Informatics in Education, 2018, Vol. 17, No. 2, 363–380

DOI: 10.15388/infedu.2018.19

Investigating Bloom's Learning for Mastery in Mathematics with Online Assessment

Timo PELKOLA[1], Antti RASILA[2], Christopher SANGWIN[3]

[1]*Department of Mathematics and Systems Analysis, Aalto University*
P.O. Box 11100, FI-00076, Finland
[2]*Guangdong Technion – Israel Institute of Technology*
241 Daxue Road, 515063 Shantou, Guangdong, China
[3]*School of Mathematics, University of Edinburgh*
The King's Buildings Edinburgh, EH9 3FD, United Kingdom
e-mail: timo.pelkola@gmail.com, antti.rasila@iki.fi, c.j.sangwin@ed.ac.uk

Received: May 2018

Abstract. In this paper we report a study in which we have developed a teaching cycle based closely on Bloom's Learning for Mastery (LFM). The teaching cycle ameliorates some of the practical problems with LFM by making use of the STACK online assessment system to provide automated assessment and feedback to students. We report a clinical trial of this teaching cycle with groups of university level engineering students. Our results are modest, but positive: performance on the exercises predicted mastery according to the formative tests to a small extent. Students also report being supportive of the use of the new teaching cycle.

Keywords: learning for mastery; online assessment; mathematics education.

1. Introduction

This research is motivated by the remarkable observation of (Bloom, 1984) that students taught by an individual tutor achieve test scores which are two standard deviations better than students who attend traditional classroom teaching. Learning for Mastery (LFM) is an educational philosophy proposed by Bloom as a partial solution to the problem of finding resources for individual tutorials. However, Learning for Mastery also has practical problems. Current automatic computer aided assessment (CAA) of mathematics has reached a level of sophistication which suggests some of the practical problems with LFM might be overcome, and this is what we set out to investigate. Can the practical problems traditionally associated with implementing Bloom's Learning for Mastery be overcome effectively with online CAA in mathematics? In this paper we report a study in which we have developed a teaching cycle based closely on Bloom's LFM, making

use of online assessment. We report a study to investigate whether we see any significant learning gains using CAA and our LFM approach.

In Section 2 we provide a theoretical background to LFM and discuss contemporary CAA of mathematics in more detail. Our precise research questions are given in Section 3. Section 4 provides details of the methodology undertaken to address our research questions. Results in Section 5 precede the final discussion.

2. Background

2.1. *Mathematics for University Engineers*

All university engineering students learn mathematics as a core part of their undergraduate education. Engineering mathematics curricula have been well-developed as an ongoing international collaboration, see (Barry and Steele, 1992, Mustoe and Lawson, 2002, Alpers, 2013). The resulting framework includes content and concepts, but goes well beyond this to include competencies. Indeed, (Alpers, 2013) opens the executive summary of the most recent framework document by arguing that "*the main message of this new edition is that although content remains important, knowledge should be embedded in a broader view of mathematical competencies.*" The phrase "*mathematical competencies*" means that a student has proficiency in a set of interrelated mathematical skills. The previous work of (Kilpatrick *et al.*, 2001; p. 116), for example, identified conceptual understanding, procedural fluency, strategic competence, adaptive reasoning and productive disposition as five important strands.

2.2. *Mastery Skills*

We separate mathematical skills (loosely) into two groups: mastery and problem solving skills (for related discussion, see (Burkhardt and Swan, 2007) and (Rasila *et al.*, 2015)). The essential distinction is that mastery skills are rarely the end goal, rather they form part of a subsequent wider task. These skills form a loose hierarchy: weak basic conceptual and procedural skills seriously hinder a student's ability to formulate and solve mathematical problems. (Skemp, 1971), for example, framed the discussion of this issue in terms of a *schema*: "*inappropriate early schemas will make the assimilation of later ideas much more difficult, perhaps impossible*", (Skemp, 1971; p. 51). Note that mastery skills are framed within a particular context and the goals of instruction.

