judges for Study 2, one of which was provided with guidance notes (see Appendix B) about what would make a good answer to the open-ended question.

Method

Participants

Participants (N = 42) were all undergraduate students at a university in England enrolled in a Mathematical Methods in Chemical Engineering module, and they took part in the study during a weekly one-hour tutorial. Participation was

entirely voluntary. Students were given the option of not taking part at all by leaving the room before the start of the administration of the instruments, or of completing the instruments but having their data subsequently removed from the study. They were told that their answers would remain anonymous, but that their lecturer would have access to the anonymous scripts in order to address general misconceptions in future lectures and tutorials.

Stimuli

Participants were all given a booklet including either a contextualised or decontextualized example (shown in Appendix C) as well as the following open-ended question:

> "Explain what a **derivative** is to someone who hasn't encountered it before. Use diagrams, examples and writing to include everything you know about derivatives."

> Participants also completed a 10-item subset (items 2, 3, 5, 7, 8, 9, 11, 15, 19 and 21) related to the concept of derivatives from the CCI[1] (Epstein, 2007, 2013).

Procedure

Participants were given 20 minutes to complete the open-ended question, following which they worked on the questions from the CCI for a further 20 minutes. Similarly to Study 1, responses to the open-ended questions were then

---

[1] The CCI items are subject to a confidentiality agreement and as such we cannot reproduce them in this manuscript. Please contact the author of the CCI, Jerome Epstein, directly.

scanned and uploaded onto the judging website. Thirty mathematics PhD students were recruited. All judges were required to complete 42 judgements, with each script being judged between 14 and 28 times ($M = 21, SD = 1.94$) by each group of judges.

Results

CJ

Due to technical difficulties with the judging website, two participants had to be removed from the analyses as their response scripts were not judged.

To investigate reliability and validity of CJ, we calculated *z*-transformed parameter estimates for each participant's response script using the Bradley-Terry model and ranked these from worst to best. In order to assess the reliability of our CJ method, the 30 judges were randomly split into two groups of 15 and we remodelled the parameter estimates for each group of judges before correlating them. We repeated this process 20 times, and found that the inter-rater reliability (Pearson's correlation coefficient) ranged from $r = .826$ to $r = .907$ ($M = .869$, Median = .871, $SD = 0.02$). In addition, we found that the Scale Separation Reliability was high, $SSR = .938$.

CCI

Percentage accuracies on the CCI ($M = 48.8\%$, $SD = 18.7\%$) were calculated. As the internal consistency (Cronbach's alpha) was low, $\alpha = .397$, we searched for a subset of items with better internal consistency. We calculated α for every possible combination of 3 to 9 items, and found that the highest was $\alpha = .562$ for just three items (4,5,7), and the next best was $\alpha = .557$ for five items (4,5,6,7,10).

We repeated the following analysis with just these five items included but it made no difference to the overall findings. It seems that the subset of items we selected did not reach an acceptable level of internal consistency.

CCI vs. CJ

The correlation between the overall CJ parameters and CCI scores was low and not significant, $r = .093$, $p = .568$. Given the low Cronbach's alpha for the CCI test, and the acceptable inter-rater reliability of the CJ test, it seems the problem is that the CCI did not perform adequately. We obtained A-level mathematics grades for 33 of the participants to investigate the validity of both CJ and CCI. A-level grades and CJ parameters were moderately correlated, $r_s = .438$, $p = .011$ (Spearman's correlation coefficient) whereas these did not correlate significantly with the CCI result, $r_s = .088$, $p = .593$. Next we correlated CJ and CCI results with students' module results, and found that CJ parameters were moderately correlated with the module results, $r = .365$, $p = .021$, whereas the correlation with CCI accuracy scores did not reach significance, $r = .277$, $p = .083$ (see Figure 3).

Performance of the CCI

The poor performance of the CCI may have arisen due to the omission of items that were not designed to test understanding of derivative. Conversely it may be that the instrument, developed in the US, does not transfer to the UK context. To investigate this we administered the entire instrument comprising 22 items to a cohort studying Foundation Mathematics, a preliminary university module for

students who do not have the prerequisite mathematics for the undergraduate courses they wish to study.

