# High Stakes Automatic Assessments: Developing an Online Linear Algebra Examination

Christopher J. Sangwin
School of Mathematics
University of Edinburgh, Edinburgh
C.J.Sangwin@ed.ac.uk

## Abstract

In this paper I investigate development of an automatically marked online version of a current paper-based examination for a university mathematics course, and the extent to which the outcomes are equivalent to a paper-based exam. An online examination was implemented using the STACK online assessment tool which is built using computer algebra, and in which students' answers are normally typed expressions. The study group was 376 undergraduates taking a year 1 Introduction to Linear Algebra course. The results of this experiment are cautiously optimistic: a significant proportion of current examination questions can be automatically assessed.

## 1 Introduction

To what extent can we produce an automatically marked online version of a current paper-based examination for methods-based university mathematics courses using contemporary technology? To what extent are the outcomes of this exam equivalent to the outcomes of a paper-based exam? And more speculatively: if we automate exams are we perpetuating "incantation" or moving towards "enlightenment"?

In this paper I report a pilot to develop, use, and evaluate an online examination for a university linear algebra course. The STACK online assessment tool [San13] is built using computer algebra and students' interactions move significantly beyond multiple choice questions with their well-known difficulties for mathematics, see [SJ17]. In particular STACK uses the computer algebra system Maxima to generate random questions; interpret students' typed algebraic expressions; establish objective mathematical properties of students' answers; and assign outcomes such as feedback and marks.

My work is based upon the epistemological position that to successfully automate a process it is necessary to understand it profoundly. It follows that automation of a process necessitates the development of a certain kind of understanding and we learn a lot about the underlying process through automation. Assessment provides students with challenge, interest and motivation: assessment is a key driver of students' activity in education. Assessment often defines the course of study, and even defines the nature of the subject itself. To many students mathematics is defined in a large part by what we expect students to do in examinations, [Bur87].

Online assessment has for many years been used widely in formative settings, see [San13]. Developing an online automatic high-stakes final examination is a natural extension of automation of formative assessment.



In: O. Hasan, W. Neuper, Z. Kovcs, W. Schreiner (eds.): Proceedings of the Workshop *CME-EI: Computer Mathematics in Education - Enlightenment or Incantation*, Hagenberg, Austria, 17-Aug-2018, published at http://ceur-ws.org

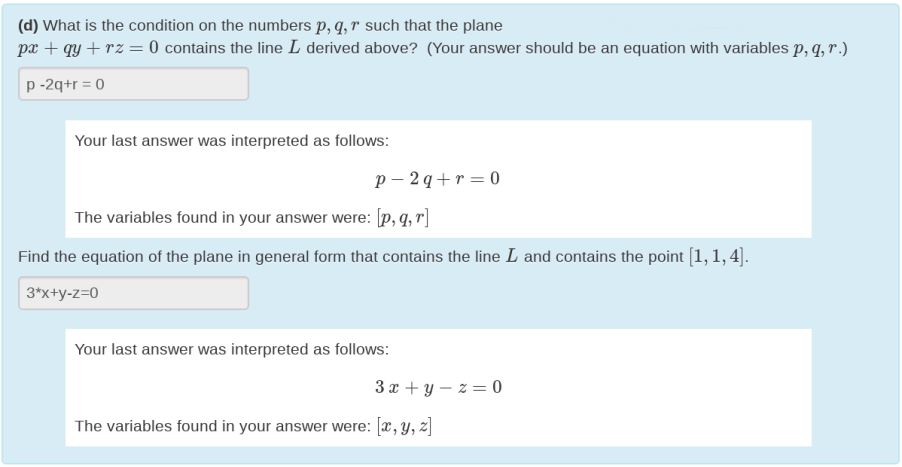


Figure 1: Question 19 of the current study in STACK

Automation also has practical benefits, reducing the marking load and potentially speeding up examination processes. However, changing written examinations, with centuries of custom and practice, is a high-stakes and high-risk undertaking. My previous joint work [SK16] examined questions set in school-level examination papers with a view to developing automatically marked online versions. The results of [SK16] were cautiously optimistic that a significant proportion of current questions could be automatically assessed. In this paper I extend this work, and create examination questions and trial their use with a large group of university students.