Mastery skills are emphatically *not* confined to the lower order tasks, such as recall of knowledge. Mathematics is highly structured: as a specific example, writing a rational expression using partial fractions require students to look ahead to anticipate the consequences of their choices. Symbolic integration, in turn, relies on choosing particular algebraic forms, including re-writing rational terms as partial fractions. In this

context, multi-step partial fractions and symbolic integration techniques are mastery skills precisely because successful implementation of these skills are not the end point for engineers.

We also include basic deductive reasoning as a mastery skill, at least to the extent that the student should understand the role of assumptions, conclusion, particular/universal statements, etc. Without these it is impossible to create even modest chains of reasoning needed to apply more complex methods and procedures, typically taught to engineering students. Furthermore, the distinction between "reasoning" and "computation" is not entirely clear. Indeed, Boole's programme was to transform some forms of logical reasoning into a computation, precisely to help mathematicians gain mastery of this notoriously difficult topic, (Inglis and Attridge, 2017).

We should also delineate via examples what is not a mastery skill. Problem solving skills are often applicable more widely, and are affective in nature (e.g. resilience) rather than framed in terms of specific knowledge schemas. Problem solving skills can often only be evaluated in terms of qualitative better-worse judgements, rather than right-wrong absolute judgements. There is a substantial body of work on the learning of teaching of problem solving skills, from the reflective work of (Polya, 1962), through the empirical studies such as the work of (Schoenfeld, 1985) and to more specialist contemporary discussion, such as pedagogy for engineers (Michalewicz and Michalewicz, 2008). Since effective problem solving is normally considered to be an important part of the end goal, we do not include these skills within mastery skills. Similarly, skills which do not form part of subsequent wider tasks are also not included within mastery skills. Depending on the goals of the course, mastery skills may include both pen/paper calculations and the use of tools like CAS or even programming environments like MATLAB.

2.3. *Teaching, Assessment and Learning for Mastery*

Different areas of mathematical proficiency require different learning strategies, e.g. conceptual and procedural abilities are typically learned though conscious practice of exercises. Assessments, particularly high-stakes examinations, are often cited as important drivers of students' learning by providing strong extrinsic motivation. We acknowledge that high-stakes school examinations have been criticised for privileging procedural items over conceptual e.g. (Iannone and Simpson, 2012; Noyes *et al.*, 2011). At universities (Tallman *et al.*, 2016) found that little had changed in the last twenty five years: the majority of items required students to recall and apply a rehearsed procedure and few required conceptual understanding or problem solving. This emphasis on procedural items is partly explained by the ease with which they can be produced and scored (Swan and Burkhardt, 2012), indeed compared to other subjects scoring reliabilities tend to be high in mathematics (Brooks, 2004). For further discussions of mathematical tasks see (Smith *et al.*, 1996), (Pointon and Sangwin, 2003), (Watson and Ohtani, 2015) and (Foster, 2013).

The review of (Bloom, 1984) considered research which compared different forms of teaching. (Bloom, 1984) reports that individual tutoring resulted in student achieve-

