Konstantina Zerva (2018).
Developing STACK assessments in Edinburgh, 2017–2019.
In: Contributions to the 1st International STACK conference 2018.
Friedrich-Alexander-Universität Erlangen-Nürnberg: Fürth, Germany.
DOI: 10.5281/zenodo.2585816

Developing STACK assessments in Edinburgh, 2017–2019

Konstantina Zerva[1]

Abstract: In this paper I report the development and use of STACK assessments at the University of Edinburgh during the academic years 2017–2018 and 2018–2019, as part of a department-wide project to improve online assessment. The primary goal is to create high quality questions, with random versions and full worked solutions, for our major first and second year mathematical courses, some with over 600 students. The paper reports our first steps in creating a framework for systematic review and quality improvement. I also report on my experiences in developing questions over the last eighteen months, providing examples and advice for anyone else working in this area.

Keywords: Online assessment; STACK

1 Background

Assessment is an important component of the teaching process and it has a significant influence on students' learning experience of higher education. Indeed, *"assessment sits at the heart of the learning process"* [Ti16]. But assessment is more than simply allocating a numerical mark to each individual student; its main purpose is to provide valuable information to students, helping them progress and succeed, an idea that is known as "assessment for learning" [Bl03]. In our era, digital technology plays an important role in every aspect of our life, and of course it has a significant impact in education as well.

In this paper I describe the development of mathematical questions to support university courses using the STACK online assessment system. My role was to support academic staff in the School of Mathematics at the the University of Edinburgh to improve and further develop the online assessment available to students in traditional university courses. Automatic assessment is commonly associated with multiple choice questions (MCQ). While MCQs certainly do have a place among the mix of assessment types, for mathematics MCQs are particularly problematic as the relative difficulty of a reversible process, e.g. integration compared to differentiation, is markedly altered in different directions [SJ17]. Motivated by the need to assess answers to students' work, the STACK online assessment system uses computer algebra to support the assessment process [Sa13]. STACK uses computer algebra to generate random questions; accept and validate mathematical answers; establish the mathematical properties of those answers and generate feedback.

[1] University of Edinburgh, UK, k.zerva@ed.ac.uk

2 Prior online assessment in Edinburgh

STACK was first used in Edinburgh prior the beginning of the academic year 2016–17, for the Lothian Equal Access Programme for Schools (LEAPS) summer school. LEAPS is an outreach programme aimed at raising awareness of the opportunities in higher education for students yet to arrive at university. We used STACK for the summer school assessments because we wanted to do a small test of the new STACK server before the beginning of semester 1. During the academic year 2016–2017 we provided weekly reading quizzes for Introduction to Linear Algebra (ILA), a large first year course described in more detail below.

During this academic year we used a variety of assessment tools, including STACK, the commercial system Maple TA and also materials provided by publishers attached to commercial textbooks for various courses. My role was to bring our assessment in-house, consolidating all assessment materials in a single system (STACK), which is linked to the University virtual learning environment (Blackboard). During the academic years 2017–18 and 2018–19 have seen a significant expansion in the use of CAA. In particular, the majority of our year 1 courses are now supported by online assessment, both within the School of Mathematics and more widely in the College of Science and Engineering. We also no longer rely on externally hosted commercial systems.

3 Use at the School of Mathematics during the academic years 2017–2018, and 2018–19

In this section I record the way STACK has been used in assessment of first and second year mathematics courses at the University of Edinburgh, during the academic years 2017-18 and 2018-19. For each of these courses I have developed and expanded the range of STACK questions available. The courses listed here are all worth “20 credits”, and since students take 120 credits per year, each course is approximately 1/3 of their work in an individual semester. Where a percentage is listed after the quiz type, e.g. (5%) this indicates the contribution of this set of quizzes to the overall grade for the course.

Introduction to Linear Algebra (ILA): Year 1, semester 1, taken by over 600 students. Compulsory for Mathematics, Informatics and Mathematical Physics degrees, and available to others.

1. **Pre-lecture quizzes** (5%). 2 per week, 25 min, 1 attempt.
2. **Weekly assessment** (5%). 1 per week, 1 hour, 1 attempt. These replaced some questions on the written weekly work, saving an estimated 65 hours per week of human marking.
3. **Mock exam**: In December 2017 and 2018 we implemented an optional mock online examination. A discussion of this is given in [Sa18].

Calculus and its Applications (CAP): Year 1, semester 2, taken by over 550 students. CAP is required for the same group of students as ILA, with some optional students.

1. **Pre-lecture quizzes** (5%). 1 per week, 30 min, 1 attempt.
2. **Skill quizzes** (5%). 1 per week, no time limit, 3 attempts, use the highest mark.
3. **Mock exam**: In April 2018 we implemented a mock online examination.