The CCI was administered students after they had received two weeks of lectures and tutorials designed to introduce differentiation. The test was a compulsory revision exercise and students were required to opt-in or out of having their results used for research purposes. A total of 79 students agreed to their results being used. Although the cohort was different to that used in the main study the outcomes enabled us to investigate why the subset of items may not have performed as expected. As for the subset of items used in the main study, internal consistency across all 22 items was low, Cronbach's α = .326. We correlated students' total scores on the CCI with their overall scores on the module and found a moderate correlation, r = .437, p < .001. We also correlated CCI scores with an online test designed to assess procedural understanding of differentiation and found smaller correlation that was not significantly different to zero, r = .222, p = .066. The low internal consistency suggests that the poor performance of the subset of CCI items also applied to the full set of CCI items, although in this case the correlation with overall module scores was moderate. To explore the effects of using a subset of ten items further we considered the students' performance on the subset of items. For these items the internal consistency was low, Cronbach's α = .288. The correlation between students' scores on the subset of ten items on the CCI and their overall scores on the module was moderate, r = .457, p < .001, and there was no significant correlation between the subset of items and scores on the online differentiation test, r = .153, p = .207. The low internal consistency for both the subset and full set of CCI items suggests that in the context of Foundation Mathematics students in

England the instrument does not measure a single construct. Despite a moderate correlation with overall mathematics performance for the module there was little correlation with the results of the online differentiation test. The online test was designed to assess procedural knowledge whereas the CCI is intended to assess conceptual understanding, but nevertheless we would expect at least a moderate correlation between these related constructs.

Figure 3 about here

Judge group differences

For the analyses reported above, an overall parameter estimate for each script was calculated from the judgements of all judges together. However, as one of the aims of this study was to investigate the impact of providing judges with guidance, we divided the judges into three groups of ten and compared the parameters estimates obtained from each group of judges. Group 1 received guidance notes with their instructions (see Appendix B) prior to the judging to inform them as to what should be considered "a good answer", whereas Groups 2 and 3 did not. Results showed that the inter-rater correlation between the two groups (2 and 3) that received no guidance, r_{23} = .898, was numerically higher than between the group (1) that received guidance and the two groups that didn't, r_{12} = .849 and r_{13} = .803. But comparing the correlations suggests that the guidance notes did not make a substantial difference to the rating of responses. There was no difference in the correlation between r_{12} and r_{23}, t (37)= 1.20, p = .238, suggesting Group 2 (no guidance) was not different to Group 1 (guidance) and Group 3 (no guidance). In addition, the difference between r_{13} and r_{23} did not

reach the Bonferroni-corrected significance, t (37)= 2.40, p = .022 (Bonferroni corrected alpha .017), suggesting Group 3 (no guidance) was not different to Group 1 (guidance) as to Group 2 (no guidance). Finally, there were also no significant differences between the correlations of each group without guidance with the guidance group, r_{12} and r_{13}, t (37)= 1.18, p = .246.

Finally, we calculated each judge's misfit value (a measure of the quality or consistency of all the judgements produced by each judge; see Pollitt, 2012), by pooling once more all of the judgements together. These misfit values were then averaged for each group of judges and submitted to a between-subject one-way ANOVA which revealed no significant differences between the three groups of judges, F (2, 27) = 1.16, p = .328. This indicates that no group of judges stood out as judging differently from the other groups.

The results of Study 2 showed that inter-rater reliability was high for the CJ method and that CJ results correlated moderately with both A-level and Module results overall. However, CJ results did not correlate with the CCI, which may be due to the low internal consistency for the CCI. In addition, results revealed that providing guidance notes to judging expert to help them decide what constituted a good answer did not substantially alter the pattern of judging.

Study 3: Children's understanding of letters in algebra

The aim of Study 3 was to extend the results of Studies 1 and 2 by investigating the usability of CJ for teaching interventions with younger students. We wanted to assess the feasibility of using open-ended questions to assess conceptual understanding of an aspect of algebra and verify that even with younger students

this form of assessment is valid and reliable, and that it compared favourably

with an existing algebra assessment.

Participants

Participants were forty-six Year 7 students (aged 11 or 12 years old) from a local

middle school. The standards of achievements for English and mathematics at

this school were above those expected nationally at the latest school inspection

in 2012. The school was of average size (approximately 500 pupils on roll) and

the number of pupils eligible for free school meals was above the national

average.

Students took part in the study voluntarily during their regular

mathematics lesson. They were given the option of doing regular classwork

rather than taking part in the study. Information sheets with an opt-out option

were sent to parents prior to the study.

Stimuli

Participants completed the following open-ended question about algebra:

"Explain how letters are used in algebra to someone who has never seen

them before. Use examples and writing to help you give the best

explanation that you can."

They also completed a 15-item subset from the *Concepts in Secondary*

Mathematics and Science - Algebra scale (items 2, 3, 4.i, 4.ii, 4.iii, 5.i, 5.ii, 5.iii, 6.i,

11.i, 11.ii, 16, 18.ii, 20 from the original scale; see Brown, et al., 1984; Hart, et al.,

1981; Küchemann, 1978). This instrument has been used extensively during the

Concepts in Secondary Mathematics and Science (CSMS) project (Hart et al.,

1981) as well as during the Increasing Competence and Confidence in Algebra and Multiplicative Structures (ICCAMS) project (see Hodgen, Brown, Küchemann & Coe, 2010; Hodgen, Coe, Brown & Küchemann, 2014; Hodgen, Küchemann, Brown & Coe, 2009). The items we used are presented in Appendix D.

