Measuring Conceptual Understanding: The Case of Teaching with Abstract and Contextualised Representations

Ian Jones, Matthew Inglis, Camilla Gilmore and Marie-Josée Bisson
Mathematics Education Centre, Loughborough University

Executive Summary.

Introduction.

The difficulty of measuring conceptual understanding presents a barrier to progress in the development and practice of high-quality mathematics education interventions. Conceptual understanding is commonly defined as deep knowledge of the underlying concepts of mathematics and how they relate to one another (Crooks & Alibali, 2014). Innovative methods for teaching mathematics are commonly claimed to impact positively on students' conceptual understanding; yet if conceptual understanding cannot be measured efficiently and reliably then robust evidence cannot be established. A recent and high-profile example of this problem is the debate over whether it is better to teach mathematical topics using abstract or contextualised representations. Some scholars have concluded that abstract representations are preferable (e.g. Kaminski et al., 2008) while others have come to more equivocal conclusions (e.g. Brown, McNeil & Glenberg, 2009). Key to these disparate conclusions is the lack of agreed and trustworthy measures of conceptual understanding (De Bock et al., 2011). As such, the current trend towards grounding mathematics curricula in real-world scenarios (ACME, 2012; MEI, 2012; Gowers, 2012; Truss, 2012) lacks an evidence base.

In the research reported here we developed a measure of conceptual understanding using a Comparative Judgement (Pollitt, 2012) approach, and demonstrated its application to the abstract vs. contextualised debate. Comparative Judgement (CJ) is a way to assess open-ended and creative mathematical work. It involves no mark schemes and no marking because such traditional methods cannot reliably be applied to assessing open-ended work (Laming, 1990). Instead two pieces of student work are presented on a screen and the assessor is asked to decide which is "better". The decision may be based on a specific objective, such as "the better understanding of fractions", or may be general, such as "the better mathematician". This is a binary decision. There is no need to decide how much better one piece of work is than the other. When many such pairings are shown to many assessors the decision data can be statistically modelled to generate a score for each student. The statistical modelling also produces quality control measures, such as checking the consistency of the assessors. Previous research has shown the comparative judgement approach produces reliable and valid outcomes for assessing the open-ended mathematical work of secondary and undergraduate students (Jones & Alcock, 2014; Jones, Inglis, Gilmore & Hodgen, 2013).

Objectives.
There were two objectives to the research reported here.

1. To apply CJ to measuring the learning outcomes of randomised controlled trials in which students are taught key concepts.
2. To provide valid and reliable evidence on the relative benefits of abstract and contextualised representations for introducing key concepts to students.

To achieve these objectives we undertook five studies. The first three studies investigated the feasibility of using CJ to measure understanding of key concepts across a range of contexts. The final two studies applied CJ to determining whether abstract or contextualised representations are superior for introducing two key concepts to students.

Studies 1a, 1b and 1c. Measuring understanding of key concepts.
Secondary school and university students completed open-ended tests on three concepts: the role of letters in simple algebra; derivatives in calculus; and *p*-values in statistics. These concepts were chosen because, unusually, validated measures have been developed in these areas and so provided a yardstick for evaluating the CJ approach. We found that student scores based on expert pairwise judgements of the open-ended tests correlated with the traditional test scores and with students' general mathematics achievement. This suggests that the CJ approach enabled the valid assessment of students' understanding of the three concepts.

Study 2. Abstract vs. contextualised representations: The case of algebra.
We first investigated whether CJ could be used to detect group differences in a randomised controlled trial. The focus was on the abstract vs. contextualised debate for the case of introducing letters in algebra to primary school students. A total of 189 students were randomly assigned to two groups and each received a series of three specially designed algebra lessons. One group was taught algebra using the MiGen software (Noss et al., 2012), which offers a broadly contextualised approach to learning mathematics; the other group was taught using the Grid Algebra software (Hewitt, 2014), which offers an abstract approach. Following the intervention, the students' understanding of the role of letters in algebra was tested using an open-ended test, which was then assessed by experts using CJ, and a traditional test. We found that the Grid Algebra group outperformed the MiGen group on both measures, although the difference between groups was larger for the open-ended test. In conclusion then, for the case of introducing algebra to primary children, the abstract approach, as exemplified by Grid Algebra, produced measurably greater learning gains. Moreover, the open-ended CJ-based test was slightly more sensitive than the traditional test at detecting this difference.

Study 3. Abstract vs. contextualised representations: The case of calculus.
We then investigated whether CJ could be used to detect group differences under more tightly controlled conditions. The focus was again on the abstract vs. contextualised debate, this time for the case of introducing differential calculus to high-achieving secondary students. 189 students were randomly assigned to

two groups and each received a series of three calculus lessons. Unlike for Study 2, the lessons were identical except that the materials drew on real-world examples (e.g. accelerating vehicles) for one group, and used only abstract representations (mathematical symbols and graphs) for the other group. Following the intervention, open-ended CJ-based post-tests and traditional post-tests were administered to measure the students' understanding of the concept of derivative. We found no difference in overall performance between the two groups on either of the measures. Thus, for the case of introducing calculus to high-achieving secondary students, neither abstract nor contextualised representations produced measurably greater learning gains.

Findings.
There are two main findings from the research. First, CJ can be used to evaluate students' conceptual understanding, and to evaluate the relative effectiveness of different teaching approaches. As such, a significant barrier to progress in the field of mathematics education can now be overcome; namely, the paucity of effective measures of students' conceptual understanding in different domains. Our contribution will enable researchers to evaluate and understand the effectiveness of various educational resources and approaches more quickly and validly than has been possible to date. This in turn will provide policy-makers and teachers with better evidence about the relative effectiveness of educational interventions.

