Running an Online Mathematics Examination with STACK

https://doi.org/10.3991/ijet.v18i03.35789

Christopher Sangwin(✉)
School of Mathematics, University of Edinburgh, Edinburgh, Scotland
C.J.Sangwin@ed.ac.uk

Abstract—In this paper I discuss running an online examination with STACK for a year 1 "Introduction to Linear Algebra" course. The COVID pandemic has continued to disrupt teaching during 2021–22, and were are still not able to return to the examination hall for a traditional exam. Instead, our examinations were still "take home". Under these circumstances we wrote a fully automatically marked final test for the course. With over 800 students on the course in 2021–22, the paper-based examination was estimated to cost about 50 person-days to mark. Clearly, reducing this cost is an attractive prospect. However, important questions remain. To what extent can we write questions which cover the learning objectives of the course? How did the students do with the test? What recommendations can we offer for similar courses, and future years? In particular, do we need to return to the exam hall?

Keywords—online assessment, examinations, STACK

1 Introduction

1.1 Background to STACK

STACK is a highly-sophisticated system designed to support automatic online assessment of mathematics, see [5]. In particular, STACK extends the quiz systems of the Moodle and ILIAS learning management systems (LMS), with a novel "question type". STACK is open-source software and so can be customized and extended, based on the needs of our users. This freedom to extend, and improve, the software has been key in STACK's success. See https://stack-assessment.org/ for further details of the project.

The basic premise is that students should enter an answer as a typed mathematical expression and then STACK uses Maxima, a computer algebra system (CAS), to support the assessment process. A key feature of STACK is the ease with which the results of computer algebra calculations, based on the student's answer, can be incorporated within in the feedback. STACK has the following key features:

- Structured random question variants can be generated.
- STACK has a unique multi-part question design, allowing different types of input beyond algebraic expressions, including multiple choice and scientific units.

- STACK carefully separates validation of students' input from more formal assessment to prevent penalizing a student for syntax errors.
- STACK establishes mathematical properties, such as algebraic equivalence with a correct answer.
- Outcomes include numerical marks, formative feedback, and data on all attempts are stored for later analysis.

An example STACK question shown in Figure 1, which illustrates some key features. To assess the student's answer, the system establishes a number of properties separately, and assembles the outcomes (marks, feedback). First note that for this question there is not a unique correct answer. We need to check that the right-hand side is zero (that is homogeneous). Then we need to check that the matrix on the left-hand side is singular, which is done by calculating the determinant of the student's matrix. The system can provide detailed feedback, either immediately, after a student has completed the quiz, or after a particular time.

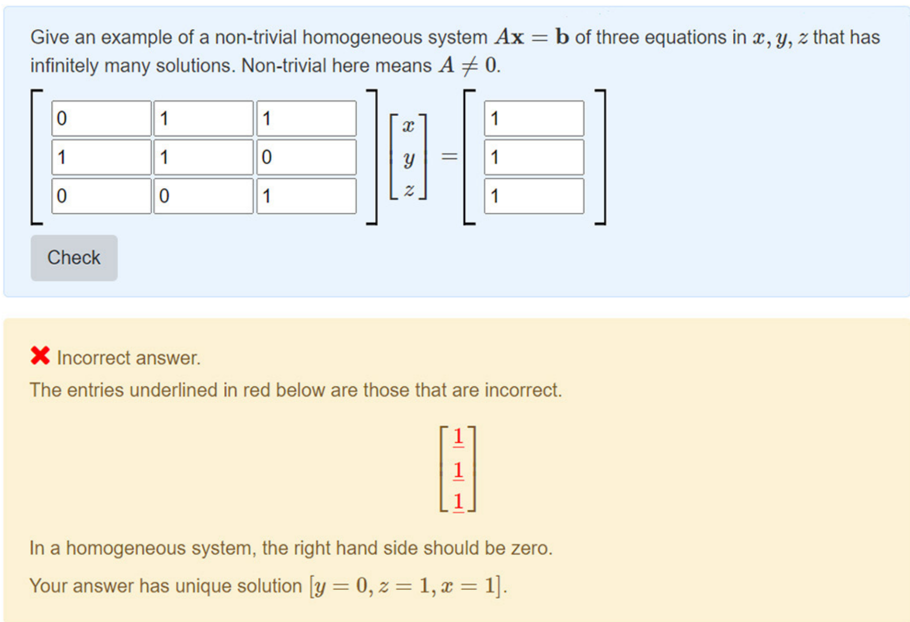

Fig. 1. A typical STACK question with feedback

In Figure 1 the last line of the feedback gives the unique solution of the student's system of equations, illustrating the extent to which STACK can incorporate calculations into feedback. In this paper we talk about examinations, and while feedback was not available until after marks for the exam were ratified by the exam board, such feedback is widely used in formative situations and is normally available immediately.

