Developing and evaluating an online linear algebra examination for university mathematics

Christopher Sangwin

Developing and evaluating an online linear algebra examination for university mathematics

Christopher Sangwin[1]

[1] University of Edinburgh, School of Mathematics; C.J.Sangwin@ed.ac.uk

In this paper I report 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, and the quantitative outcomes are moderately correlated with the paper examination.

Keywords: Mathematics, undergraduate study, high stakes tests.

Introduction

To what extent can we produce an automatically marked version of a 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? In this paper I report a pilot to develop, use, and evaluate an online examination for a university linear algebra course.

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. To many students mathematics is defined in a large part by what we expect students to do in examinations, (Burkhardt, 1987).

The STACK online assessment tool (Sangwin, 2013) is built using computer algebra and students' interactions move significantly beyond multiple choice questions with their well-known difficulties for mathematics, see (Sangwin & Jones, 2017). 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. Online automatic assessment has for many years been used widely in formative settings, see (Sangwin, 2013). Developing high-stakes final examinations is a natural extension of automation of formative assessment. Automation also has practical benefits, such as reducing the marking load, better test reliability, and potentially speeding up examination process. However, changing written examinations, with centuries of custom and practice, is a high-stakes and high-risk undertaking. When I examined school-level examinations (Sangwin & Kocher, 2016) the results were cautiously optimistic: a significant proportion of current questions could be automatically assessed. In this paper I extend this work and create online examination questions and trial their use with a large group of university students.

Methodology

Changing examination processes is both high-stakes and high-risk. Furthermore, there are serious ethical problems with running an experiment in these circumstances without serious and authentic trial exams. Hence, for this study I added a mock online examination to an existing course: 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 (Poole, 2011) 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 in the short period between the end of teaching and the scheduled examination. Students would have expected more detailed formative feedback than provided by the genuine exam, and the genuine exam takes approximately 35 person-days to mark. 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, and the potential use of internet resources 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, which themselves constitute important results and generate key points for discussion. The results consist of a report on the extent to which current questions can be faithfully automated, and I give a report on students' attempts.

Results

The existing paper-based ILA examination takes 180 minutes and consists of Section A: compulsory questions worth 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 teaching goal was to provide students with an online examination which was as close as possible to the forthcoming paper-based summative course examination. The research goals were to evaluate the effectiveness of this, and to provide evidence for a discussion of equivalence with the genuine 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 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 very strict interpretation. Specifically, where the original intention of the examiners included "with justification", I only awarded those marks which could be given for 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*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 automatic feedback was provided. Ultimately I used two papers (120 marks each) to create the online exam. In this 59 marks of the online exam were 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. I think this is a remarkably high proportion, and discuss this in more detail below. However, the online versions as implemented for this study do lack some partial credit and do not (in this experiment) implement follow on-marking, which in some Section B questions is substantial. This is not a limitation of the system itself, but rather in the time available to implement more elaborate automatic marking schemes.

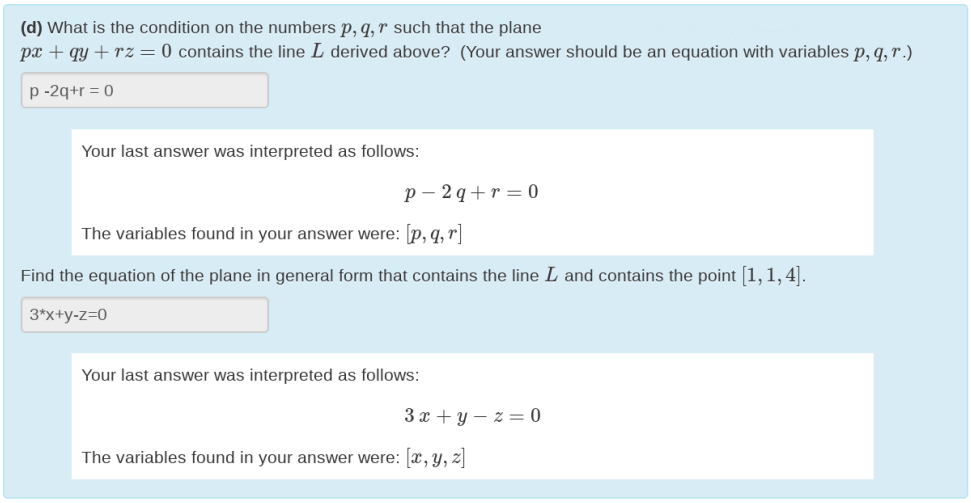

Figure 1: Question 19d of the current study in STACK

Note that STACK requires students to type in an algebraic expression as their answer, and an example question is shown in figure 1. 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.

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 are shown in figure 2. The "online exam" refers to scores on the online mock. The "study group paper" is the achievement of the study group on the genuine paper examination. "Paper examination" refers to the whole cohort of ILA in the genuine paper exam. 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.