## 2 Methodology

For this study I added a mock online examination to Introduction to Linear Algebra (ILA). This is a year 1, semester 1, mathematics course worth 20 credits taken by mathematics, computer science and other undergraduate students. Students normally take 120 credits per year, in two semesters. The course is defined by [Poo11] Chapters 1 to Chapter 6.2, with a selection of the applications included and selected topics omitted. ILA had over 600 students, of whom 578 took the final written examination and had a non-zero examination mark.

Students had requested exam practice, but it was impractical to administer and mark students' attempts (approximately 35 person-days for the genuine exam) in the short period between the end of teaching and the scheduled examination. In context, a mock examination was likely to be taken seriously by a significant proportion of the student cohort as a valuable practice and learning opportunity. Since the mock examination did not contribute to the overall course grade there was no incentive for students to cheat, or to be impersonated. Introduction to Linear Algebra, has an "open book" examination and so possible access to materials is less of a threat to this experiment than would be the case for a closed-book examination. The lack of certainty over who was sitting the online tests, the circumstances of participation, the potential use of internet resources and so on is certainly a compromise. Such uncertainty does not affect the extent to which I could produce questions at a technical level, or the effectiveness of the scoring mechanism in the face of students' attempts.

The results consist of a report on the extent to which current questions can be faithfully automated, and I give a preliminary report on students' attempts.

## 3 Results

The existing paper-based ILA examination takes 180 minutes and consists of Section A: compulsory questions totalling 40 marks, and Section B: four questions each of 20 marks from which we take the student's best three marks. Students may use any standard scientific calculator but graphical calculators with matrix functions are not permitted.

The primary goal was to provide students with an online examination which was as close as possible to the forthcoming paper-based summative course examination. ILA has been running for many years, with a stable (but not invariant) syllabus, and I had access to examinations going back to December 2011 (two per year: the main exam and an equivalent resit paper). I therefore decided to remove the oldest exam papers from easy access

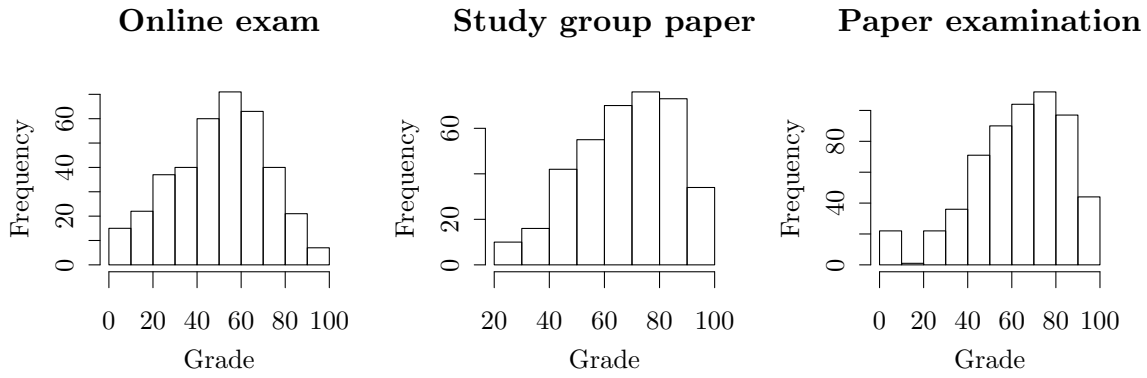


Figure 2: Histograms of achievement in the online mock and paper based examinations

through the course website and base the online examination on those questions. Using as few papers as possible helps provide a representative online examination. Technically it is difficult to operate a "best 3 out of 4" mark scheme in the STACK online system and in any case for a formative mock exam this makes little sense.

In deciding how to allocate marks I have taken a strict interpretation. Specifically, where the original intention of the examiners included "with justification", I only awarded a minimum number of marks for giving the answer only online. For example, Q5 on our online exam asked the following.

5. Is it possible for $A$ and $B$ to be $3 \times 3$ rank 2 matrices with $AB = 0$? True/False.

The original paper awarded 7 marks for the answer and justification, whereas only one mark was awarded for the correct answer. I did ask students to provide typed free-text justifications even though these would not be marked and no feedback was provided.

Ultimately I used two papers (120 marks each) to create the online exam with 59 marks of the online exam coming from Dec-11 and 50 marks from Aug-12. I took one question from Dec-13 to add a mark to Section A to make the online exam total 110 marks. Of the paper-based questions selected for the online exam, 44 marks are not awarded online. These missing marks are for justification which cannot, at this time, be automatically assessed. This resulted in Section A having fewer marks than would be the case with a paper based submission. Of the 240 marks available on the Dec-11 and Aug-12 papers, 109/240 marks 45% were automated in a way faithful to the original examinations. However, the online versions do lack some partial credit and do not (in this experiment) implement follow on-marking, which in some Section B questions is substantial.