Second, we have informed the abstract vs. contextualised representations debate by providing evidence on relative effectiveness in two contexts. For the case of algebra we compared two technology-specific approaches to teaching using abstract and contextualised representations. We found that an abstract approach using the Grid Algebra software was more effective for learning about letters in algebra than a contextualised approach using the MiGen software. For the case of differential calculus we compared two specially-designed sets of teaching resources. We found that an abstract approach, using formal representations such as symbols and graphs, and a contextualised approach, using real-world representations such as accelerating cars, were equally effective for learning about the concept of derivative. We conclude that the role of abstraction and contextualisation when teaching mathematics is nuanced, and effectiveness depends on the concept being taught, the approach used, and perhaps the age and prior achievement of learners. Importantly, the CJ approach enabled us to overcome the measurement problem that has limited the findings of previous research.

Background.
A common distinction is made in mathematics education between procedural and conceptual knowledge (Hiebert & Lefevre, 1986; Skemp, 1976). Procedural knowledge involves memorising facts and applying algorithms, whereas conceptual knowledge involves understanding mathematical concepts and the relationships between them (Star, 2005). Procedural knowledge is relatively straightforward to measure using familiar test questions, but measuring conceptual understanding is more difficult and time-consuming, and the outcomes are not always trustworthy (Crooks & Alibali, 2014).

The difficulty of measuring conceptual understanding presents a barrier to progress in the development and practice of high-quality mathematics education interventions. Innovative methods for teaching mathematics are commonly claimed to impact positively on students' conceptual understanding; yet if conceptual understanding cannot be measured efficiently and reliably then robust evidence cannot be established. A recent and high-profile example of this problem is the debate over whether it is better to teach mathematical topics using abstract or contextualised representations. Some scholars have concluded that abstract representations are preferable (e.g. Kaminski et al., 2008) while others have come to more equivocal conclusions (e.g. Brown, McNeil & Glenberg, 2009). Key to these disparate conclusions is the lack of agreed and trustworthy measures of conceptual understanding (De Bock et al., 2011). As such, the current trend towards grounding mathematics curricula in real-world scenarios (ACME, 2012; MEI, 2012; Gowers, 2012; Truss, 2012) lacks an evidence base.

In the research reported here we adapted and deployed a novel measure of conceptual understanding based on the Comparative Judgement (CJ) method described in the following section.[1] To investigate the method's validity and cost-effectiveness we conducted five studies, including two randomised controlled trials that compared teaching approaches using abstract and contextualised representations.

Comparative Judgement (CJ).
CJ is based on a long-standing psychological principle that people are better at comparing two objects against one another than they are at comparing one object against specified criteria (Thurstone, 1927). When applied to educational assessment, CJ offers an alternative to traditional educational testing based on scoring rubrics (Pollitt, 2012). The basics of CJ are straightforward. Experts are presented with pairs of student work and asked to decide which of the two has demonstrated the better understanding of a given concept (an example is shown in Figure 1). The outcomes from many such pairings presented to several experts are then statistically modelled to produce a score of the 'quality' of each piece of work.

[1] We used the online CJ engine www.nomoremarking.com, which is free to use for educational and research purposes.

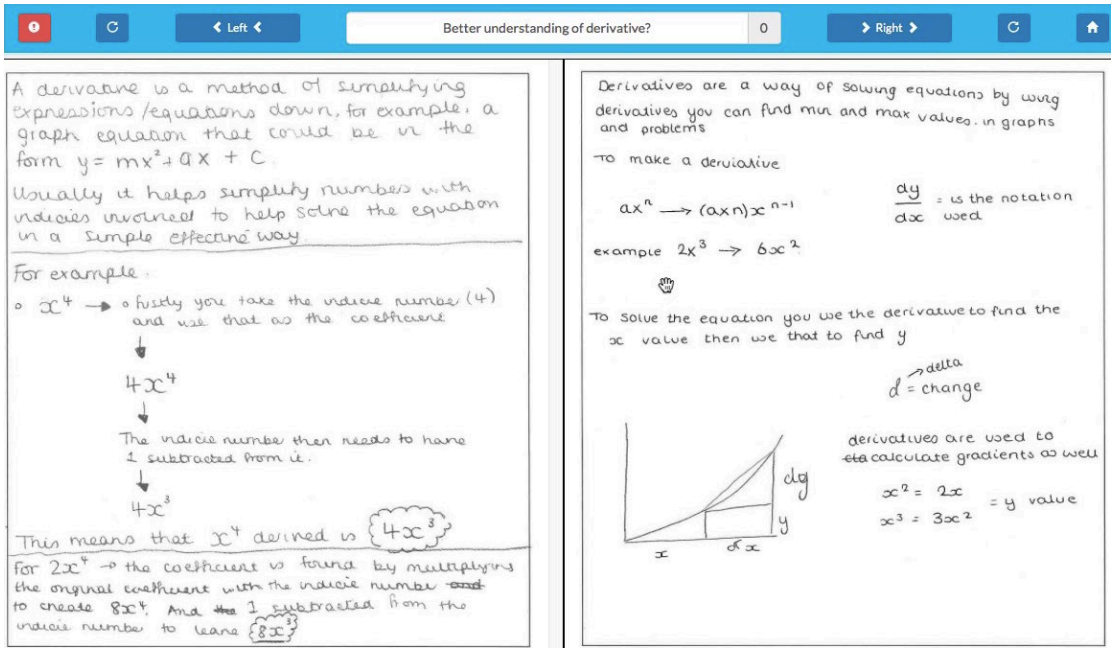


Figure 1: Example of two pieces of student work presented via a web browser.

Previous research has demonstrated the usefulness of CJ for assessing procedural knowledge (Jones, Swan & Pollitt, 2014), problem solving (Jones & Inglis, 2015), conceptual understanding for the case of fractions (Jones et al., 2013), and multivariate calculus (Jones & Alcock, 2014). The research reported here built on this work by applying CJ to determine the effectiveness of different teaching interventions for improving students' understanding of the concepts of letters in algebra and derivatives in calculus.