The role of letters in each item was categorised by Küchemann (1978), as shown in Table 1. An item is classified as *letter evaluated* (three items) if the letter's numerical value can be determined by trial and error without handling it as unknown. Items classified as *letter ignored* (five items) are those which can be solved numerically without need to "handle, transform or even remember the expression" (p. 25). *Letter as specific unknown* (three items) and *letter as generalised number* (two items) items are those containing a letter representing one value or a range of values respectively, and in both cases the value or values need not be evaluated in order to solve the item. Finally, items classified *letter as variable* (one item) imply a systematic relationship between two letters within an expression or two expressions containing the same letter.

Table 1 about here

Procedure

Students were given approximately 15 minutes to complete the open-ended question before working on the subset of the Algebra scale for a further 15 minutes. Responses were scanned and uploaded onto the judging website, and 10 mathematics PhD students were recruited as experts to judge which answer

showed the best conceptual understanding. Judges completed 46 judgements each and each script was judged between 19 and 21 times.

Results

CJ

As with Studies 1 and 2, we calculated *z*-transformed parameter estimates for each participant's response script using the Bradley-Terry model and ranked these from worst to best. Inter-rater reliability for the CJ instrument was assessed by randomly allocating the 10 judges to two groups of 5, and by calculating the correlations between the parameter estimates remodelled for each group. This was repeated 20 times, and results showed that the inter-rater reliability ranged from r = .678 to r = .837 (M = .745, Median = .742, SD = 0.04). In addition, Scale Separation Reliability was high, SSR = .843.

Percentages of correct answers on the subset of items from the Algebra instrument were calculated (M = 48.1%, SD = 19.4%, Cronbach's α = .770) and correlated with our *z*-transformed parameter estimates from the Bradley-Terry model. Results showed a moderate correlation, r = .428, p = .003. In addition, both the CJ parameter estimates and the accuracy scores on the Algebra subset significantly correlated with the students' current mathematics levels of achievement, r_s = .440, p = .002 and r_s = .555, p < .001 respectively (Spearman's Correlations; see Figure 4).

Figure 4 about here

In Sum, the results of Study 3 showed that our CJ measure correlated moderately with an existing instrument, and that the inter-rater reliability was high.

General Discussion

The aim of this manuscript was to add to the growing body of literature on comparative judgement by assessing the validity and reliability of this method compared to existing validated instruments designed to assess the conceptual understanding of various topics in mathematics. Three studies were carried out on the topics of statistics (p-values), calculus (derivatives), and algebra (the concept of letters as variables) and the results of these are summarised in Table 2.

Table 2 about here

In Studies 1 and 3, CJ scores correlated significantly with the existing instrument whereas this was not the case in Study 2. The lack of correlation between CJ and the existing instrument in Study 2 may have been due to the poor performance of the subset of items selected from the CCI, which was reflected in CCI's low internal consistency obtained for the instrument. Furthermore, the CCI results did not significantly correlate with achievement data. In contrast, CJ correlated significantly with the achievement data in all three studies thereby evidencing its validity. The correlations in all cases were moderate, however, this was not surprising as the achievement data was based

22

on results obtained throughout a mathematics course, which therefore included many mathematics topics and not just the topic evaluated by our CJ method.

The rank order obtained by three different groups of judges was investigated in Study 2 and showed that different groups of judges made similar judgements. We also found that providing guidance notes made very little difference. It would seem that simply using their expertise is sufficient for judges to arrive at reliable decisions. In addition, the high inter-rater reliabilities and internal consistencies computed in Studies 1 to 3 further highlighted the reliability of CJ.

Overall we found that it seems to be relatively quick and efficient to assess conceptual understanding using CJ. The number of decisions to use and how many judges to recruit are important issues left to the researcher. A general rule of thumb is to have at least 10 times the number of judgements to the number of scripts. This is the number we have used in the studies reported in this manuscript, and this was also the case in previous studies (Jones & Alcock, 2012; Jones et al., 2013; Jones et al., 2014). Overall this led to good validity and reliability. We also found that it is preferable to have at least 10 judges per study, in order to access the "collective expertise" of judges.

Limitations

In this section we consider possible limitations to using CJ to measure conceptual understanding. We start by considering threats to the validity of what is measured before considering some practical considerations.