Proofs and Problem Solving (PPS): Year 1, semester 2, taken by over 280 students. PPS is required for mathematics students, and is an introduction to mathematical proof.

1. **Pre-lecture quizzes** (5%). 1 per week, 25 min, 1 attempt.

Mathematics for the Natural Sciences (MNS) 1a/1b
Engineering Mathematics (EM) 1a/1b: Year 1, semesters 1 and 2. MNS is a first year course for students of Chemistry and related disciplines (150 students). EM is a first year course for Engineering (300 students). These courses are closely related, with many shared STACK quizzes.

1. **Practice quizzes** (0%). 3 per week, unlimited attempts, unlimited time, score 60% to unlock the next quiz (practice then assessed).
2. **Assessed quiz** (5%). 1 per week, no time limit, 3 attempts, use the highest mark.

Several Variable Calculus and Differential Equations (SVCDE): Year 2, semester 1, taken by over 200 students. Required for Mathematics and Mathematical Physics students.

1. **Practice quizzes** (0%). 1 per week, unlimited attempts, unlimited time, score 50% to unlock the assessed quiz.
2. **Assessed quiz** (15%). 1 per week, no time limit, 1 attempt.

Mathematics for Physics 1: Year 1, semester 1, taken by around 200 students. Required for Physics students.

Fundamentals of Algebra and Calculus (FAC): Year 1, semester 1, taken by 110 students. FAC is our first attempt at a completely online course. This is a major initiative, involving over 50 STACK quizzes, and it is reported elsewhere see [Ki18].

The courses: ILA, CAP, PPS, EM, MNS, SCVDE, FAC and Mathematics for Physics represent a substantial range of courses with large numbers of students. We have used a variety of quiz policies, and a mix of formative and summative assessments. The pre-lecture quizzes tend to be short, with 2-5 questions and a general survey encouraging students to explain what they found difficult in the pre-lecture reading. The assessed quizzed tend to be longer, with 6-12 questions and are designed with a summative purpose. In section 5 I describe how we review the effectiveness of these online assessments.

4 Reflections on question authoring

In this section I will provide some reflections and comments about the process of authoring questions. The purpose of doing this is to provide realistic expectations for anyone planning a substantial project in developing online assessment materials, either as the question author or managing the process.

To begin with, there are many things that somebody needs to consider before starting writing any questions. It is important to have clear statements of the course content and a clear conception of the purposes of assessments during the course. For instance, are the assessments formative, or will these also have a strong summative component? Does the question author write the questions? Set specific goals for numbers of questions needed and clear deadlines of when they need to be ready. Also, think about the type of questions and whether a mix of questions types is necessary, e.g. see the taxonomy of [Sm96].

Of course, much of this (i.e. knowing what I should be doing) has little, if anything to do with online assessment, rather it comes down to professionalism in the design of high-quality courses for students to study. What is specific to CAA is the necessity for a longer time to develop and review questions before they are released to students. It is essential to have materials ready for technical authoring well in advance. This presents a particular challenge for course organizers who are, perhaps, unaccustomed to working in teams. The danger is that the course organizer will articulate their requirements in rather vague terms, "I need something on Green's Theorem for next week". Even when a course organizer provides a question, with a detailed, written solution, it is usually necessary for the question author to have a good understanding of the mathematics, especially when this is part of the randomization.

It is actually very difficult to quantify the workload involved in writing questions. It can take a day to write a particularly complex question, or as little as 20 minutes to modify an existing template. The whole endeavour is therefore rather difficult to quantify precisely. As a rule of thumb, given questions with solutions, it is not unreasonable to average six questions in a working day. However, learning the materials, writing the questions, writing solutions, worrying about randomisation take additional time.

Additionally, STACK is a complex, technical system, one undergoing improvements and changes. In MNS1a, for the first time, we were confronted with the challenge of writing simple statistical questions for non-specialists. The problem was that many questions require a floating point number as an answer and students (unwittingly) did not provide enough information for us to reach a reliable judgment on whether they understood the statistics. As a result of these difficulties, we developed a new "numerical" input type for STACK. This input type has options to specify the expected level of numerical accuracy at the *validation stage*. For example students can be required to provide a minimum number of significant figures of accuracy in their answer, and expressions will be rejected as invalid otherwise. Moving this check into the validation has significantly improved students' experience for

this kind of assessment, and made writing reliable questions much easier. While such developments do take time to make their way onto production servers, this improvement exemplifies the requirements driven design and development of STACK itself.