Objectives and Key Outcomes.

There were two main objectives to the research. The first was to make a methodological contribution that has the potential to transform how conceptual understanding in mathematics is operationalised and assessed. The second was to provide robust evidence to inform the abstract vs. contextualised debate, and so impact on how mathematics lessons are designed and taught.

The first objective was met in two ways. First, a series of three studies were conducted to establish that CJ can be applied to the measurement of conceptual understanding of letters in algebra, derivatives in calculus, and *p*-values in statistics. The studies were successful, demonstrating that CJ measured these concepts validly and reliably, outperforming traditional approaches to assessing conceptual understanding (Bisson, Gilmore, Inglis & Jones, 2016). Second, two experiments, one conducted with primary students and another with secondary students, demonstrated that CJ offers a reliable and suitably sensitive method for comparing group outcomes in randomised controlled trials.

The second objective has been met through two teaching interventions. In one study, we introduced algebra to medium- to high-achieving 9 and 10 year olds using two existing software packages. The software packages, Grid Algebra (Hewitt, 2014) and MiGen (Noss et al., 2012), were both designed to assist

students with the transition from arithmetic to algebra, but are grounded in different underlying philosophies that align to abstract and contextualised conditions respectively. We found that the children taught algebra using Grid Algebra outperformed those taught using MiGen on a post-test that was assessed using our CJ method, and on a traditional assessment. In the other study, we introduced calculus to high-achieving 15 and 16 year olds in two carefully controlled conditions. The lesson materials were identical except that the contextualised condition made plentiful use of real-world examples, such as accelerating vehicles, whereas the abstract condition used only mathematical symbols and graphs. We found no difference in outcomes across the abstract and contextualised conditions for the case of calculus.

The Research.

The research involved five studies: a series of three initial studies to validate and refine the CJ method for assessing conceptual understanding; a study that introduced algebra to primary children using one of two educational software packages; and a study that introduced calculus to secondary students using abstract or contextualised lesson materials.

Studies 1a, 1b and 1c.

There were three components to Study 1, corresponding to the mathematical concepts of (a) the role of letters in simple algebra, (b) derivatives in calculus, and (c) *p*-values in statistics (Bisson et al., 2016). The participants were 46 students at the start of secondary schooling (letters in algebra), 42 undergraduates enrolled on a Mathematical Methods in Chemical Engineering module (derivatives), and 20 undergraduates enrolled on an Applied Statistics module (*p*-values).

Each component followed a similar procedure: students were administered a specially-designed open-ended test, and a traditional test from the research literature. The traditional tests were as follows. For letters in algebra we used 15 items from the Concepts in Secondary Mathematics and Science – Algebra Scale (Brown et al., 1984); for derivatives we used 10 items from the Calculus Concept Inventory (Epstein, 2007); for *p*-values we used 13 items from the Reasoning about *p*-values and Statistical Significance Scale (Lane-Getaz, 2013). The open-ended and traditional tests targeted the same mathematical concept in each study. General mathematics achievement data were also collected for each student. Experts were recruited from two universities, using our contacts from previous projects, to comparatively judge the open-ended test (30 mathematics PhD students were recruited for the calculus tests, 10 mathematics PhD students for the algebra tests, and 10 psychology PhD students for the statistics tests). Once judging was complete, the inter-rater reliability and internal consistency (namely, the scale separation reliability, a coefficient similar to Cronbach's α) of the CJ outcomes were estimated. The CJ scores were then correlated with the traditional test scores and students' general mathematical achievement in order to evaluate the validity of the approach. The results, which are summarised in Table 1, provided empirical support that the CJ outcomes were reliable and valid

for all three mathematical concepts and all three groups of students.[2] In other words, our novel approach successfully tapped understanding of each concept, and the outcomes were stable across independent groups of expert judges.

	Letters in algebra	Derivatives	*p*-values
Reliability			
Inter-rater reliability	.745	.869	.749
Internal consistency	.843	.938	.882
Validity			
Correlation with traditional test	.428	.093	.457
Correlation with achievement measure	.440	.365	.555

Table 1: Summary of reliability and validity for all three concepts explored in Studies 1a, 1b and 1c.

For the calculus tests, the judgement decisions from all 30 judges were included in the analysis. However, initially these judges were randomly allocated into one of three groups. Group 1 received guidance on what makes a good answer (see Appendix) whereas Groups 2 and 3 received no guidance. This enabled us to investigate whether providing guidance impacts on the quality of judgements made. We found that the guidance made no difference; that is, Group 1's judgements agreed with those of Groups 2 and 3. The correlations of CJ-based test scores between the three groups of judges were r_{12} = .85, r_{13} = .80 and r_{23} = .90, and these were not significantly different to one another. We conclude from this that providing judges with guidance does not impact on outcomes, which is perhaps unsurprising given that PhD students' knowledge of mathematics can be expected to be far in advance of first year engineering undergraduates. This adds to the evidence for the validity of the CJ approach.

Study 2.

In Study 2, a teaching intervention was conducted to establish whether the CJ approach to measuring conceptual understanding could be used to detect group differences in a randomised controlled trial. We focused on teaching a concept that was unlikely to have been encountered by the students beforehand to minimise the effect of prior knowledge on learning outcomes. The concept was the role of letters in algebra,[3] which we taught to primary students two years before they would have normally encountered it at the start of secondary school.

Lesson design. Two comparable sets of three lessons were specially designed for the study. In the lessons, primary students were taught linear equations containing up to two letters (e.g. $5 = 3 + x$, $y = 4x$) using one of two software packages. The Grid Algebra package builds on children's knowledge of written

[2] An exception to this was the traditional test for derivatives, which failed to perform reliably and validly in the context it was applied here. This explains the poor correlation with the CJ scores.