An example question is shown in Figure 1, illustrating "validity" feedback which was available during the exam. Validity feedback is normally available to a student, and provides information on syntax errors and other input problems helping reduce the extent to which students are penalized on a technicality. For ILA, online course work quizzes were already implemented using STACK. All students were expected to sit 30 online quizzes using the STACK system as part of the ILA course before the mock examination, and would be thoroughly familiar with how to enter answers into the system.

The online examination was made available to students to do in their own time for a period of one week in December 2017, between the end of formal teaching and the scheduled paper-based exam. Students could choose when to sit the online examination, but were given one attempt of 180 minutes to do so to simulate examination practice. All data was downloaded from the online STACK system, and after ratification by the exam board, combined with overall achievement data. Students were assigned a unique number to ensure anonymity, and the data loaded into R-studio for analysis.

There were 395 attempts at the mock online exam in December 2017. One student who was granted a second attempt for technical reasons had their first attempt disregarded, giving 394 attempts. There were no other significant technical problems affecting the conduct of the online examination. For the online exam (including those who scored zero) the mean grade was 47.9% with standard deviation of 23.2%. The coefficient of internal consistency (Cronbach Alpha) for the online exam was 0.87. There was a moderate positive correlation between time taken (M=132 mins, SD=48.6 mins) and the online exam result (M=47.9%, SD=23.2) $r(392) = 0.517$, $p < 10^{-16}$, as might be expected. Despite a small number of outlier questions, the mock online exam appears to have operated successfully in its own right as a test.

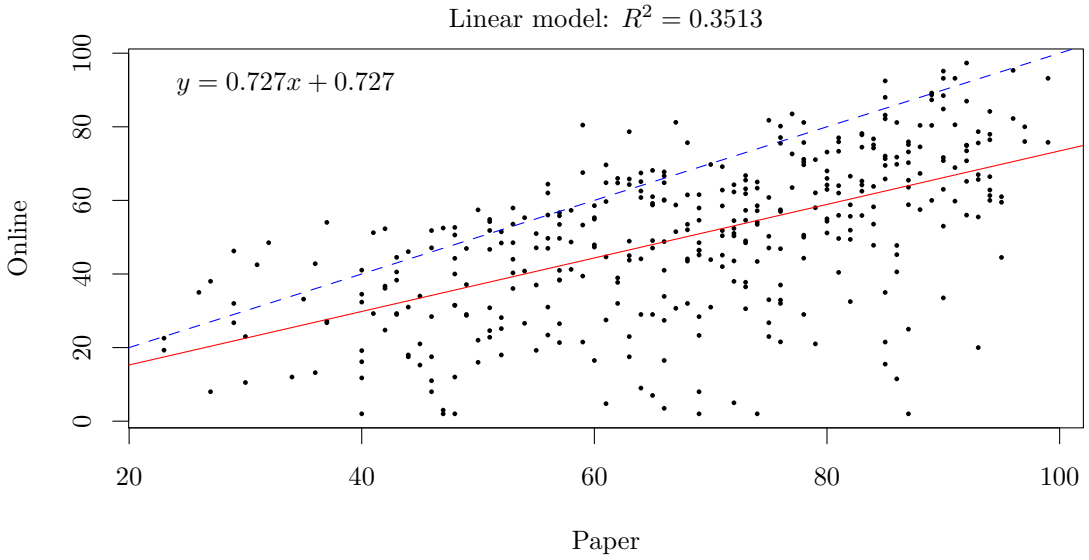


Figure 3: Online mock exam grades *vs* paper exam grades for the study group

| | | |
|---|---|---|
| Q4 | 276 | (M=187, SD=167) |
| Q6 | 211 | (M=125, SD=117) |
| Q8 | 242 | (M=204, SD=155) |
| Q10 | 249 | (M=128, SD=124) |
| Q12 | 220 | (M=201, SD=194) |
| Q15 | 227 | (M=142, SD=145) |
| Q22 | 213 | (M=173, SD=145) |
| Q26 | 194 | (M=67, SD=69.4) |

Table 1: The number of free text responses, together with mean number of characters $M$ and the standard deviation of the response length

The final mark for ILA is made up of coursework (20%) and a final paper-based exam (80%). There were 394 attempts at the online mock examination, and all but one of these students also sat the paper-based examination. Note that 17 students scored 0 for the online exam, perhaps indicating students who looked at the online questions but made no serious attempt at them. Technically there is a difference between students who never sat the online exam, and those who opened the exam and scored 0. For the analysis I excluded the 17 students who scored 0 in the online exam: this leaves the *study group* of $N = 376$ students with paper and mock exam information.