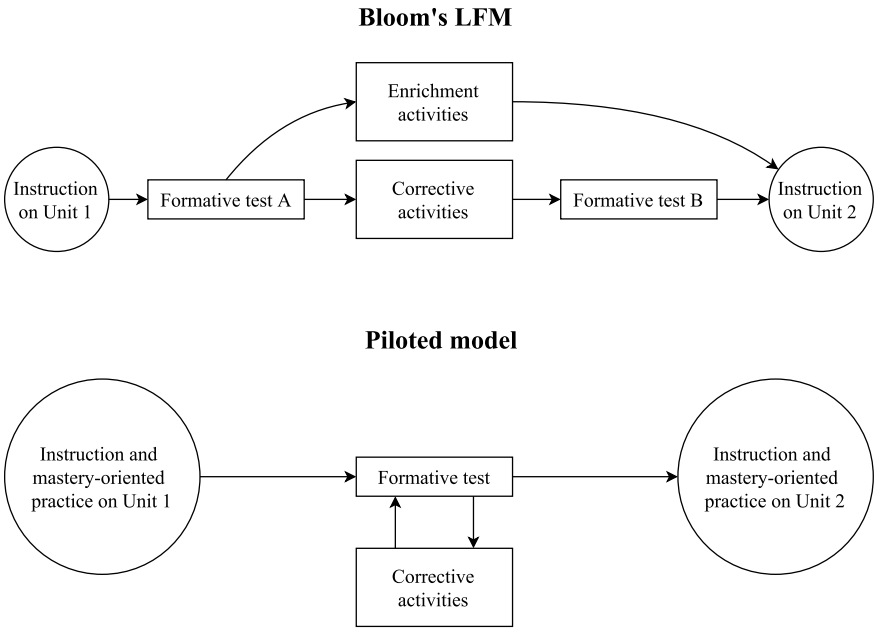

Fig. 1. A comparison of Bloom's Learning for Mastery (LFM) cycle and our model.

ment which is two standard deviations better than that of students who attend traditional classroom teaching. To close this gap (Bloom, 1984) devised and evaluated a teaching intervention called Learning for Mastery (LFM). In LFM students are regularly tested by using formative tests and students are required to demonstrate a correct answer to 90% of the test problems, i.e. demonstrate "mastery". When a student falls short of mastery further teaching and testing is repeated, several times if necessary. Bloom's Learning for Mastery has been well-studied, with a number of independent studies reporting significant positive effects, e.g. (Anderson *et al.*, 1995), and see (Hattie, 2012) for a review.

One of the practical impediments to LFM is the difficulty faced by the teacher who has to orchestrate the work of many students who are potentially all at different stages. They also potentially have to devise different but related formative tests. In traditional settings such extensive testing is still impractical. Certainly in typical university entry-level mathematics courses, with hundreds of students, this will be impossible. Online assessment has the potential to remove this practical barrier. However, mastery learning can lead into surface-oriented learning strategies, especially if formative testing is mainly based on multiple choice questions. Our interest in this topic arose because of the potential we see with contemporary online assessment in mathematics.

The current research is based on experiences gained in previous projects, such as (Rasila *et al.*, 2010) where we started to work with the online learning system STACK as a tool for learning basic calculation techniques for engineering students, and (Majander and Rasila, 2011) where we tried to use formative assessment (much in the sense of Bloom) to improve students' motivation to participate in the course activities. However,

this previous work lacked the corrective measures associated with Mastery Learning, which we report here. Besides improving learning outcomes, we are also interested in finding objective assessment methods suitable for distributed and distance learning (cf. (Rasila and Malinen, 2016)).

2.4. *Online Assessment for Mathematics*

Computer aided assessment is well established and widely used to support the teaching and learning of mathematics. There is over a quarter of a century of experience developing automatic online assessment of mathematics: an early review is given by (Beevers *et al.*, 1991) and a more recent review in (Sangwin, 2013).

The current technical state of the art in online assessment of mathematics focuses on accepting a final answer from students and automatically establishing mathematical properties. This goes well beyond relying on multiple choice (and similar question types) which have the well-known difficulties discussed by, e.g. (Sangwin and Jones, 2017). For example, if a student enters an algebraic expression the teacher will have specified in advance that the computer should seek to establish algebraic equivalence with the correct answer. They may also additionally, and separately, seek to establish that it is written in a particular algebraic form, such as factored. Normally, there are a variety of correct answers, e.g. $(x - 1)(x + 2)$ or $(x + 2)(x - 1)$ could be acceptable. Here $(x + 1)(x - 2)$ is not equivalent to the correct answer, but is in factored form, and it is typical of systems to create specific feedback for students.

The following features are now typical in many, if not most, mathematical systems.