One possible threat to validity is judge bias. Earlier we discussed the need for a group of experts to undertake the judging to ensure that the outcome

reflects *collective* expertise. However, it is still possible that systematic bias common to all or most of the judges will affect outcomes. For example, non-mathematical features, such as the quantity written, neatness of presentation and literacy levels, might positively prejudice judges when making decisions. We explored the possible impact of quantity written on judges' decisions by taking file size of the scanned responses as a proxy for quantity written and calculating the Spearman correlation between the rank order given by parameter estimates and that given by file size. There were no significant correlations between parameter estimate and file size for understanding of p-value, $r_s = .281$, $p = .230$, derivative $r_s = .134$, $p = .409$, or letters in algebra, $r_s = .154$, $p = .306$, suggesting that the quantity of written text did not have a substantial impact on judges' decision making across the three studies. Further work is required to identify further possible sources of bias. For example, we are currently investigating how learners' literacy levels might impact on judges' decisions.

We have made the argument that CJ offers a cheaper and more efficient approach than traditional test development, which can take years to design, test and refine. However there are contexts in which CJ may be less efficient. For example, if many studies are to be conducted at large scale that focus on a specific concept, such as the fairly widespread use of the CCI to assess undergraduate understanding of calculus (Epstein, 2013), then the costs of judging may become prohibitive. Conversely, once an instrument has been designed and validated, the costs of administering and scoring it are relatively cheap, especially for multiple-choice or other objective formats that can be scored by computer. However, such scenarios are relatively rare and as noted

24

instrument development still introduces a delay to being able to evaluate learner understanding.

There also remain situations where a concept of interest might be better measured using a traditional, instrument-based approach. One example for the case of younger learners is understanding of mathematical equivalence for which a psychometrically robust instrument has been developed (Rittle-Johnson, Matthews, Taylor & McEldoon, 2011). The instrument is underpinned by decades of rigorous research into children's developing understanding of equivalence and, while the instrument has been critiqued for a disconnect between theory and measurement (Crooks & Alibali, 2014), it is unlikely to be improved upon using the CJ approach described here. Similarly, research into understanding misconceptions and how they might be triggered and overcome, might not be best served using the approach described here. For example, the phenomenon of "natural number bias" (Van Hoof, Lijnen, Verschaffel & Van Dooren, 2013) has been researched in detail using specialised tasks that can be adapted to investigate specific hypotheses (e.g. Durkin & Rittle-Johnson, 2015). Traditional instruments lend themselves well to the fine-grained mapping of misconceptions, whereas CJ is better suited to testing the relative effectiveness of interventions for improving understanding of a given concept more broadly. There may also be situations in which CJ can be used to complement other methods. For example, we might investigate the role a known misconception plays on general understanding of a concept by administering specific tasks to probe the misconception and a suitable CJ test to measure general understanding.

Conclusion

We believe the CJ method of measuring conceptual understanding has important advantages compared to creating and validating an instrument. We hope that the series of studies reported here contributed to demonstrating the usefulness of CJ as well as its validity and reliability. Two of the studies reported here (Studies 1 and 2) involved undergraduate students, who would have been accustomed to answering open-ended questions and explaining their reasoning. Conversely, Study 3 was conducted with Year 7 students where this type of question, especially in mathematics classrooms, is not the norm. It is remarkable therefore that CJ performed as well as it did even with this younger cohort. Future studies will investigate using CJ with primary school aged students, as well as investigating the sensitivity of CJ to measure the impact of different teaching interventions.

References

Baroody, A. J., Feil, Y., & Johnson, A. R. (2007). An alternative reconceptualization of procedural and conceptual knowledge. *Journal for Research in Mathematics Education, 38*, 115-131.

Brown, M., Hart, K., & Küchemann, D. (1984). Chelsea Diagnostic Mathematics Tests (pp. 1–8). Retrieved from http://iccams-maths.org/CSMS/

Byrnes, J. P. (1992). The conceptual basis of procedural learning. *Cognitive Development*, *7*, 235–257.

Byrnes, J. P., & Wasik, B. A. (1991). Role of conceptual knowledge in mathematical procedural learning. *Developmental Psychology, 27*, 777-786.

Code, W., Piccolo, C., Kohler, D., & MacLean, M. (2014). Teaching methods comparison in a large calculus class. *ZDM*, *46*, 589–601.

Crooks, N. M., & Alibali, M. W. (2014). Defining and measuring conceptual knowledge in mathematics. *Developmental Review, 34*, 344-377.

Durkin, K., & Rittle-Johnson, B. (2015). Diagnosing misconceptions: Revealing changing decimal fraction knowledge. *Learning and Instruction, 37*, 21-29.

Epstein, J. (2007). Development and Validation of the Calculus Concept Inventory. In *Proceedings of the Ninth International Conference on Mathematics Education in a Global Community* (pp. 165–170).

Epstein, J. (2013). The Calculus Concept Inventory - measurement of the effect of teaching methodology in mathematics. *Notices of the American Mathematical Society, 60*, 1018-1027.

Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical software*, *12*(1), 1-12. Retrieved from http://www.jstatssoft.org/

Hart, K., Brown, M. L., Küchemann, D. E., Kerslake, D., Ruddock, G., & McCartney, M. (Eds.). (1981). *Children's understanding of mathematics: 11-16*. London: John Murray.

Hiebert, J. & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hibert (Ed). *Conceptual and Procedural Knowledge: The Case of Mathematics* (pp. 113-133). New York and London: Routledge.

Hodgen, J., Brown, M., Küchemann, D., & Coe, R. (2010). Mathematical attainment of English secondary school students: a 30-year comparison. *Increasing attitudes and attainment in secondary school mathematics: Evidence from a large-scale study in England.* In D. Durant (chair) Symposium conducted at the meeting of British Educational Research Association, Warwick, UK.

Hodgen, J., Coe, R., Brown, M., & Küchemann, D. (2014). Improving students' understanding of algebra and multiplicative reasoning: Did the ICCAMS intervention work? In S. Pope (Ed.), *Proceedings of the 8th British Congress of Mathematics Education 2014* (pp. 1–8).

Hodgen, J., Küchemann, D., Brown, M., & Coe, R. (2009). Children's understandings of algebra 30 years on. *Research in Mathematics Education, 11*, 193–194.

Jones, I., & Alcock, L. (2013). 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.

Küchemann, D. (1978). Children's understanding of numerical variables. *Mathematics in School, 7*, 23–26.

Lane-Getaz, S. J. (2013). Development of a reliable measure of students' inferential reasoning ability. *Statistics Education Research Journal, 12*(1), 20–47. Retrieved from http://iase-web.org/Publications.php?p=SERJ

NCTM. (2000). *Principles and Standards for School Mathematics.* Reston, VA: National Council of Teachers of Mathematics.

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

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.

Rittle-Johnson, B., & Schneider, M. (2014). Developing conceptual and procedural knowledge of mathematics. In R. Cohen Kadosh & A. Dowker (Eds.), *Oxford handbook of numerical cognition.* Oxford: Oxford University Press.

Rittle-Johnson, B., Siegler, R., & Alibali, W. A. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*, 346–362.

Skemp, R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching, 77*, 20-26.

Thurber, R. S., Shinn, M. R., & Smolkowski, K. (2002). What is measured in mathematics tests? Construct validity of curriculum-based mathematics measures. *School Psychology Review*, *31*, 498.

Thurstone, L. L. (1994). A law of comparative judgment. *Psychological Review, 101*, 266–270.

Van Hoof, J., Lijnen, T., Verschaffel, L., & Van Dooren, W. (2013). Are secondary school students still hampered by the natural number bias? A reaction time study on fraction comparison tasks. *Research in Mathematics Education, 15*, 154-164.

Table 1: Classification and levels of subset of items from the Algebra scale described by Küchemann (1978).

Item	Letter…	Item	Letter…
1	ignored	4c	specific unknown
2	as variable	5	evaluated
3a	ignored	6a	evaluated
3b	ignored	6b	evaluated
3c	specific unknown	7	generalised number
4a	ignored	8	generalised number
4b	ignored	9	specific unknown

Table 2: Summary of validity (correlation coefficient with validated instrument and/or achievement data) and reliability measures (inter-rater reliability and Scale Separation reliability or Cronbach's alpha) for all three studies for the CJ method as well as the validated instruments.

		Study 1 (*p*-values)	Study 2 (derivatives)	Study 3 (algebra)
CJ				
Validity	Instr.	.457	.093	.428
	Achiev.	.555	.365	.349
Reliability	Inter-rater	.749	.869	.745
	SSR	.882	.938	.843
Instrument				
Validity	Achiev.	.553	.277	.448
Reliability	Cronbach α	.539	.397	.770

Note. CJ = Comparative judgement; Instr. = Correlation with existing instrument (RPASS-7, CCI and Algebra for Studies 1 to 3 respectively); Achiev. = Correlation with achievement data; SSR = Scale Separation Reliability.

Figure captions

Fig. 1: Example display screen for comparative judgement website.

Fig. 2: Correlations between CJ parameter estimates, the RPASS-7 subset scores and students' module results.

Fig. 3: Correlations between CJ parameter estimates, CCI subset results and students' A-levels and module results.

Fig. 4: Correlations between CJ parameter estimates, results for the subset of items from the Algebra instrument from the Concepts in Secondary Mathematics and Science project, and students' mathematics achievement levels.

Appendix A

Subset of items from the Reasoning about p-value and Statistical Significance scale (RPASS-7)

Scenario 1:

A research article reports that the mean number of minutes students at a particular university study each week is approximately 1000 minutes. The student council claims that students are spending much more time studying than this article reported. To test their claim, data from a random sample of 81 students is analysed using a one-tailed test. The analysis produces a p-value of .048.