For the study group, the online exam results had (M=50.2, SD=21.3) and paper exam (M=68.0, SD=17.3). For all students who sat the ILA paper exam (M=63.1, SD=21.6). Histograms of achievement in the online mock and paper based examinations are shown in Figure 2. There is a significantly larger failure rate (score less than 40%) in the online examination, and a significantly lower mean. These differences could be explained by the level of engagement: the online exam carried no credit, and students may have lost motivation when tired.

A scatter plot of the online mock exam grades *vs* paper exam grades is shown in Figure 3, together with a linear regression model. The dashed line shows the (ideal) linear relationship in which the online mock examination has identical outcomes with the paper-based exam. Notice the online exam scores are clearly below those of the paper exam, supporting the hypothesis that students may have lost motivation when tired and not performed to their full potential in the online mock exam. The mock exam grades and paper exam grades were moderately correlated, $r(374) = 0.593$, $p < 10^{-15}$.

The number of non-empty free text responses to each of the "justify" questions is given in Table 1, together

with the mean and standard deviation of the response length (in number of characters). It is clear reading through the free-text responses that over 200 students took the exercise seriously, providing sensible (and often correct) justifications in good English. For the Section A questions in paper there were 59 marks available, whereas in the STACK exam only 24 marks were awarded. I did not expect students to make serious use of the free-text entry. The fact students entered sensible justification to many of these questions, and received no marks, could easily account for the difference in mean scores between the paper-based and online exam. There were a large number of empty responses (as there are on paper as well), together with some incoherent utterances, and some plaintive messages. I did not assess these free-text responses, or subject them to comprehensive analysis for the purposes of this paper. However, in a genuine online examination such responses could be assessed (1) manually in the traditional way on-screen, (2) using automatic assessment technology such as described in [BJ10, Jor12], or (3) using comparative judgement for longer passages, see [JSP14, Pol12].

## 4 Discussion

The implementation of the mock online examination for linear algebra was a modest success. There were no serious technical problem during the conduct, and no students complained of inaccurate or unfair marking. The results of the online examination were broadly comparable with a paper-based exam, with the consistently lower online performance explained by a combination of (1) potential disengagement in a low stakes setting, (2) lack of assessment of students' justification, (3) lack of partial credit and follow through marking. Both partial credit and follow through marking are technically possible in STACK, but are expensive (in staff time) to implement. The results give us confidence to use such assessments in higher-stakes settings in the future.

This research has done nothing to address serious practical problems associated with online examinations. Problems include the need for invigilation to reduce plagiarism and impersonation, and security to eliminate communication during the exam (such as answer sharing) or access to unauthorised resources. These examination conduct problems must be solved, but they have nothing to do with mathematics.

Automatic assessment is an area of mathematics which would particularly benefit from tools which automate explanation, justification and reasoning. In particular "proof checking" software, as applied to students' understanding, is necessary to move beyond assessing only a final answer, as shown in Figure 1, to a full mathematical answer. In this study, only students' final answers were subject to automatic assessment which is a serious limitation. However, progress is being made to assess working especially in the area of reasoning by equivalence as discussed briefly in [SK16].

I was surprised at the large extent to which existing questions could be automatically assessed with the current tools, based on computer algebra, faithfully. However, there is nothing sacrosanct about current examination questions. Why should the online examination be exactly the same as a paper-based examination? Current questions are written explicitly for the paper-based format, and it is sensible to seek to write questions which take advantage of the online format as appropriate. Many Section A are true/false, but the justification of a "false" response is via appropriate examples. Computer algebra is ideally suited to assessing answers, such as counter examples, which expect the teacher to perform some time-consuming and error-prone calculation. For this research I did not rephrase questions to "give me examples, such that ...", but this would be one option.