- Questions are randomly generated in a structured way using computer algebra systems (CAS). Normally the question and steps in a fully worked solution are reverse engineered from the teacher's answer. Quiz management components can also randomly select from a question bank to create an activity for an individual student.
- Students provide the final answer in the form of a mathematical expression, e.g. an equation, rather than responding to multiple choice questions. It is not yet typical to automatically assess a complete argument or proof.
- Objective mathematical properties of answers are automatically established, e.g. algebraic equivalence with a correct answer.
- Outcomes are automatically generated (including feedback) which fulfil the purposes of formative and summative assessment.
- Data on all attempts at one question, or by one student, are stored for later analysis.

The ability to randomly generate similar questions is particularly important for mastery learning. Previous experience suggests the high value to students of the corresponding worked solutions, which provide a model from which students can base their answer to subsequent similar versions.

Many example systems provide the features we have outlines above. This project made used of the STACK online assessment system described in Section 4.3. STACK

is based on a computer algebra system, as is a commercial alternative MapleTA https://www.maplesoft.com/products/Mapleta/ (retrieved Jan 2018). The NUMBAS project https://www.numbas.org.uk/ (retrieved Jan 2018) aimed for light-weight portable code, and does not call a third party computer algebra system. There are important differences between the systems, e.g. STACK consciously separates out "validity" from "correctness". Feedback on validity is always provided by STACK, whereas correctness might be immediately assessed in a formative setting but delayed during an online examination. In some questions floating point numbers might be forbidden as approximations, in other questions it may be impossible to establish if a student has answered correctly if they provide too few significant digits. Information on the context helps students understand what type of answer is expected and this has been found to significantly reduce the extent to which students are penalized on a technicality. Many other systems have a single feedback mechanism, combining information on context validity with the overall assessment. At the current stage of development each project has its particular strengths, and particular features. WebWork http://webwork.maa.org/ (retrieved Jan 2018), for example, has a large question bank of tested materials. Most systems have the features listed above in common.

While these systems do not (yet) fully assess complete solutions provided by students, we are aware of a number of parallel developments to implement checking of "line by line" working in many procedural situations. STACK has this feature for algebraic arguments, as do other software such as the SOWISO project https://calculus.sowiso.nl/ (retrieved Jan 2018). In the near future checking of line by line reasoning, and simple logic, is likely to also become standard.

3. Research Questions

In this paper we report an action research study to investigate the following research questions:

1. To what extent is STACK suitable for implementing Learning for Mastery?
2. Can mastery be predicted from the STACK exercise data for formative tests?

Lastly, we are interested in how students react to the STACK online tests used in our learning model.

4. Methodology

4.1. *Adapting Mastery Learning for an Online Environment*

LFM suggests pairing formative assessment with appropriate correctives and we are interested in whether the traditional practical problems with implementing LFM can be overcome effectively with online assessment. In pursuing our investigation of LFM in

an online context we have adopted a *Design Research* paradigm, aligned with the criteria of (Collective, 2003).

- The learning environments and the developing theories are connected and intertwined.
- Research and development take place through iterative and continuous cycles of design, enactment, analysis, and redesign.
- Research on design should yield to sharable theories that can facilitate communication between the practitioners and educational designers about possible implications.
- Research must account for how designs function in authentic settings.
- The development of such accounts relies on methods that can document and conceptualize processes of enactment to outcomes of interest.

Indeed, putting these characteristics in the context of our study we have combined a theoretical idea of (Bloom, 1984) with the design of contemporary online learning environments, taking advantage of automatic feedback correctives required by the theory but only recently available in a practical setting. Our study takes place in a real-world setting. Indeed, unlike (Bellhäuser *et al.*, 2016) who reported randomised control trials, we implemented our LFM scheme in a mainstream core course.

Bloom's Learning for Mastery model was adapted in our study using weekly online exercises and formative tests to assess mastery in core skills. As a result, the methodology in this study differed in some ways from the original LFM model. In LFM, mastery is assessed only with formative tests, which usually come in the form of invigilated multiple-choice questionnaires with different versions for reattempts. In Bloom's original implementation this was limited to two attempts. In this study mastery was assessed with online exercises. The same formative test was used for each attempt, with the possibility of a random versions of the quiz generated for each attempt. The formative test was given the name "practice exam" during the course, since this term was more familiar to the students.