Question 1.1 Assume a student had conducted a two-tailed test instead of a one-tailed test on the same data, how would the p-value (.048) have changed?

  a. The two-tailed p-value would be smaller (i.e., the p-value would be .024).
  b. The two-tailed p-value be the same as the one-tailed (i.e., the p-value would be .048).
  c. The two-tailed p-value would be larger than the one-tailed (i.e., the p-value would be .096

Scenario 2:

The district administrators of an experimental program are interested in knowing if the program had improved the reading readiness of first graders. Historically, before implementing the new program, the mean score for Reading Readiness for all first graders was 100. A large random sample of current first graders who attended the new preschool program had a mean Reading Readiness score of 102. Assess the following actions and interpretations of district researchers.

Question 2.1 Interpretation: In their presentation to the district administration, the researchers explained that when comparing the observed results to the general population, the stronger the evidence that the reading readiness program had an effect, the smaller the p-value that would be obtained.

  a. Valid interpretation
  b. Invalid interpretation

Question 2.2 Interpretation: After checking the conditions necessary for inference, the district researchers found they had statistically significant results. They interpreted the small p-value to mean that the cause of the results obtained was clearly due to chance.

  a. Valid interpretation
  b. Invalid interpretation

Scenario 3:

A researcher conducts a two-sample test. He compares the mean hair growth results for one group of students who agreed to try his treatment to a second group's mean who do not use the treatment. He hopes to show that there is a statistically significant difference between the two group means. How should this researcher interpret results from this two-sample test?

Question 3.1 Interpretation: If the group that had the treatment has more hair growth (on average) compared to the no treatment group and the *p*-value is small, the researcher interprets the *p*-value to mean there would definitely be more hair growth in a population who uses his treatment.
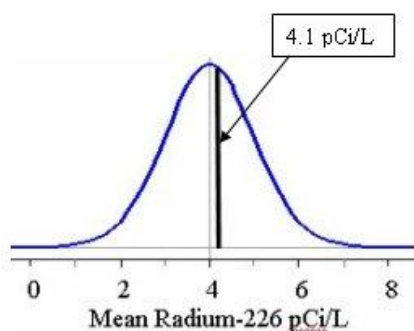
   a.   Valid interpretation
   b.   Invalid interpretation

Question 3.2 Interpretation: Assume the conditions for inference were met. The researcher interprets the *p*-value as an indicator of how rare (or unusual) it would be to obtain the observed results or something more extreme, if the hair treatment had no effect.

   a.   Valid interpretation
   b.   Invalid interpretation

Scenario 4:

Radium-226 is a naturally occurring radioactive gas. For public safety, the Environmental Protection Agency (EPA) has set the maximum exposure level of Radium- 226 at a mean of 4 pCi/L (picocuries per litre). Student researchers at a southern Florida university expected to show that Radium-226 levels were less than 4 pCi/L. However, these student researchers collected 32 soil specimens with a mean Radium-226 measured at 4.1 pCi/L. Students checked the necessary conditions and conducted a hypothesis test at the .05 level. Estimate the *p*-value given the sketch below of the distribution of means and the observed mean of 4.1 pCi/L.

33

Question 4.1 Interpretation: Based on the estimated *p*-value, the students' sample mean was statistically significant.

   a.   Valid interpretation
   b.   Invalid interpretation

Question 4.2 Interpretation: The estimated *p*-value for the students' sample is greater than .05.

   a.   Valid interpretation
   b.   Invalid interpretation

Scenario 5:

Suppose you have a new driving school curriculum which you suspect may alter performance on passing the written exam portion of the driver's test. You compare the mean scores of subjects who were randomly assigned to control or treatment groups (20 subjects in each group). The treatment group used the new curriculum. The control group did not. You use a 2-sample test of significance and obtain a *p*-value of 0.01.

Question 5.1 Statement: The small *p*-value of .01 is the probability that the null hypothesis (that there is no difference between the two population means) is false.

   a.   True Statement
   b.   False Statement

Question 5.2 Statement: The probability that the experimental (i.e., the alternative) hypothesis is true is .01.

   a.   True Statement
   b.   False Statement

Question 5.3 Statement: Assume you had obtained an even smaller *p*-value (than .01). A smaller *p*- value...

   a. is stronger evidence of a difference or effect of the new driving school curriculum.
   b. is weaker evidence of a difference or effect of the new driving school curriculum.
   c. suggests no change in the difference or effect of the new driving school curriculum.

In this section, there are no scenarios. Just chose the best answer for each question.

