

## Automatic assessment of steps in students' work

E R Cerval-Peña cervale@for.mat.bham.ac.uk

D F M Hermans D.F.M.Hermans@bham.ac.uk

C J Sangwin sangwinc@for.mat.bham.ac.uk

School of Mathematics, University of Birmingham, UK

Assessment is a key process in the learning cycle. Indeed, assessment drives students' learning, from both formative assessments and low-stakes class work, through to high-stakes public examinations. Learning technology can support students' learning, e.g. enabling visualization of graphs, through to providing direct answers to mechanical computations. Technology can also be used to support the assessment process itself, and computer-aided assessment (CAA) of mathematics has a long history. This is a large field which includes systems for collating and organizing students' work to help a human undertake marking in an efficient and ethical way. Our interest is confined to *automatic marking* and in particular we are interested in automatically establishing the mathematical properties of students' answers. Driven by the difficulty of automatically assessing students' open-ended mathematical answers, many CAA implementations make use only of numerical input or very simple question types such as multiple choice questions (MCQs), or modifications of such "teacher-provided response" types. However, since 2000 a number of systems have used a computer algebra system (CAS) as a library of functions with which to support this process and such CAA appears to be a technical improvement over the MCQ format for reasons such as the following:

1. The CAS supports structured randomly-generated questions with corresponding worked solutions. Students can, and do, ask for repeated practice of similar examples. Different questions may also deter plagiarism or reduce impersonation.
2. Mathematical properties of students' answers can be established. This is significant: the teacher moves away from deciding whether the "answer looks correct" to articulating precisely which properties are relevant and the acceptability of an answer satisfying only a subset of them. Such judgements form the keystone of all assessment, whether formative, summative or as part of a research instrument. See [3].
3. Feedback based on these properties and which incorporates mathematical manipulations of answers can be provided almost immediately. A concrete example of a STACK question is shown in Figure 1. This has two linked parts, each of which requires the student to enter a mathematical expression using the typed linear syntax of the CAS Maxima. The first part is "valid" but "wrong" and feedback is provided, the second part has been rejected as "invalid" and feedback also provided. Repeated attempts are encouraged. Students may then act on this feedback rather than waiting for the delay inevitable in a traditional paper and pencil environment.

1. Find the derivative of  $x^3$  with respect to  $x$ .

Your last answer was interpreted as:

$$\frac{x^4}{4}$$



**Incorrect answer.**

It looks like you have integrated instead!

Your mark for this attempt is 0. ❌ With penalties, and previous attempts, this gives 0 out of 1

2. Find the equation of the tangent line to  $x^3$  at  $x=2$ .

Your last answer was interpreted as:

$$4*(x-2)+8$$



**Your answer should be an equation, but is not.**

Figure 1: A STACK question

However, there are new and significant difficulties. Students must enter an expression into a machine and this is non-trivial: see [6]. Students may also change their strategy to that of “guess and check”. Perhaps most seriously, in such CAA the only evidence available on which to base outcomes is often the *final answer*. In particular, the working is not available. All technology for gathering students’ responses as part of an assessment system effect what is gathered and how it is gathered. The MCQ format (whether automated or on paper) is an extreme example where the format distorts questions in a manner which is usually unacceptable. Such distortion is inherent in all formats, including traditional paper-based tests.

With traditional paper-based work, there is a significant delay in providing feedback to students. As a result it is common to provide “partial credit” for correct working and sometimes to perform “follow-through” marking to ensure students are rewarded for what they can do, rather than penalizing a small mistake by ignoring all subsequent work. Such follow-through marking may send out mixed messages to students about the importance of accuracy as a goal. As an extreme example: they may be able to accrue significant credit, whether in a formative or summative environment, and yet provide no correct answers whatsoever!

The rapidity of the feedback in CAA could allow the student to re-attempt any part where their response does not meet *all* the relevant criteria. Hence, “follow-through” marking would not appear to be necessary in such an environment and the importance of accuracy as a goal could be reinforced by the computer-based format. Evaluation of the STACK CAA system (see [4, 5]) has highlighted two key issues: first how to enter a mathematical expression (see [6]), and second how assessment of “steps in a calculation” might be implemented in a way which genuinely enhances students’ learning.