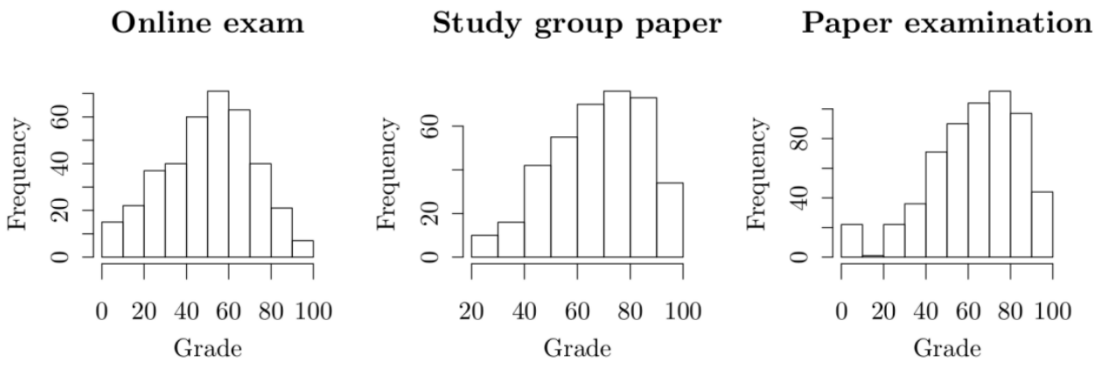

Figure 2: Histograms of achievement in the online mock and paper based examinations

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 blue dashed line shows the (ideal) linear relationship in which the online mock examination has identical outcomes with the paper-based exam. The mock exam grades and paper exam grades were moderately correlated, $r(374)=0.593$, $p<10^{-15}$. 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. Indeed, students scoring less than 40% in the online exam and more that 70% in the genuine exam are very likely not to have taken the online test seriously, perhaps confident (with good reason) about their ability.

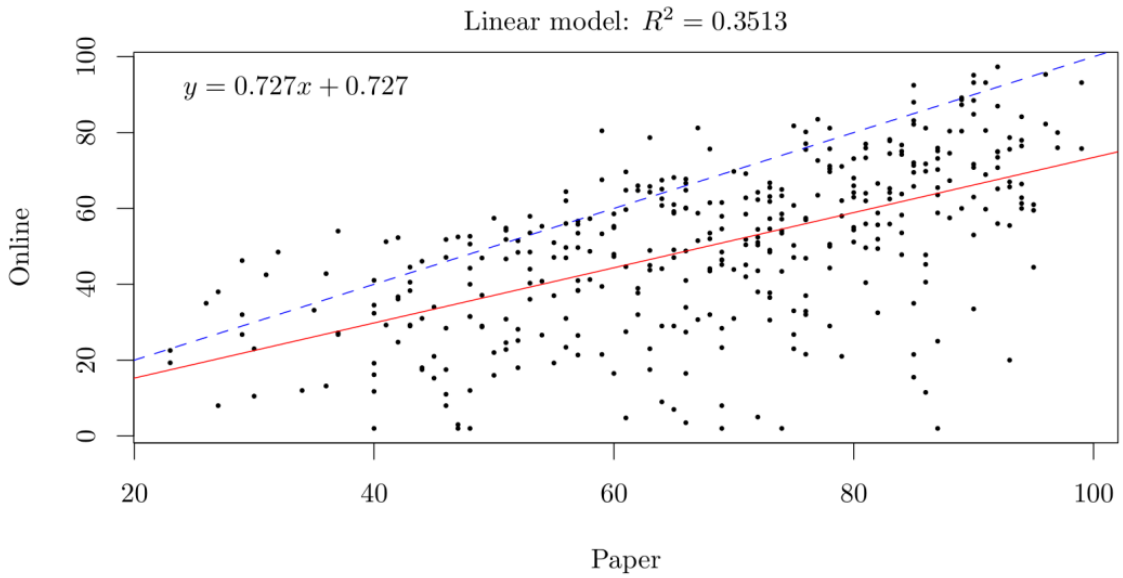

Figure 3: Online mock exam grades vs paper exam grades for the study group

The number of non-empty free text responses to each of the "justify" questions is worthy of mention here. While no planned evaluation of the responses was part of this study, 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 included in this study there were 59 marks available in the paper format, whereas in the online 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 or feedback, 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 (Butcher & Jordan,

2010; Jordan, 2012), or (3) using comparative judgment for longer passages, see (Jones, Swan, & Pollitt, 2014; Pollitt, 2012).

Discussion

The implementation of the mock online examination for linear algebra was a modest success. There were no serious technical problems 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 (which is typically rather generous), and (3) lack of partial credit and follow through marking.

This research has done nothing to address serious practical problems associated with online examinations in general. 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. Indeed, while technology has the potential to support examination processes, there is technology specifically designed to undermine traditional examinations as well. For example the Ruby Calculator (https://rubydevices.com.au retrieved September 2018) is designed to aid unauthorized communication during exams. Either these examination conduct problems must be solved, or we need new models of assessing students. But these examination conduct problems have nothing to do with mathematics as a subject.