[3] We deliberately avoided the term "variable" because representing a varying quantity is just one of the roles letters play in school algebra (Küchemann, 1978).

arithmetic to introduce algebra through symbols and expressions, and formed the basis of the abstract condition. The MiGen package introduces algebra as a notational system for recording and describing repeating geometric patterns, and formed the basis of the contextualised condition. An example screenshot from the two software packages can be seen in Figure 2. All lessons were co-designed and delivered by the same, highly-experienced and respected teacher (Jan Parry). The lessons were videotaped and observed by a researcher to ensure consistency across schools and conditions (abstract/contextualised).

Participants and measures. A total of 257 Year 5 students (ages 8 and 9) were recruited and randomly allocated to two groups. A battery of tests was administered to the students prior to the teaching intervention. The tests measured numeracy using questions 8 to 44 of the Numerical Operation subtests from the Welchsler Individual Achievement Test (WIAT-II UK: Welchsler, 2005); mathematics anxiety using questions 2, 3, 5 and 7 from the Child Maths Anxiety Questionnaire (Ramirez et al., 2013); writing skills using the Written Expression subtest of the WIAT-II UK; and non-verbal reasoning performance using Sets A, B and C from the Raven's Educational UK Edition Standard Progressive Matrices Plus Version (Raven, 2008). No differences were found in any of these measures across the two groups of students. Following the three lessons, students were administered an open-ended test containing the question:

> *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 explanations that you can.*

This was followed by a single side of blank paper for the students to write their answer. An example student response is shown in Figure 3. The open-ended test responses were comparatively judged by ten mathematics PhD students (recruited from Studies 1a and 1b). Their judgement decisions were statistically modelled to produce a score representing the 'quality' of each response. The outcomes were found to be reliable (internal consistency, $\alpha = 0.86$; inter-rater reliability, $r = 0.70$). A traditional algebra post-test was also administered consisting of 12 items from the Concepts in Secondary Mathematics and Science – Algebra Scale (Brown et al., 1984), and its internal consistency was found to be acceptable (Crobach's $\alpha = 0.64$).

Analysis. We analysed which teaching condition (abstract/contextualised) led to greater learning of the concepts as measured by both the open-ended and traditional post-tests. The analysis was designed to take account of student differences (numeracy and writing levels, school attended and so on) in order to identify how much variance in learning gains could be attributed to the teaching experiments rather than extraneous variables.

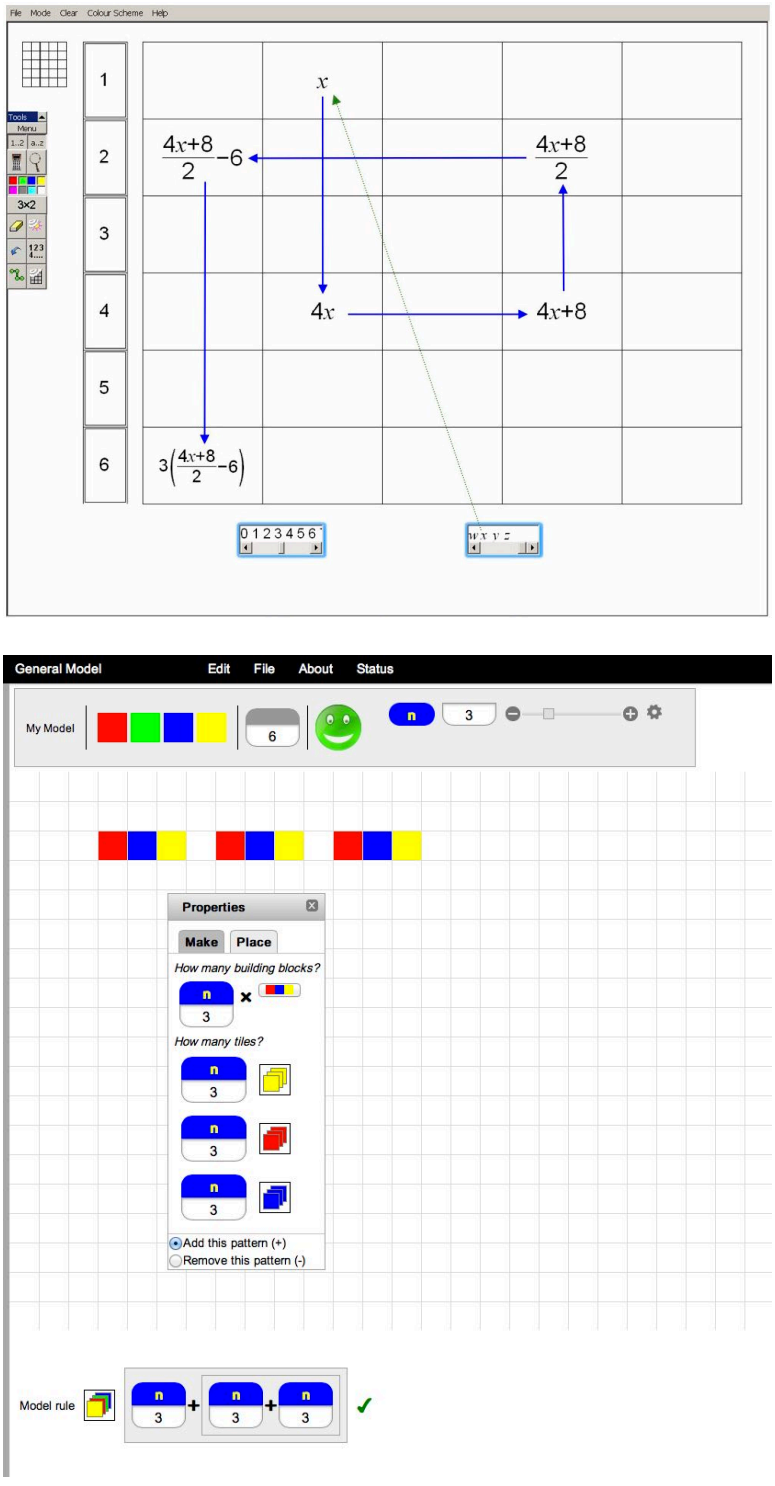


Figure 2: Screenshots from the Grid Algebra (top) and MiGen (bottom) software packages used in the algebra teaching experiments.