1.2 Background to Introduction to Linear Algebra

Introduction to Linear Algebra (ILA) is a university mathematics course taken in year 1, semester 1, by a wide range of undergraduate students at a leading UK university. ILA is a large course, e.g. 578 took the written exam in 2017. ILA is compulsory for students on Mathematics degrees, including the flagship BSc degree and all joint honors degrees in mathematics. ILA is also compulsory for students on all Computer Science degrees, and is also taken by a wide range of other year 1 students as an optional course. ILA attracts 20 "credits" in the UK university system, in which students take 120 credits total per year.

The basic topics of the course include vectors in two and three dimensions, including vector algebra. Solving linear equations and systematic Gaussian elimination. Span and linear independence. Matrices and matrix algebra, including inverse matrices and elementary matrices. Subspaces, bases, dimension and rank. Linear Transformations. Orthogonal matrices, and orthogonal bases. Eigenvalues/vectors, determinants, and diagonalization. Orthogonal complements and Orthogonal diagonalization.

In the 2019–2020 academic year, and earlier (i.e. pre-COVID), ILA used a "flipped classroom" approach with pre-reading and interactive assessment. There were 20 online "reading quizzes" which tested if students understood the pre-reading from the textbook [4] before the lecture. There were weekly written assessments, with formative feedback and a weekly online assessment (STACK) to assess more procedural parts of the module. The final grade was made up of 80% for a traditional examination and 20% from weekly coursework, including STACK quizzes.

For the 2020–2021 academic year, and later (i.e. post-COVID), everything moved rapidly online. As a result, we replaced "flipped classroom" lectures with online quizzes. Rather than record, and edit, 50 min online lectures we were emboldened with the prior experiences of other colleagues running fully online courses, see [1]. In particular, we use the metaphor of "putting the book inside the quiz" and used online STACK quizzes to organize the course materials, including numerous short video clips containing the substantive course content. This approach is described more fully in [8].

Further, there was no opportunity for a traditional exam, and so the assessment balance was changed. From 2020–21 the assessment was 40% for written hand-in work, 40% for online quizzes (STACK) and 30% for an online "synoptic assessment" which was in lieu of a traditional examination. The course textbook was changed to [3], a freely available open text, as a response to concerns that supply chain problems internationally would prevent distribution of a traditional printed textbook.

The above provides the context in which ILA took place. In particular, this paper reports the design, use, and outcomes from the online "synoptic assessment" component of ILA: an online examination.

2 Use of online examinations in ILA

ILA takes place in semester 1, from September until December. The pre-pandemic version of ILA had a traditional examination before Christmas. There was consistent demand from students for a practice (mock) examination, and starting in December 2017 we developed and ran a "mock online examination", see [6]. In particular, that previous work addressed two questions.

1. To what extent can we produce an automatically marked version of a paper-based examination for methods-based university mathematics courses using contemporary technology?
2. To what extent are the outcomes of this exam equivalent to the outcomes of a paper-based exam?

The results of that trial were a modest success. In particular, there were no serious technical problems. No students complained of inaccurate or unfair marking and the results of the online examination were broadly comparable with results obtained on the paper-based exam. There were some unsatisfactory features, for example there was no assessment of students' justification (which is typically rather generous), and a lack of partial credit and follow through marking. STACK is able to award partial credit in many situations, but not all and certainly not with the flexibility of a human. As a practice for students in the period between the end of teaching, and the actual examination these were acceptable compromises: the ILA exam typically took about 35 person-days of work for about 600 scripts and a paper-based practice was entirely impractical.