This paper looks at an attempt to address this last point, using a newly-developed system (see [1])

as a medium for some of the discussions. The system considered lies within STACK and allows students to enter their attempt at the solution to a problem line by line, including comments describing their actions at each step; this working is assessed in four different ways which are explained below. Located in the United Kingdom, we consider the education of undergraduate mathematics and engineering students.

## **Nature of practice and its role in learning**

Mathematics courses in the early years of science, engineering and mathematics degrees face a dual challenge: to develop mathematical insight and to embed basic skills. CAS supported CAA can help with both. It allows a greater exposure to open ended, exploratory questions [2] which require more creative thinking rather than the application of routine procedures. But it is in the development of basic skills and the application of standard procedures that CAA has been used extensively at many universities. It is not sufficient to merely provide question banks. Different learning outcomes can be obtained using different assessment strategies. At Birmingham, AiM based [7] question banks of over 1 000 questions, covering core topics from differentiation and integration to linear vector-spaces, are used in a variety of ways:

1. Access to questions, sorted by topics is provided for genuine practice, with feedback provided on request and the ability to ask for a new variant of the same question.
2. Questions sorted in assessed quizzes, typically covering a week's exposure of lectures, are used to encourage practice. A penalty system, reducing the maximum obtainable mark by 33% for each attempt is meant to encourage students to think carefully before submitting the answer. It also encourages critical self-evaluation in that the students take primary responsibility for checking their own answers rather than comparing them with model solutions. Solutions are provided after the deadline for the assessment has passed.
3. Questions sorted in assessed quizzes, again covering a week's exposure of lectures each, are used to encourage engagement with the whole of the course content. Students are allowed as many attempts as needed but need to achieve the correct outcome on all questions in all quizzes to gain any mark at all. Students are therefore expected to have studied all the basic skills or procedures in the course. This addresses the observation that many students fail to recognise the importance of building a solid foundation for the future and favour a minimal effort approach to passing each module.

These are just illustrations of how CAA can be embedded in assessment strategies to enhance learning and teaching in a course. In all of these, the ability to provide step-wise feedback will be an added advantage. In too many cases, students are still lost as to how to get started on a question, or lose track of what they were working towards. Certain CAA packages have already incorporated options for the students to choose between supplying a final answer and gain 100% or to have the question broken into steps at a penalty. The feedback given at each step and/or

the ability to provide partial marks, can, within an appropriate assessment strategy, constitute a significant further enhancement to the level of engagement between a student and the curriculum.

The ultimate reality in many universities is that far too little written work can be marked and returned to students in a reasonable time-frame. CAA provides a complimentary activity which couples instantaneous feedback with a consistency which is often lacking in larger class environments where multiple markers are used. Where it is widely believed that assessment can drive learning, another factor in encouraging regular practice is the knowledge of instantaneous and informative feedback.

## Steps in working

In any piece of mathematical work students move from the question to their solution via a series of *steps*. Often these are based, at least initially, on an example or a known algorithm which they have been taught is an effective approach to the particular type of problem. Some areas of mathematics lend themselves more readily to differing approaches, such as proof, whilst others are frequently perceived as being highly algorithmic. When such work is marked there is often a mark scheme based on some ‘ideal’ route to the solution of a problem. To provide a guide to how one might automate assessment of students’ working it is important to consider what causes a particular piece of work to appear correct.

When solving a problem in an area such as ordinary differential equations (ODEs) students may follow the ideal route either directly or with some additional, or even superfluous, steps, but there are certain checkpoints which suggest the student is following the correct path.