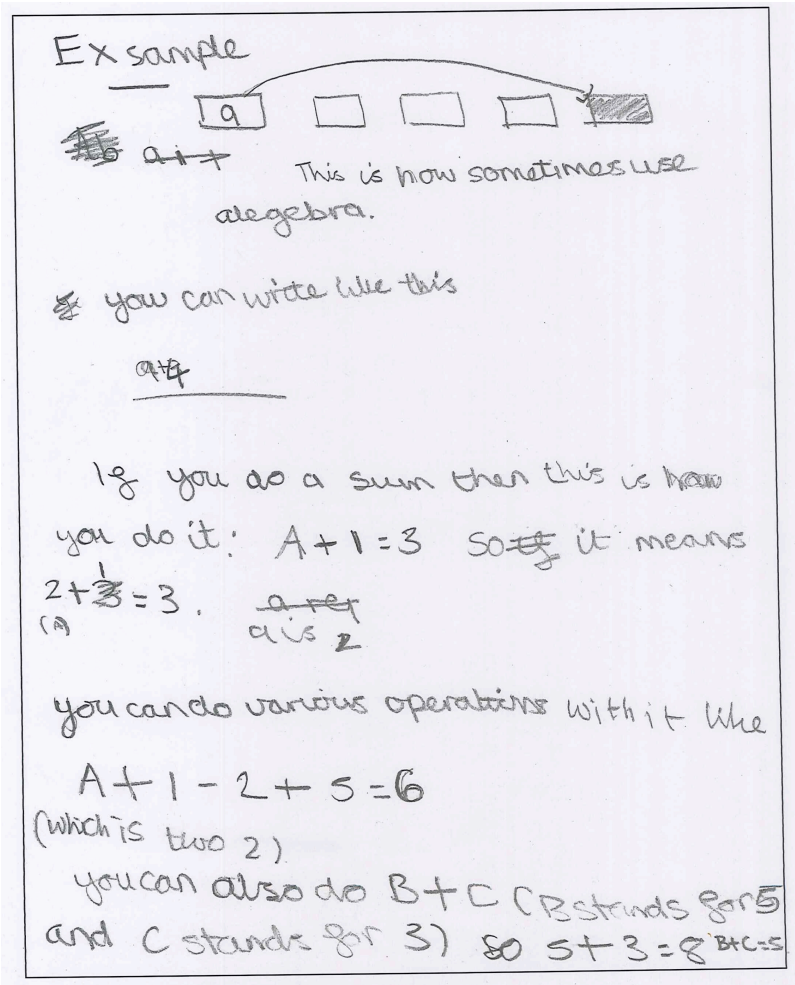


Figure 3: Example student response to the open-ended post-test (algebra).

Out of the 257 participants recruited, 39 were excluded because they were part of an initial trial run after which the lessons were amended, and a further 29 were excluded due to absence for at least one session. This left 189 participants who were included in the analysis: 98 in the Grid Algebra group and 91 in the MiGen group. A mean score for the open-ended test and for the traditional test was calculated for each group. The open-ended test mean score for the Grid Algebra group (M_{CJ_GA} = 0.26) was higher than that for the MiGen group (M_{CJ_MG} = −0.27); similarly, the traditional test mean score for the Grid Algebra group (M_{T_GA} = 5.05) was higher than that for the MiGen group (M_{T_MG} = 4.45). To investigate these differences further we constructed a multilevel model that took account of covariates (numeracy, mathematics anxiety, writing, non-verbal reasoning) and school attended. This revealed that students taught using Grid Algebra (abstract) learned more than students taught using MiGen (contextualised) according to both the open-ended test (Cohen's d = 0.40) and the traditional test (Cohen's d = 0.23). The CJ approach resulted in a larger effect size between groups than the traditional test, and only the CJ approach reached statistical significance. This suggests the CJ approach was slightly more sensitive than the traditional test at detecting the difference in understanding of the concept of letters in algebra across the two teaching conditions.