Question 6.1 A research article gives a *p*-value of .001 in the analysis section. Which definition of a *p*-value is the most appropriate? The *p*-value is...

a.   the value that an observed outcome must reach in order to be considered significant under the null hypothesis.
b.   the probability that the null hypothesis is true.
c.   the probability of observing an outcome as extreme or more extreme than the one observed if the null hypothesis is true.
d.   the probability that the observed outcome will occur again.

Question 6.2 If a researcher was hoping to show that the results of an experiment were statistically significant they would prefer:

a.   a large *p*-value
b.   *p*-values are not related to statistical significance
c.   a small *p*-value

Question 6.3 A researcher conducts an experiment on human memory and recruits 15 people to participate in her study. She performs the experiment and analyzes the results. She obtains a *p*-value of .17. Which of the following is a reasonable interpretation of her results?

a. This proves that her experimental treatment has no effect on memory.
b. There is evidence of a small effect on memory by her experimental treatment.
c. There could be a treatment effect, but the sample size was too small to detect it.
d. She should reject the null hypothesis.

Appendix B
Guidance notes for judges

Guidance to assessors

Question

Explain what a **derivative** is to someone who hasn't encountered it before. Use diagrams, examples and writing to include everything you know about derivatives.

Guidance on a good answer

A "good answer" is a self-contained complete story. It is very unlikely that a stream of consciousness will result in a coherent story. Some rough working will be necessary to order the ideas. But, under exam/test conditions (such as this) it may be difficult to plan or revise work.

You should expect to see the formal definitions of

- derivative at a point $x = a$;
- derived function $f'(x)$.

These make use of limits. There are a number of related *concepts.*

- The idea of a tangent line and the *gradient of the tangent line.* The *tangent line* to a curve at a point (x, y) on that curve is the straight line through (x, y) which gives the *best local approximation* to the curve.
- Instantaneous rates of change, including velocity and acceleration.

Appropriate diagrams could be used to relate the formal definition to the concept of tangent line.
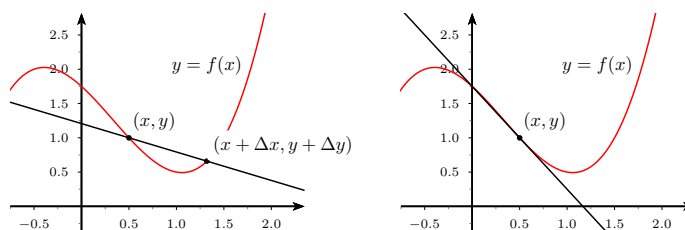
The solution should have a uniform level of detail. I.e. spell out the tricky bits, but omit details of very simple calculations.
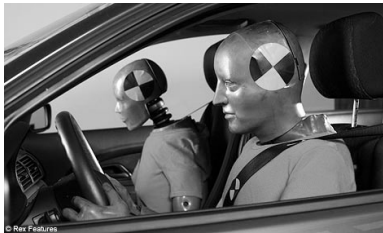
It is very helpful to have some examples which should be simple but also generic enough to capture most (ideally all) of the important concepts, and processes. Not all functions have a derivative, an example such as $|x|$ might help to illustrate this.

A good answer will both distinguish and relate the formal definition to the actual practical process of finding the derivative, which are the familiar techniques of differential calculus.

The story should be *complete.* A complete piece of mathematics contains a mixture of formal algebraic calculation and logical reasoning. Remember algebra is primarily abbreviation, and so should form part of a sentence. However, the mathematics is more important than handwriting, spelling or grammar: concentrate most on the *mathematics.*
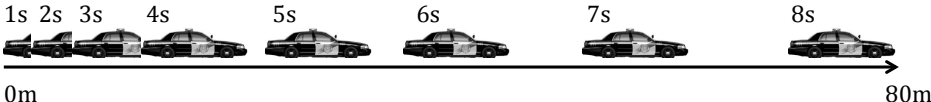
Appendix C
The (i) contextualised and (ii) decontextualised examples provided to participants who completed the open-ended question on derivatives.

Think about when you have travelled in a car.
How does it feel when the car is moving?
How does it feel when the car speeds up?
How does it feel if the car brakes sharply?

In this lesson we will think about describing the movement of a car using mathematics.

Imagine a parked police car suddenly zooming off. After the first second, it has hardly moved any distance at all. It gets faster and faster, and after 8s, it has moved 80m, quite a long way. How might the change of speed feel to the police officers in the car?