The learning units were slightly extended to readjust the workload from the formative tests. Also, some of the higher-order learning objectives in the course were not covered by the formative tests or online exercises, as automatic assessment of these are difficult without, in our view, fatally compromising the test validity. Since the online component of the course covered mostly procedural skills, a new "guided discovery" type of project work was introduced for the exercise sessions to provide students with a balance of assessments during their course. This consisted of four paper-based assignments and a final report about the mathematics of harmonic oscillation.

Our current study used the automatic feedback generated by STACK questions as the primary corrective. The formative test items were also paired with thirdparty videos of similar worked examples, which were made available after the first submission of the test. Indeed, (Hodges and Murphy, 2009) found that vicarious experience was one of the most important factors, and these videos provide some vicarious experience in an online environment that might be provided in person during traditional lecturing. We also believe that students who had already gained mastery would gain some benefit from taking the formative test anyway. Lastly, we note that in the absence of a control group and proper pre/post-tests, the effectiveness of mastery learning itself was not considered in this study.

4.2. *Courses Selected for the Study*

Calculus I (MS-A010x) is a six-week (5 ECTS credits) compulsory course for science and engineering students at Aalto University covering single variable differential and integral calculus and ordinary differential equations. The course is offered separately for each degree programme, but with similar content. MS-A0106 (for student majoring in mechanical and construction engineering) and MS-A0107 (for students majoring in chemical engineering) were selected for this study, which took place as part of the continuing Aalto Online Learning (A!OLE) strategic development project coordinated at the Aalto University School of Science. The courses consisted of four hours of lectures and exercise sessions per week, weekly online exercises, paper-based assignments, formative tests at the end of learning units and a paper-based final exam. The course was divided into two three-week learning units, with the first unit covering limits, series and differential calculus and the second unit integral calculus and ODEs.

The course content included analysis of sequences and series, approximation of functions by series. Students were expect to be able to differentiate and integrate basic functions, and use these techniques in simple applications. The course included first order linear and separable differential equation, and second order linear differential equations with constant coefficients. Specifically, the course used (Adams and Essex, 2013) as the textbook and the defacto syllabus including material from chapters 1,2, and 3–12 inclusive.

4.3. *Online Assessment of Mathematics with STACK*

Our study adopted the STACK online assessment system. STACK has sustained development and use for over a decade with significant contributions of code from Aalto University Finland (see (Sangwin, 2013; Chapter 8) and, for very recent work (Harjula *et al.*, 2017)), the United Kingdom Open University and latterly the University of Edinburgh in Scotland. STACK was originally developed for Moodle but has been ported to ILIAS (see `http://www.ilias.de`, retrieved Jan 2018) and is used in other systems, including Blackboard, through the LTI protocol. See `https://stack.maths.ed.ac.uk/demo` (retrieved Jan 2018). STACK was developed by the last author, and the experimental study reported in this paper was undertaken by the first two authors at an independent institution. The key features of STACK include its mathematical sophistication, and the full authoring interface which aims to give teachers a wide range of options in a way which still makes writing learning materials practical.

STACK is used reliably with thousands of users on over 700 registered Moodle sites. For example, at the United Kingdom Open University during the academic year 2015–2016, students attempted over 880,000 questions on seven modules. The STACK question type accounted for approximately 15% of all questions used, and is second only to multiple choice in popularity (at 35% of all questions). There are a number of large international projects such as the Abacus `https://abacus.aalto.fi/` (retrieved Jan 2018) multi-lingual material bank which makes use of STACK, (Rasila,

2016). Other projects include (Barbas and Schramm, 2016), (Mäkelä *et al.*, 2016) and (Paiva *et al.*, 2015), and publishers are increasingly supplementing textbooks with online assessments such as (Coletta, 2010) which has 600 online homework problems written with STACK.