$x \frac{dy(x)}{dx} + 4y(x) = x$ $\frac{dy(x)}{dx} + \frac{4}{x}y(x) = 1$ Integrating factor = $e^{\int \frac{4}{x} dx}$ Integrating factor = $x^4$ $x^4 \frac{dy(x)}{dx} + 4x^3y(x) = x^4$ ...	$x \frac{dy(x)}{dx} + 4y(x) = x$ Integrating factor = $e^{\int \frac{4}{x} dx}$ Integrating factor = $x^4$ $\frac{d}{dx} (x^4y(x)) = x^4$ ...	$x \frac{dy(x)}{dx} + 4y(x) = x$ Integrating factor = $x^4$ $x^4y(x) = \frac{x^5}{5} + c$ ...
--	---	--

Table 1: Possible sets of checkpoints for the solution of an ODE

Table 1 shows three possible checkpoint sets for a particular problem and demonstrates the need to decide how proscriptive is it appropriate to be about the contents of a student’s solution. There are students who will solve a problem by writing only a bare minimum of steps; the decision must take into account what is considered to be excessive brevity, and which steps are not essential to a valid solution. For the purposes of the work described here, checkpoints were chosen by minimising the number of essential steps required, on the basis of teaching and assessment experience.

However, a minimal solution is not necessarily the same as a model solution. A minimal solution may be used as a mark scheme to assess whether students’ work contains the required steps and

uses an appropriate method. A model solution is what might be provided to students and which clearly shows the progress from the problem to the solution. The disparity here is not unusual since in teaching there is often this same difference between ‘you *must* include’ and ‘you *should* include’; that is, the essentials of a solution and the elaborations of the same.

When a piece of work is marked there are a number of different types of feedback that can be provided, and which the new STACK system gives, based on the mathematics and associated comments given; figure 2 shows three of the main types.

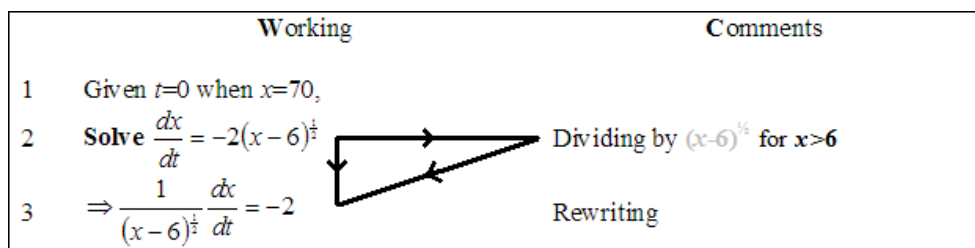


Figure 2: A example of different feedback types

The horizontal line refers to the *appropriateness* of the action chosen, or in other words, ‘has the student chosen a sensible next step?’ This can be provided on the basis of whether the comment matches that from the equivalent point in the model or minimal solution. All we can say at this point is that if the student follows the action of the teacher then they will follow a method which works. Deviating may result in a correct alternative or may be a mistake.

The diagonal line refers to the *application* of the action; ‘has the student correctly done what they said they would?’ This can be provided by checking the equivalence of the current line of working with what the associated comment implies it should have been. An example of a misapplication error would be if a student said they would integrate, but instead differentiated, the previous line of working

The vertical line refers to mathematical *derivation* of the current step from the previous one; ‘have they performed a mathematically permissible operation?’ This requires application of any conditions given in the comment, such as extrapolating an integrating factor from the ODE on the previous line, and then checking equivalence between the adjusted previous and current lines of working. A simple example of a derivation error would be if a student said they would find the integrating factor of a certain ODE but miscalculated the coefficient of  $y(x)$  to be integrated.

One final type of feedback given that is not shown in figure 2 is the most simple, *correctness* of their final solution; ‘have they got the right answer?’ Some assessors in mathematics might consider this to be sufficient information when marking formative work, but it is the authors’ belief that use of all four types of feedback/assessment helps students more readily identify and correct misconceptions or misrecollections about their working.

## User's responses

In the early stage of designing the software, colleagues from four different universities in England (Birmingham, Coventry, Durham and Loughborough) were interviewed about their experiences teaching ODE-related courses. Mention was also made of the intention to create the new version of STACK, and the lecturers' opinions sought on this.