Maxima has some serious limitations, and we have yet to resolve all of these. For example dealing with complex numbers $a + ib$, there are conventions in mathematics that Maxima does not always respect. Take, for example $a + ib$. The default Maxima behaviour is to re-write this as $ib + a$, respecting the default order on a and b. Negative numbers also affect the display. On some occasions we would like a complex number to be considered as a single entity, on others as the sum of real and imaginary parts. Converting between different forms, especially when a complex number appears within a large expression can be problematic. Indeed, wherever there are multiple equivalent representations of mathematical expressions, including roots in the denominator or roots in general, there can be subtle difficulties in getting the precise display form required.

Organising and finding questions will become a problem in the future. Large question banks are difficult to browse, and have no index. Also, questions end up duplicated between courses. This makes longer term maintenance problematic. While this is not, strictly speaking, a STACK issue it does impinge on teachers and questions authors' experience of using the system.

5 Reflections on reviewing our assessments

One of the disadvantages of online assessment is the lack of a feedback loop for the teacher. In particular, there is no direct incentive for them to review the marked work, and since they no longer mark work by hand, active steps need to be taken to review what students have done and take action as a result. In this section I will provide some comments about the process of reviewing questions.

Before starting to review the quizzes it is important to consider the following:

1. How many quizzes do the students have every week?
2. What is the purpose of each of them?
3. What quiz policies should we use? E.g. for how long should the quiz be open, how many attempts should the student get, and do the marks contribute to the overall course grades?
4. How much time would you like the students to spend dealing with each quiz?

There are many possibilities and to create a coherent, sensible course and the course organizers have to choose from a sometimes bewildering array of options.

It is very important, when a quiz first opens, to check the early attempts for unnatural patters (a question that gives zero marks to everybody). It could be an indication that there is

something wrong with this question. The best way to deal with errors in a live quiz is to fix the question and then regrade everybody. Avoid to override marking on individual attempts.

After having a significant amount of attempts it is worth looking whether the quiz, as a whole entity, behaves in an appropriate way.

- Does the histogram of marks match the target histogram?
- Does the histogram of time taken match the target histogram?

General distribution of marks:
It is quick and easy to inspect the **general distribution of marks** [2]. For formative assessment, where mastery is the goal, we expect the majority of the students to achieve full marks (see Figure 1, top left). For summative assessment, where the students are allowed only 2-3 attempt, we expect a more evenly shaped distribution of marks (see Figure 1, bottom left). If the students have only one attempt we expect a bell shaped distribution (see Figure 1, right).

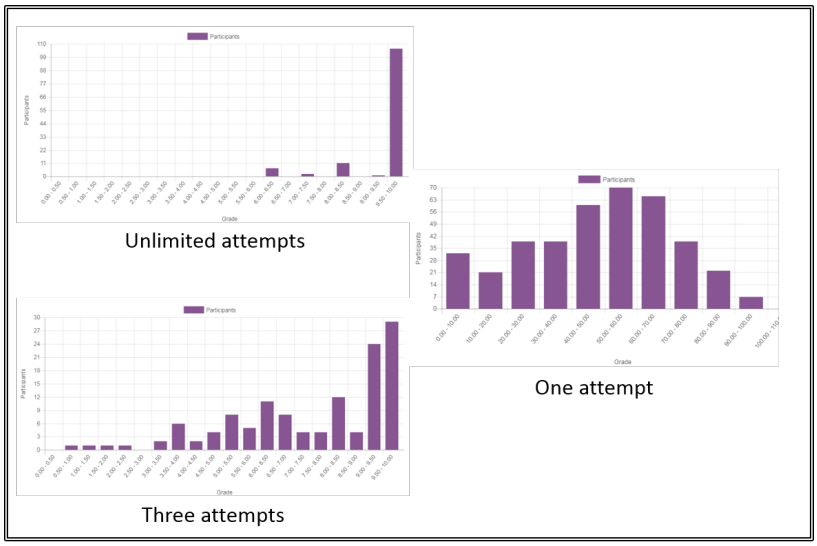

Fig. 1: Mark distribution for various number of attempts.

Time distribution:
It is essential to have a clear idea of how much time the students need to spend in each quiz. Even in the case that we do not have time limits, we need to consider the time that each quiz takes. An average time of 10 min may indicates that the quiz is too easy or it doesn't have a lot of questions. An average time of more than 1.5 hours may indicates that the quiz is too difficult. There are no strict guidelines about the time; it will vary depending on the purpose of each quiz (see Figure 2).