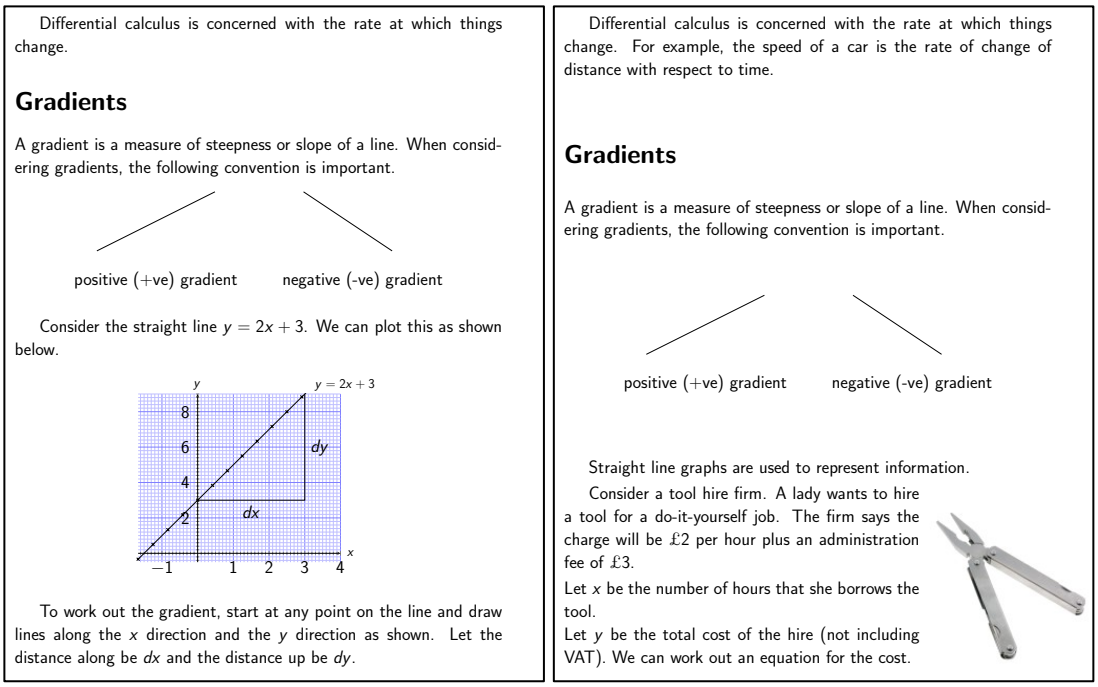


Figure 4: Examples from the lesson materials for the calculus lessons in abstract (left) and contextualised (right) conditions.

Students' performance on the numerical operations pre-test was significantly correlated with their performance on the subsequent open-ended and traditional tests. Importantly, performance on the writing pre-test was also significantly correlated with their performance on the open-ended test. This is unsurprising given that a good response to the open-ended test requires communicating understanding. This could be seen as a weakness of the CJ method: it clearly assesses writing skills as well as the mathematical concept of interest. However, by controlling for writing ability in our statistical model, we were able to show that while students with better writing ability performed better on the open-ended test, they were still judged to have a greater understanding of letters in algebra if they were in the Grid Algebra group. This result suggests that researchers who use CJ methods to assess conceptual understanding in RCTs should consider controlling for writing skills, especially when studying younger children, in order to improve the sensitivity of their measure.

Study 3.

Study 3 was similar to Study 2 but with a more advanced concept taught to older students. The concept was derivatives in differential calculus, which we taught to high-achieving Year 11 students (ages 15 and 16) who had not yet encountered calculus.

Lesson design. Whereas for Study 2 we selected two existing software packages that embody different approaches to teaching mathematics, for Study 3 we designed two sets of lesson materials that were identical except for the use abstract and contextualised examples. This enabled a precise exploration of the relative benefits of abstraction and contextualisation for learning advanced mathematics.

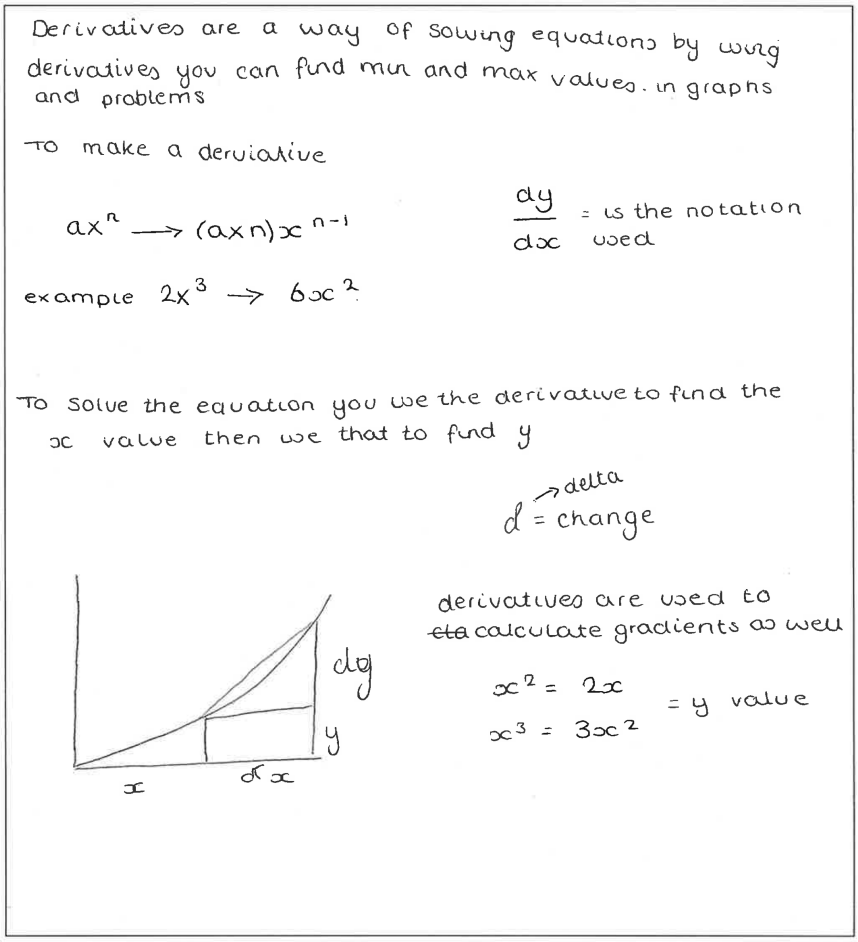


Figure 5: Example student response to the open-ended post-test (calculus).

In the lessons, secondary students were taught differential calculus of simple functions (e.g. $y = x^2$). In the abstract condition the materials exclusively used formal mathematical representations (symbols, graphs), and in the contextualised condition the materials made extensive use of real-world scenarios (e.g. accelerating vehicles). An example of the materials for each condition can be seen in Figure 4. As for Study 2, all lessons were co-designed and delivered by the same, highly-experienced and respected teacher (Rod Bond), and the lessons were videotaped to ensure consistency across sessions and conditions.