However, the process of developing this mock exam and using it annually in 2017, 2018 and 2019, provided us with valuable experience and welcome confidence. The prior experience enabled us to propose and run STACK-based online "synoptic assessments" for ILA in 2020, and in 2021. We plan to retain this for the foreseeable future. The phrase "synoptic assessment" is used rather than "examination" because the Institution attaches a technical meaning to examination which requires conduct under specific examination conditions. A synoptic assessment was not conducted under these regulations. The rest of this paper evaluates the design, conduct and results from the assessment conducted in 2020–21 (Dec 2020) and 2021–22 (Dec 2021). For students who did not pass the main assessment in December, there is an "Alternative assessment", i.e. a resit exam, in August. The alternative assessment is taken by fewer students each year who are not typical of the cohort, and the assessment is required to be essentially identical in form (but not content) to the original assessment. This provided us with further experience, but these assessments are not reported here.

2.1 Developing online examinations with STACK

To create the mock examination reported in [6] we used two papers containing (120 marks each) to create the online exam. Of the 240 marks available on the Dec-11 and Aug-12 past examination papers, 109/240 marks (45%) were automated with STACK questions in a way faithful to the original examinations. In one respect this is a remarkably high proportion. However, missing marks are for justification which cannot, at this time, be automatically assessed. Acknowledging this, and in fairness to students at a particularly unsettling and uncertain time, we opted for a semi-automatic online assessment in December 2020. This assessment contained a mix of STACK questions, together with questions where students had to type/upload photographs of their answers. The students' answers were then assessed online in the traditional way. In December 2021 we wrote, and used, a fully automatic synoptic assessment using STACK questions. This was a natural progression

In both cases, a range of questions covering the learning objectives of the course were written, developed and checked. These contained a mix of styles of questions including conceptual multiple choice questions, calculations (e.g. "find the eignenvalues and associated eigenvectors of M"). Increasingly, and inspired by prior research such as [2] and [7], we are making use of proof comprehension questions.

Students were given two hours to complete the online assessment, but were free to choose when to sit the assessment within a five day window. Our institution has long experience of writing open-book examinations which take place in a traditional invigilated setting. The online assessment here were "take home" assessments conducted without invigilation. Clearly, the lack of invigilation is a threat to the integrity of examinations, but no examinations were invigilated during the pandemic and so the conduct of the synoptic assessment was in line with other take home traditional examinations, in which a PDF paper was downloaded by students. For this reason we have little to say here about plagiarism, impersonation and other forms of cheating. To mitigate these problems invigilation is probably necessary, and invigilated online exams could take place in traditional settings.

2.2 The 2020–21 assessment, December 2020

The 2020–21 online assessment contained a mix of STACK and human marked questions. Of the 60 marks, 24 were human-marked. A typical human-marked question is shown in Figure 2, and this is very similar to questions on a traditional examination. Two follow-up questions are shown in Figure 3. These five parts form one coherent sequence of questions on the same topic. However, the questions shown in Figure 3 have many correct answers, and would require a human to undertake a significant computation to mark. Typically this is a feature which is avoided in traditional examinations, however valuable such questions might be educationally. Questions requiring significant computation from the examiner are simply never set in traditional exams.

In 2020–21 the assessment functioned very well. The results from 679 students gave a mean of 61%, and standard deviation 18%, which is perfectly acceptable and within departmental norms. It is worthy of note, that most students uploaded photos and did not type.

Semi-automatic marking brings some significant benefits. In particular, staff appreciated separation of "accuracy" from "method". For example, one question asked students to calculate the eigenvalues and associated eigenvectors of a three by three matrix. The semi-automatic approach separated establishing accuracy (Did the student calculate correct eigenvalue/vectors?) which was automatically assessed by STACK from method (Am I confident the person knows how to calculate eigenvalue/vectors?) which was marked by a human. This separation is liberating. Humans did not have to worry about following through mistakes in algebra or integer arithmetic, they only had to decide the level of confidence they had that a student knew the method because accuracy marks are awarded by STACK.