The lecturers interviewed all showed interest in the idea of the new software, one colleague listing properties which he would like to see, including many which were in the software design: "it would be useful to have some sort of interactive thing as you progress through the question... if you get them to show their work, you want to be able to interact with the different steps of the work".

Once a simple version of the software was developed it was trialed with three small groups of students from different universities. There was at least a week between each trial, which allowed some changes to be made after each one depending on certain feedback received in the focus groups run after the students had explored the software for up to 90 minutes.

After the first trial students commented that they wanted the system to reflect paper-based working more closely. Specifically they wanted to be able to refer back to any previous line of working rather than just the default of the preceding line; also, students wanted to be able to hide/unhide any preceding lines of working, such as ones where incorrect working had been carried out. These changes had been successfully implemented by the time the second trial ran.

There were many common areas of comment among the three groups, all of which were taken into consideration in the creation of the next version of the software. For example, students seemed to agree that getting used to the software took time and would be greatly facilitated by some kind of "help book" to explain how to enter responses etc. However, in lieu of such help files, the existing *Theory* and *Worked Example* options were felt to be useful in this manner, as well as in providing support for mathematically struggling students.

One area that caused much discussion amongst the students was the question of how often they would prefer feedback. This question was put to some of the students who were interviewed before the software trial, as well as to the three focus groups. Though there was no consensus on this matter it seemed that general opinion was swayed towards providing 'on-demand' feedback rather than line-by-line or at-the-end, since this option made the other two available to the student should they wish it. This had not been the case with those students who were interviewed but exposure to the system had, by their own admission, changed some minds about the preferability of line-by-line feedback.

Overall, students were positive about the software and seemed to feel that the principles behind its design were appropriate; for example, the method of displaying the working alongside the student's comments was felt to be clear and helpful. One student, despite struggling initially to work with the necessary syntax, went so far as to ask when the software would be available for him to purchase for personal use.

## Conclusion

This paper reports our experiences of automatically assessing student's work with a particular CAS-supported assessment system. We report a new model for allowing the entry and assessment of line-by-line commented working, and some early results from trials given. This is a rather complex process, in which both the student and teacher need to be explicit at a fine grained level of detail not usually considered in the traditional paper based approach.

Additional studies are being planned to gain much more data on student usage of the system and the different types of feedback it offers. This is with the aim of being able to identify those types of automated feedback students prefer, and which appear to have the greatest impact on performance in the questions.

Whilst there is still more research to be done, it is hoped that the system and the concepts of learning it represents will stimulate discussion on formative assessment practices, both automatically and manually.

## Acknowledgements

E. Cerval-Peña would like to thank the UK Higher Education Academy and the sigma Centre for Excellence in Teaching and Learning for funding his doctoral studies.

## References

- [1] E. R. Cerval-Peña. Computer-aided formative assessment of ordinary differential equations: a new approach to feedback. In *Proceedings of the Mathematical Education of Engineers Conference*, Loughborough University, 6–9th April 2008.
- [2] C. J. Sangwin. [On Building Polynomials](#). *The Mathematical Gazette*, 89(516):441–451, November 2005.
- [3] C. J. Sangwin. [STACK: making many fine judgements rapidly](#). In *CAME*, 2007.
- [4] C. J. Sangwin. [What is a Mathematical Question?](#) In *Proceedings of the JEM conference, Lisbon, Feb 2007*, 2007.
- [5] C. J. Sangwin and M. J. Grove. [STACK: addressing the needs of the “neglected learners”](#). In *Proceedings of the First WebALT Conference and Exhibition January 5-6, Technical University of Eindhoven, Netherlands*, pages 81–95. Oy WebALT Inc, University of Helsinki, ISBN 952-99666-0-1, 2006.
- [6] C. J. Sangwin and P. Ramsden. [Linear syntax for communicating elementary mathematics](#). *Journal of Symbolic Computation*, 42(9):902–934, 2007.
- [7] N. Strickland. Alice interactive mathematics. *MSOR Connections*, 2(1):27–30, 2002. <http://itsn.mathstore.ac.uk/newsletter/feb2002/pdf/aim.pdf> (viewed December 2002).