Participants and measures. A total of 260 Year 11 students were recruited and randomly allocated to two groups. A battery of tests was administered to the students prior to the lessons using the intervention materials. The tests measured numeracy using questions 20 to 54 of the Numerical Operations subtests from the WIAT-II UK; mathematics anxiety using questions 2, 5, 8, 9, 10, 13, 14, 19, 21 and 22 from the Mathematics Anxiety Scale UK (Hunt, Clark-Carter & Sheffield, 2011); writing skills using the Written Expression subtest of the WIAT-II UK; and non-verbal reasoning performance using Sets C, D and E from the Raven's Educational UK Edition Standard Progressive Matrices Plus Version (Raven, 2008). No differences were found across the two groups of students on

any of the tests. Following the three lessons, students were administered an open-ended test containing the 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.*

This was followed by a single side of blank paper for the students to write their answer. An example student response is shown in Figure 5. The traditional post-test was 10 items from the Calculus Concept Inventory (CCI, Epstein, 2007). As in Study 1b the internal consistency of the CCI was found to be poor (Cronbach's α = 0.11), casting doubt on its validity as a measure of understanding of derivative in the context of this study. We therefore do not report on the CCI results further.

Analysis. Some participants were excluded from the analysis due to attending different intervention lessons to the ones randomly assigned to them (N = 39), absence from lessons (N = 14), non-completion of pre-lesson activities (N = 9), prior knowledge of calculus (N = 8), or opting-out of the study (N = 8). This left 189 participants who were included in the analysis: 97 in the abstract condition and 92 in the contextualised condition. The open-ended test responses were comparatively judged by ten mathematics PhD students (recruited from Studies 1a and 1b). Their judgement decisions were statistically modelled to produce a score representing the 'quality' of each response. The outcomes were found to be reliable (internal consistency, α = 0.82; inter-rater reliability, r = 0.67). The open-ended test mean score for the abstract condition (M_{CJ_A} = 0.07) was higher than that for the contextualised condition (M_{CJ_C} = -0.10). To investigate whether this difference was statistically significant we conducted an analysis that took account of pre-test results (numeracy, mathematics anxiety, writing, non-verbal reasoning).[4] This revealed that there was no difference in the learning gains between students in the abstract and contextualised conditions (p = 0.31).

As in Study 2, students' performance on the numerical operations pre-test related significantly to their performance on the subsequent open-ended test. Unlike for the algebra experiment, no other variable, including performance on the writing pre-test, was related to their performance on the open-ended test. This contrast with Study 2 may have arisen due to the difference in ages of the students across the studies, or because the older students were also all high achievers: older and higher-achieving students could perhaps be expected to have a greater and more uniform mastery of written communication.

Key findings.

There are two key findings from the research, one methodological and one theoretical.

The methodological finding is that the CJ approach we applied to measuring conceptual understanding in mathematics appeared to be successful. The

[4] Preliminary analyses showed no variation in performance by school, therefore the analysis was more straightforward than for the algebra experiment.

findings from Studies 1a, 1b and 1c demonstrated its validity for measuring understanding of concepts in algebra, calculus and statistics across different age ranges and educational contexts. The findings from Studies 2 and 3 demonstrated the method's applicability to testing for group differences in randomised controlled trials. In the algebra experiment we found that the open-ended test was slightly more sensitive to group differences than was the traditional test. In the calculus experiment we found that the open-ended test performed validly and reliably (although no group differences were detected), whereas the traditional test failed as a valid measure of conceptual understanding because of its poor internal reliability.

Moreover, the open-ended tests required only a few minutes to design, whereas the traditional tests took months (Küchemann, 1980) or years (Epstein, 2007) to develop, trial and refine. The CJ approach described here therefore provides an efficient and robust method to measure any concept of interest and has great potential for use in a range of educational applications, including the evaluation of teaching interventions and randomised controlled trials. This could help to enable the research community to provide timely and robust evidence on the effectiveness of educational approaches, and therefore move the field of mathematics education forward.

The theoretical finding is that the benefits of being taught mathematical concepts using abstract or contextualised resources is more nuanced than some researchers (e.g. Kaminski et al., 2006) and policy initiatives (e.g. Truss, 2012) suggest. For the case of teaching algebra to middle- and high-achieving primary students, an abstract approach, as embodied by the Grid Algebra software, was more effective for learning than a contextualised approach. However, this may have been due to other factors that varied between the software packages; for example, MiGen requires students to learn idiosyncratic notation to use the software, whereas Grid Algebra requires students to use complicated formal notation. For the case of teaching differential calculus to high-achieving secondary students, we designed contextualised and abstract materials that were extremely similar, except with regards to the nature of the representations used. In this case neither approach resulted in greater learning gains over the other.

Therefore it seems that the benefits of abstraction and contextualisation interact with other variables such as age, prior achievement of the learners, pedagogic approach, and the concept being taught. Nevertheless, in contrast to previous studies (e.g. De Bock et al., 2011; Kaminski et al., 2006), our approach avoided using tests that were procedural, or that closely resembled the teaching materials used, to make claims about conceptual understanding. CJ enabled us to overcome the specific limitation that has hampered recent research and fuelled controversy regarding the relative efficacy of using 'pure' and real-world examples in mathematics education.

Implications.
Based on the research reported here we offer the following recommendations related to the use of CJ for measuring conceptual understanding, and the role of contextualisation when introducing new concepts to students.

Comparative Judgement: Recommendations.
The CJ method offers a valid, reliable and efficient technique for measuring students' understanding of a target concept. We have demonstrated its application to three mathematical concepts in this research: letters in algebra, derivatives, *p*-value. Understanding of any given target concept can, in principle, be similarly measured by developing an appropriate open-ended test question. A particular and potentially important role for CJ that we have explored is evaluating the outcomes of randomised controlled trials in terms of gains in conceptual understanding.