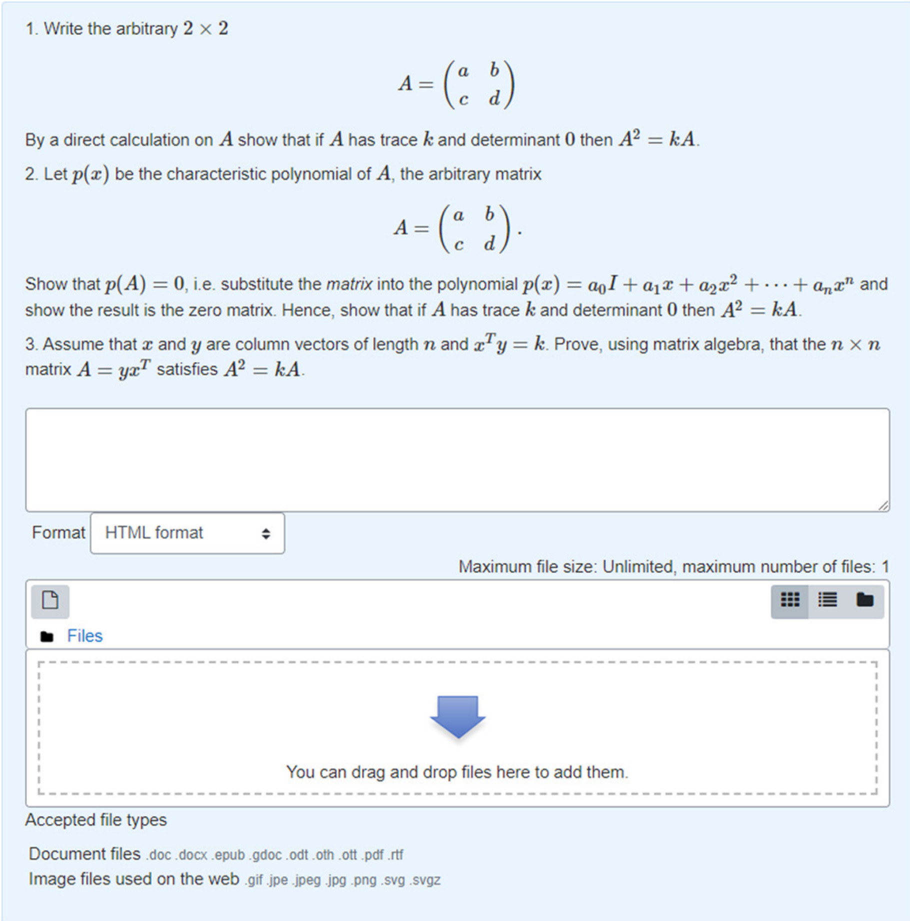

Fig. 2. Question 13 from ILA in 2020–21 showing human marked questions

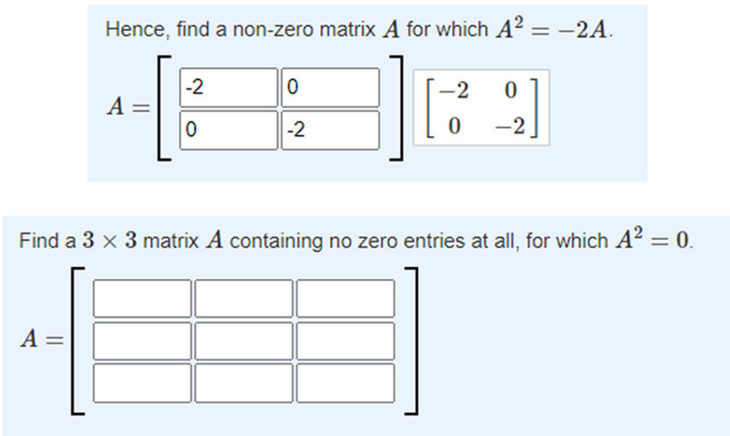

Fig. 3. Questions 14 and 15 from ILA in 2020–21 showing open-ended STACK questions

2.3 The 2021–22 assessment, December 2021

The 2021–22 online assessment contained only automatically assessed (STACK) questions. This test consisted of a mix of question types including calculations, conceptual multiple choice, give example questions (e.g. Figure 3) and proof-comprehension questions. Figure 4, from the December 2021 assessment, shows a "give example" style question using counter examples. This question is based on a known misconception, and 34% of the students answered "true" (which is incorrect). Only 40% of the cohort were able to supply a correct counter example. Indeed, students found such questions much more difficult than those requiring a predicable calculation.