1s 2s 3s   4s   5s   6s   7s   8s

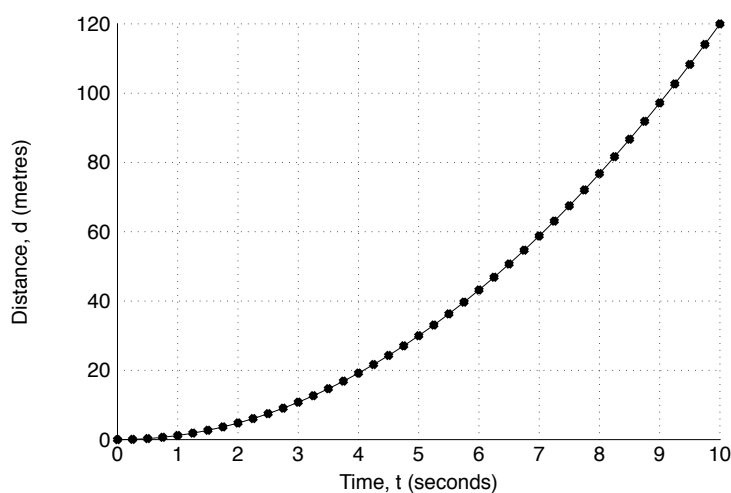0m   80m

Let's think about an accelerating car mathematically.

The graph below shows the distance travelled, d metres, plotted against time, t seconds, for a car accelerating from rest.

The table shows the distance every 1 second for the interval $0s \leq t \leq 10s$.

However, the graph has been drawn by plotting many more points than are shown in the table.

Time, t seconds	Distance, d metres
0	0.0
1	1.2
2	4.7
3	10.6
4	18.9
5	29.5
6	42.5
7	57.8
8	75.5
9	95.6
10	118.0

Appendix C continued

The graph below shows a function in which y is plotted against x.

The table shows the value of y for every increase of 1 in x for the interval $0 \le x \le 10$.

However, the graph has been drawn considering many more points than are shown in the table.

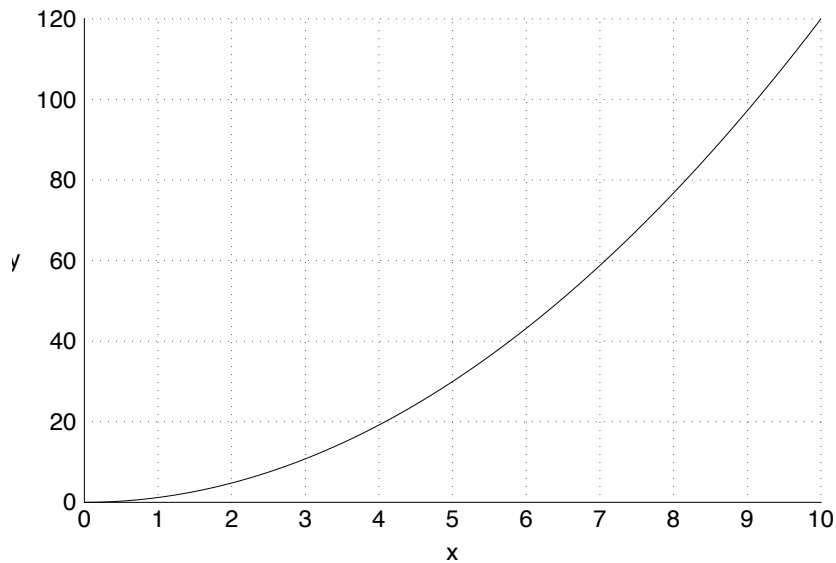| x	y
0	0.0
1	1.2
2	4.7
3	10.6
4	18.9
5	29.5
6	42.5
7	57.8
8	75.5
9	95.6
10	118.0

Appendix D
Subset of items from the Algebra scale of the "Concepts in Secondary Mathematics and Science" project

---

1. Write down the smallest and the largest of these: smallest   largest

    $n + 1$, $n + 4$, $n - 3$, n, $n - 7$ ……………   ……………

---

2. Which is larger, $2n$ or $n + 2$?   ……………

    Explain: ………..

---

3. **4 added to n** can be written as **$n + 4$.**
   Add 4 onto each of these:
   $$8 \qquad\qquad n + 5 \qquad\qquad 3n$$
   ……………   ……………   ……………

---

4. If $a + b$ =   43   If $n - 246 = 762$ If $e + f$ = 8

    $a + b + 2$ = …………… $n - 247$ = …………… $e + f + g$ = ……………

---

5. What can you say about a if $a + 5 = 8$ …………………………………………………

---

6. What can you say about u if u = $v + 3$
                        and v = 1   …………………………………………………

   What can you say about m if m = $3n + 1$
                        and n = 4   …………………………………………………

---

7. What can you say about c if $c + d$ = 10
                        and c is less than d …………………………………………………

---

8. When is the following true - always, never, or sometimes?
                        underline the correct answer:

    $L + M + N = L + P + N$ Always.   Never.   Sometimes, when ……………………………….

---

9. Cakes cost c pence each and buns cost b pence each.

   If I buy 4 cakes and 3 buns,
   what does
                        $4c + 3b$ stand for ?   …………………………….