Perhaps the single most important design issue is writing an appropriate test question. The question must explicitly target the concept of interest, and also provoke a wide-enough variety of student responses that judges are able to make meaningful pairwise decisions. Researchers should also consider controlling for students' writing skills to increase the sensitivity of CJ-based measures. This is particularly the case for younger students whose writing skills are less advanced. Judges need to be experts in their field, clear about the judging task, and undertake the judging task sincerely. However, beyond this requirement, there appears to be no benefit to attempting to impose a consensus as to what kinds of student response should be preferred over others. So long as the judges are experts they can be assumed to know conceptual understanding when they see it.

Abstract and contextualised representations: Recommendations.
Across both Study 2 and Study 3 we found that, on average, students in the abstract condition scored more highly than those in the contextualised condition. However, this difference was only statistically significant for the case of algebra in which abstract and contextualised representations were exemplified by two very different types of technology-based interventions. Moreover, even for the algebra intervention the overall group difference was small. There was no significant group difference for the case of differential calculus in which abstract and contextualised representations were more tightly controlled. Therefore researchers, teachers and policy-makers should be sceptical of the wide-spread belief that new mathematical topics are best introduced to students using applied or 'realistic' contexts. Conversely, the relative benefit of using abstract representations may be very small at best, and not there at all in some contexts. Therefore the mathematics education community should be similarly sceptical of strong claims that abstract representations are always best when introducing new topics to students.

Acknowledgement.
The Nuffield Foundation is an endowed charitable trust that aims to improve social well- being in the widest sense. It funds research and innovation in

education and social policy and also works to build capacity in education, science and social science research. The Nuffield Foundation has funded this project, but the views expressed are those of the authors and not necessarily those of the Foundation. More information is available at www.nuffieldfoundation.org

References.

ACME. (2012). *Post-16 Mathematics*: *A Strategy for Improving Provision and Participation*. London: Advisory Committee on Mathematics Education.

Bisson, M.-J., Gilmore, C., Inglis, M., & Jones, I. (2016). Measuring conceptual understanding using comparative judgement. *International Journal of Research in Undergraduate Mathematics Education*. Online first DOI: 10.1007/s40753-016-0024-3

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

Brown, M. C., McNeil, N. M., & Glenberg, A. M. (2009). Using Concreteness in Education: Real Problems, Potential Solutions. *Child Development Perspectives, 3*, 160–164.

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

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.

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.

Gowers, T. (2012). *How Should Mathematics be Taught to Non-Mathematicians?* Blog post, https://gowers.wordpress.com/2012/06/08/how-should-mathematics-be-taught-to-non-mathematicians/

Hewitt, D. (2014). A symbolic dance: the interplay between movement, notation, and mathematics on a journey toward solving equations. *Mathematical Thinking and Learning, 16*, 1–31.

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.

Hunt, T. E., Clark-Carter, D., & Sheffield, D. (2011). The development and part validation of a U.K. scale for mathematics anxiety. *Journal of Psychoeducational Assessment, 29*, 455–466.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*, 1774–1787.

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: can comparative judgement help? *Educational Studies in Mathematics, 89*, 337–355.

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, Germany: IGPME.

Jones, I., Swan, M., & Pollitt, A. (2014). Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education, 13*, 151–177.

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

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

Küchemann, D. (1980). *The Understanding of Generalized Arithmetic (Algebra) by Secondary School Students* (unpublished doctoral dissertation). Chelsea College, London.

Laming, D. (1990). The reliability of a certain university examination compared with the precision of absolute judgements. *The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 42*, 239–254.

Lane-Getaz, S. J. (2013). Development of a reliable measure of students' inferential reasoning ability. *Statistics Education Research Journal, 12*, 20–47

MEI. (2012). *Integrating Mathematical Problem Solving: Applying Mathematics and Statistics Across the Curriculum at Level 3. End of Project Report*. London: Mathematics in Education and Industry.

Noss, R., Poulovassilis, A., Geraniou, E., Gutierrez-Santos, S., Hoyles, C., Kahn, K., … Mavrikis, M. (2012). The design of a system to support exploratory learning of algebraic generalisation. *Computers & Education, 59*, 63–81.

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

Ramirez, G., Gunderson, E. A., Levine, S. C., & Beilock, S. L. (2013). Math Anxiety, Working Memory, and Math Achievement in Early Elementary School. *Journal of Cognition and Development, 14*, 187–202.

Raven, J. (2008). *Standard Progressive Matrices Plus Version*. Pearson Education: London, UK.

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

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education, 36*, 404–411.

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

Truss, E. (2012). *Elizabeth Truss Calls for a Renaissance in Maths*. Norfolk: Speech to the National Education Trust. Retrieved from https://www.gov.uk/government/speeches/elizabeth-truss-calls-for-a-renaissance-in-maths

Wechsler, D. (2005). *Welchsler Individual Achievement Test Second UK Edition (WIAT-II)*. Pearson Assessment: London, UK.

Appendix

Guidance provided to Group 1 of the experts who judged the open-ended calculus tests.

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.

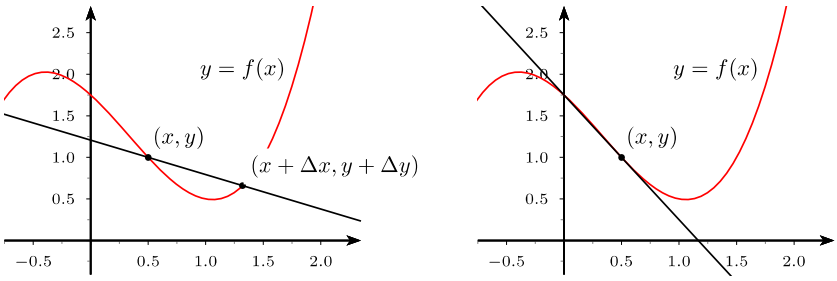
You should expect to see the formal definitions of

- 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*.