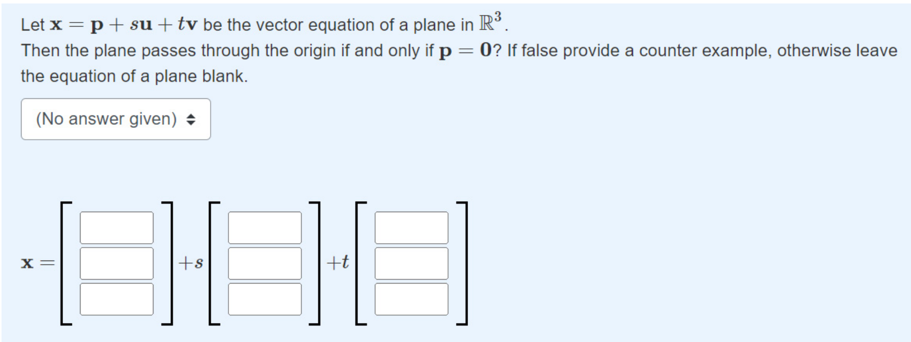

Fig. 4. Question 9 from ILA in 2021–22 showing counter-examples

The results in 2021 had a mean of 69%, and standard deviation 19%. This is a slightly higher mean than the (informal) target with a wider standard deviation. Re-balancing the assessment with one more proof-comprehension and correspondingly less computation would probably restore the achievement statistics more within departmental norms.

3 Discussion

Questions need to be designed for the format (STACK) and writing proof comprehension questions is something of an art, see [7] for a detailed discussion. On reflection, this is probably due to our lack of familiarity and experience with this style of assessment. Indeed, writing any assessment questions requires a knowledge of the subject matter, the student group, the aims of the course, together with a deep understanding of the tool used for assessment. The online format allows questions which would require significant computation from the examiner, and hence are not used in traditional examinations. While proof comprehension enables assessment of some learning objectives associated with mathematical proof, we cannot automatically assess freeform answers and this is clearly a significant drawback. The semi-automatic approach is a sensible compromise for better coverage of intended learning outcomes and so some human marking will be reinstated and retained for the foreseeable future.

Questions need to be carefully written, and additional work is needed to automate the STACK assessments. Unlike formative assessment, detailed feedback need not be automated *in advance* of any assessment. Indeed, a review of students' attempts can take place once the assessment is finished to develop partial credit, and double check the marking algorithms. Table 1 shows how much person-effort was needed to mark the assessments. The (*) in 2021–22 is for checking and awarding some partial credit for unanticipated responses. There is a considerable time saving for staff in running a STACK online exam, especially for larger courses such as ILA.

Table 1. Time taken to mark assessments

Year		No. Students	Days to Mark	Mins/Script
2019–20	Traditional exam	650	35	22
2020–21	STACK and human	718	22	13
2021–22	STACK only	814	2(*)	1

It has taken from December 2017 to December 2021 to gain experience and sufficient confidence to undertake a serious attempt at a fully automatically marked online examination. In our view, the fully automatically marked synoptic assessment in Dec 2021 was perfectly respectable in the sense that the information supplied by the test enabled us to decide whether students understood linear algebra in ways comparable to a traditional examination. This work demonstrates we could write fully online examinations for mathematics, both for linear algebra and more widely.

4 References

[1] Kinnear, G., Wood, A. K. & Gratwick, R. (2021). Designing and evaluating an online course to support transition to university mathematics. International Journal of Mathematical Education in Science and Technology, 0(0), 1–24. https://doi.org/10.1080/0020739X.2021.1962554

[2] Mejia-Ramos, J. P., Lew, K., de la Torre, J. & Weber, K. (2017). Developing and validating proof comprehension tests in undergraduate mathematics. Research in Mathematics Education, 19(2), 130–146. https://doi.org/10.1080/14794802.2017.1325776

[3] Nicholson, W. K. (2019). Linear algebra with applications. Lyryx.

[4] Poole, D. (2011). Linear algebra: A modern approach (Third). Brooks/Cole, Cengage learning.

[5] Sangwin, C. J. (2013). Computer aided assessment of mathematics. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199660353.001.0001

[6] Sangwin, C. J. (2019). Developing and evaluating an online linear algebra examination for university mathematics. Proceedings of CERME 11, (TWG 21: Assessment in Mathematics Education).

[7] Sangwin, C. J. & Bickerton, R. (2021). Practical online assessment of mathematical proof. International Journal of Mathematical Education in Science and Technology.

[8] Sangwin, C. J. & Kinnear, G. (2021). Coherently organised digital exercises and expositions. PRIMUS. https://doi.org/10.31219/osf.io/jhngw

5 Author

Christopher Sangwin joined the University of Edinburgh in 2015 as Professor of Technology Enhanced Science Education. His learning and teaching interests include educational technology and automatic assessment of mathematics using computer algebra.

Article submitted 2022-10-03. Resubmitted 2022-12-17. Final acceptance 2022-12-19. Final version published as submitted by the authors.