This analysis raises the question of whether we, as a mathematics community, believe current mathematics examinations are a valid test of mathematical achievement. Do current examinations actually represent valid mathematical practice, as undertaken by researchers, industrial mathematicians and for pure recreation as an intellectual pursuit? Construct validity is a central educational concern, but it is not relevant to the research question of whether we can actually automate current exams. My personal views about the nature of mathematics broadly align with those expressed in [Pol54] and [Lak76]. That is, that setting up abstract problems and solving them lies at the the heart of mathematics. [Pol62] identified four patterns of thought to help structure thinking about solving mathematical problems. His *"Cartesian"* pattern is where a problem is translated into a system of equations, and solved using algebra. Note that the algebraic manipulation is the technical middle step in the process: setting up the equations and interpreting the solutions are essential parts to complete this pattern. My previous work [SK16] examined questions set in school-level examination papers and found that line-by-line algebraic reasoning, termed *reasoning by equivalence* [NBC04], is the most important single form of reasoning in school mathematics. However, many examination questions do not relate to a problem at all, rather they instruct students to undertake a well-rehearsed set of techniques, isolated from any problem. Many of the questions in the ILA examinations also rely on predictable methods which can be well-rehearsed. Predictable methods predominate in school examinations, such as those considered in my previous research in [SK16]. Current

examinations tend towards "incantation" by students, and there is a real danger that national examination boards, universities, and others with responsibilities for examinations will replicate traditional examinations online without a critical reassessment of the purpose of mathematics education.

## 5 Conclusion

The increasing use of software tools in online assessment will affect mathematics education. It is likely that automatic online examinations for mathematics in school, as well as for some methods-based university courses, will become feasible and will be used in the very near future. A pragmatic combination of computer algebra supported assessment and automatic assessment of short answer questions will assess a significant proportion of current questions automatically. Traditional expert marking, and use of comparative judgement will potentially widen the scope of exams at the expense of complete automation. A further pragmatic approach will be to split courses into two summative assessment components: largely skill-based questions can be automatically assessed online, with the justification and rhetorical discussion in a traditional written examination. Replication of traditional examinations online without a critical reassessment of the purpose of mathematics education would be a wasted opportunity to define the subject through valid assessments.

#### 5.0.1 Acknowledgements

The online questions were created with the help of Dr Konstantina Zerva, of the University of Edinburgh.

## References

[BJ10] P. G. Butcher and S. E. Jordan. A comparison of human and computer marking of short free-text student responses. *Computers and Education*, 55(2):489–499, September 2010.

[Bur87] H. Burkhardt. What you test is what you get. In I. Wirszup and R. Streit, editors, *The Dynamics of Curriculum Change in Developments in School Mathematics Worldwide*. University of Chicago School Mathematics Project, 1987.

[Jor12] S. Jordan. Student engagement with assessment and feedback: Some lessons from short-answer free-text e-assessment questions. *Computers and Education*, 58(2):818–834, 2012.

[JSP14] I. Jones, M. Swan, and A. Pollitt. Assessing mathematical problem solving using comparative judgement. *International Journal of Science and Mathematics Education*, 13(1):151–177, 2014.

[Lak76] I. Lakatos. *Proofs and refutations*. Cambridge University Press, 1976.

[NBC04] J. F. Nicaud, D. Bouhineau, and H. Chaachoua. Mixing microworlds and CAS features in building computer systems that help students learn algebra. *International Journal of Computers for Mathematical Learning*, 9(2):169–211, 2004.

[Pol54] G. Polya. *Mathematics and Plausible Reasoning. Vol.1: Induction and Analogy in Mathematics. Vol 2. Patterns of Plausible Inference*. Princeton University Press, 1954.

[Pol62] G. Polya. *Mathematical discovery: on understanding, learning, and teaching problem solving*. Wiley, London, UK, 1962.

[Pol12] A. Pollitt. The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3):281–300, 2012.

[Poo11] D. Poole. *Linear Algebra: a modern approach*. Brooks/Cole, Cengage learning, third edition, 2011.

[San13] C. J. Sangwin. *Computer Aided Assessment of Mathematics*. Oxford University Press, Oxford, UK, 2013.

[SJ17] C. J. Sangwin and I. Jones. Asymmetry in student achievement on multiple choice and constructed response items in reversible mathematics processes. *Educational Studies in Mathematics*, 94:205–222, 2017.

[SK16] C. J. Sangwin and N. Köcher. Automation of mathematics examinations. *Computers and Education*, 94:215–227, 2016.