I think it is remarkable that 109/240 (45%) of the marks available on paper were automated in a way faithful to the original examinations. Further, by selecting existing questions from two existing past papers I was able to create a fully online exam, with broadly similar syllabus coverage. However, this result can be interpreted as a comment on the mechanical nature of the subject, and of the assessments we use in the traditional examination. If the assessment of students' answers can be automated, then certainly the underlying processes can be automated by the computer algebra system. Why then are students still learning to perform these mechanical processes, e.g. in the context of ILA row reduction, and calculation of determinants and eigenvalues/vectors? Both partial credit and follow through marking are technically possible in STACK, but are expensive (in staff time) to implement. To take this work further we need tools which automate assessment of explanation, justification and reasoning. In particular "proof checking" software, as applied to students' understanding, is necessary to move beyond assessing only a final answer to a full mathematical answer. In this study, only students' final answers were subject to automatic assessment which is a serious limitation to the award of partial credit and method marks. However, progress is being made to assess working especially in the area of reasoning by equivalence as discussed briefly in (Sangwin & Kocher, 2016). The work on reasoning by equivalence is essential for assessing questions in calculus and algebra, two other pillars of pure mathematics. For this reason, I am confident the cautious optimism expressed here about linear algebra exams also extends to mathematics more broadly in year 1 and 2 university methods-based examinations and mathematics examinations at the school/university interface.

I was surprised at the extent to which existing questions could be automatically assessed. 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 are tuned to, or indeed take advantage of, the online format as appropriate. For many true/false questions the justification requires appropriate examples. Computer algebra is ideally suited to assessing answers, such as counter-examples, which expect the teacher to perform some time-consuming and potentially error-prone calculation. For this research I did not rephrase such questions to "give me examples, such that ...", but this would be one option. Specifically we could certainly have

5. Give examples A and B of non-zero 3*3 matrices for which $AB = 0$.

A computer algebra system is ideally a much better tool for assessing such answers than a human marker. Hence, it would be much more sensible to design an online examination with the format in mind. Indeed, often human examiners do not ask students to "give examples of", only because of the work entailed in marking these by hand. That said, to establish face-validity for the online examination it is useful to understand the extent to which we can assess existing questions and to establish that the technical assessment processes are equivalent.

The practical benefits of online automatic examinations include increased reliability, reduction in costs and in swifter marking times. This is very attractive to all stake holders in the process, including students, teachers and end-users of the results. It is highly likely that automatic examinations will become the norm in the near future. Online examinations will happen, but there is no need for them to be restricted to multiple choice formats. Indeed, as a community of educators we can do much better than that. However, 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.

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 (Polya, 1954) and (Lakatos, 1976). That is, that setting up abstract problems and solving them lies at the heart of mathematics. (Polya, 1962) 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 (Sangwin & Kocher, 2016) examined questions set in school-level examination papers and found that line-by-line algebraic reasoning, termed *reasoning by equivalence*, 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 questions in the ILA examinations also rely on predictable methods which can be well-rehearsed.

Informal discussions with colleagues, particularly during the thematic working group 21 during CERME, strongly suggest that online examinations are a concern for many working in mathematics education. The question of validity of all examinations, on paper and online, is central as is the difficult question of retaining validity if an examination format changes. Changing to online examinations provides some unique opportunities but it will be essential for stake-holders to retain confidence in any new assessment regimes, regardless of any significant merits the format brings.

Using sophisticated assessment tools such as STACK we can create a fully automatically marked examination, which is broadly equivalent to current paper-based examinations at the technical level and in terms of outcomes for students. With other tools, we can create a more rounded online examination, perhaps incorporating some human assessment of free-text justification. However, the attempt to automate assessment of students' answers reveals much about what we really ask students to do in examinations.

References

Burkhardt, H. (1987). What you test is what you get. In I. Wirszup & R. Streit (Eds.), *The dynamics of curriculum change in developments in school mathematics worldwide*. University of Chicago School Mathematics Project.

Butcher, P. G., & Jordan, S. E. (2010, September). A comparison of human and computer marking of short free-text student responses. *Computers and Education, 55(2)*, 489–499. doi: 10.1016/j.compedu.2010.02.012

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

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

Lakatos, I. (1976). *Proofs and refutations*. Cambridge University Press.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19(3)*, 281–300. doi: 10.1080/0969594X.2012.665354

Polya, G. (1954). *Mathematics and plausible reasoning. vol.1: Induction and analogy in mathematics. vol 2. patterns of plausible inference*. Princeton University Press.

Polya, G. (1962). *Mathematical discovery: on understanding, learning, and teaching problem solving*. London, UK: Wiley.

Poole, D. (2011). *Linear algebra: a modern approach* (Third ed.). Brooks/Cole, Cengage learning.

Sangwin, C. J. (2013). *Computer aided assessment of mathematics*. Oxford, UK: Oxford University Press.

Sangwin, C. J., & Jones, I. (2017). Asymmetry in student achievement on multiple choice and constructed response items in reversible mathematics processes. *Educational Studies in Mathematics, 94* , 205–222. doi: 10.1007/s10649-016-9725-4

Sangwin, C. J., & Kocher, N. (2016). Automation of mathematics examinations. *Computers and Education, 94* , 215–227. doi: 10.1016/j.compedu.2015.11.014