[2] https://docs.moodle.org/34/en/Quiz_grades_report

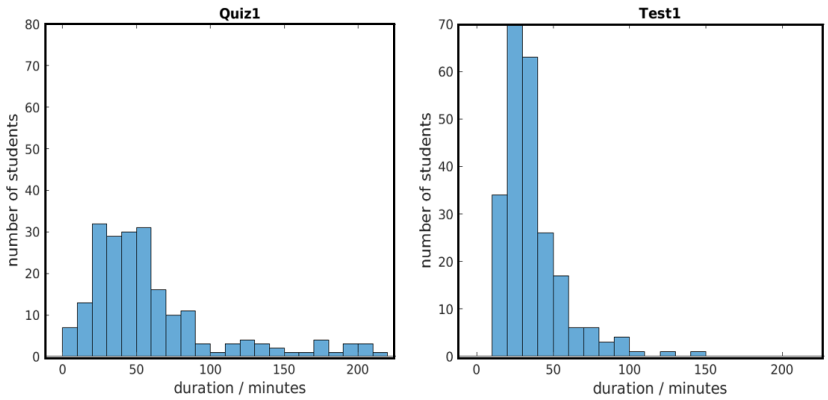

Fig. 2: Time distribution of a formative quiz (left) and a summative quiz (right).

Of course Moodle quizzes offer a whole range of statistics, for the whole quiz as an entity, and for each individual question (look the documentation site `https://docs.moodle.org/dev/Quiz_report_statistics`) so it is possible to achieve a more sophisticated review.

Engaging course organizers in reviewing the students' work is also sometimes a challenge. Some course organizers are very engaged, reviewing the time taken, success rates, marks distributions and so on. Other course organizers are less active.

6 Conclusions/What next?

This major two-year project is now eighteen months completed. At this stage most of the year 1 courses have new, and expanded, online assessment support provided by STACK. Students on these courses have a consistent online assessment experience (compared to the variety of systems we used to use), and are asked to complete more work online. Our students are therefore doing more regular, more extensive work on mathematical tasks with significantly less ongoing staff effort expended marking the work. For the large courses where we are now using STACK, the cost saving in weekly staff time is significant, but with the one-off capital expense of developing the questions themselves.

This project has been relatively conservative in providing online quizzes for traditional on-campus university courses. We have worked with course organizers, and supplemented or replaced traditional exercises without fundamentally seeking to change pedagogy or teaching practice. We have not entirely replaced hand-in work, but rather we are using staff time to assess proof and extended arguments and are using STACK to assess more methods-based tasks. This is a pragmatic division, which still requires some ongoing human marking. We have experimented with mock examinations using STACK in ILA (see [Sa18]) but we have not yet used STACK for a large examination. Through this experience we have developed

confidence and competence with online assessment. More ambitious projects, such as that reported in [Ki18], require very close cooperation between learning technologists such as myself, and course organizers. Such projects also require the effort to be recognized by management, either through support for posts such as mine, or in recognizing the effort needed from academic staff.

However we are now faced with a law of diminishing returns. We have completed work on all our year 1 and year 2 calculus and other methods-based courses. These courses have many assessments which are relatively easy to automate with STACK. Assessments in group theory, real analysis and other more advanced courses will be a lot harder to automate with STACK because of the need for assessment of proof. We are making a start with these courses, and expanding into statistics and computer programming, particularly combining STACK assessments with CodeRunner [LH16] (using Python, MATLAB and R). We are also investigating expanding the range of assessment types to make use of the match question type pmatch, [Jo15]. The in-built question types together with STACK and other specialist tools are now able to provide rich and engaging assessment experience for students in an expanding range of subject areas.

References

[Bl03] Black, P.; Harrison, C.; Lee, C.; Marshall, B.; William, D.: Assessment for Learning: putting it into practice. Open University Press, Maidenhead, U. K., 2003.

[Jo15] Jordan, B.: Qualification Reform: A guide to what is happening around the UK. Technical report, Universities and Colleges Admissions Service, 2015.

[Ki18] Kinnear, George: Delivering an online course using STACK. In: Proceedings of the STACK Conference. 2018.

[LH16] Lobb, R.; Harlow, J.: Coderunner: a tool for assessing computer programming skills. ACM Inroads, 7(1):47–51, March 2016.

[Sa13] Sangwin, C. J.: Computer Aided Assessment of Mathematics. Oxford University Press, Oxford, UK, 2013.

[Sa18] Sangwin, C. J.: High stakes automatic assessments: developing an online linear algebra examination. In: Proceedings of 11th Conference on Intelligent Computer Mathematics. Hagenberg, Austria, 2018.

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

[Sm96] Smith, G.; Wood, L.; Coupland, M.; Stephenson, B.: Constructing mathematical examinations to assess a range of knowledge and skills. International Journal of Mathematics Education in Science and Technology, 27(1):65–77, 1996.

[Ti16] Timmis, S.; Broadfoot, P.; R., Sutherland; Oldfield, A.: Rethinking assessment in a digital age: opportunities, challenges and risks. British Educational Research Journal, 42(3):454–476, June 2016.