4.4. *Description of the Procedure*

New STACK questions were developed for the formative tests and weekly online exercises and the same set of questions were used on both courses. The mastery threshold was set to a minimum of 75–80% of the points available in the weekly exercise or formative test. Neither the online exercises or formative tests were strictly compulsory, but in line with much teaching in mathematics, both contributed a small proportion of the final course grade. Including reattempts, only points above each mastery threshold were awarded, and we refer to this as the *mastery bonus point scheme*. We intended and anticipated that all students would achieve mastery, and so these points will be above 80% and will contribute a small proportion of the final course grade. As a result, these points had a minimal effect differentiating course grading, and should be considered primarily as formative assessment.

The online exercises and formative tests had slightly different functions and were setup accordingly. While both gave feedback on the progress of a student's learning, the online exercises were meant for initial practice, while the formative test was to ensure that mastery in those skills had actually been gained and retained. In both cases questions could be reattempted an unlimited number of times without penalties, but in exercise questions the feedback was immediate, whereas in the formative tests it was deferred until submission of the entire test.

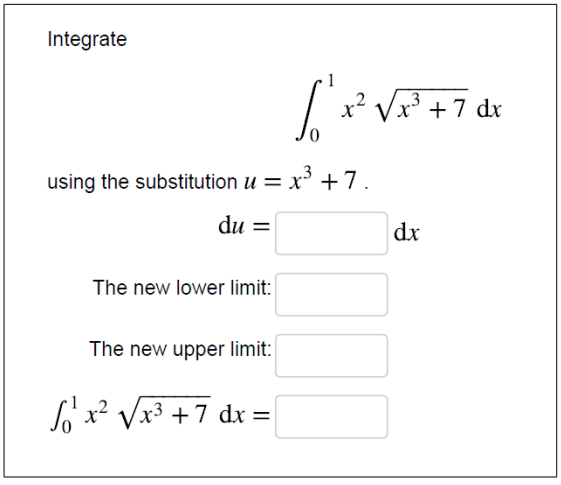

Fig. 2. An example of a STACK question used on the course.

The weekly online exercises were due on a Sunday, and consisted of five questions related to the lectures of the week. The formative tests could be taken at the end of the learning unit before the next lecture or course exam, and the use of calculators, textbooks or other accessories were discouraged although not controlled. The first formative test included four and the second five items.

5. Results

95 of 134 enrolled students in MS-A0106 and 118 of 198 enrolled students in MS-A0107 consented to the use of their data for this study. Of these students, 176 in the first learning unit and 168 in the second had opened all the weekly quizzes and the formative test at least once, which was counted as an attempt. These were used for predictive modeling.

Individual STACK item scores and numbers of attempts were extracted from the Moodle learning management system using a purpose-made export tool. This data was then imported to R for analysis. Both the data from STACK and course feedback was used to determine the suitability of STACK for implementing mastery learning. We implemented predictive modelling with various different classification methods and pre-processing with the help of the caret R package. Performance was evaluated with ten-fold cross validation with three repeats. Similar results were achieved with many of the methods. The results from logistic regression ('glm' in caret) are presented here.

Mastery was defined as achieving a score of 4 out of 5 (80%) or 3 out of 4 (75%) on a weekly quiz or formative test. *Initial mastery* denotes the percentage of students who had achieved mastery on the first attempt, and *eventual mastery* those who achieved

Table 1
Percentages of students who had gained mastery

	W1	W2	W3	FT1	W4	W5	W6	FT2
Initial mastery								
MS-A0106	13%	22%	7%	47%	6%	11%	8%	52%
MS-A0107	22%	25%	10%	44%	7%	13%	3%	38%
Both	18%	24%	8%	45%	6%	12%	6%	44%
Eventual mastery								
MS-A0106	92%	95%	85%	95%	79%	90%	90%	98%
MS-A0107	90%	90%	83%	90%	85%	90%	87%	94%
Both	91%	92%	84%	92%	82%	90%	89%	96%
Difference								
MS-A0106	80%	73%	78%	47%	73%	78%	82%	46%
MS-A0107	67%	65%	74%	47%	78%	77%	84%	56%
Both	73%	69%	76%	47%	76%	77%	83%	52%

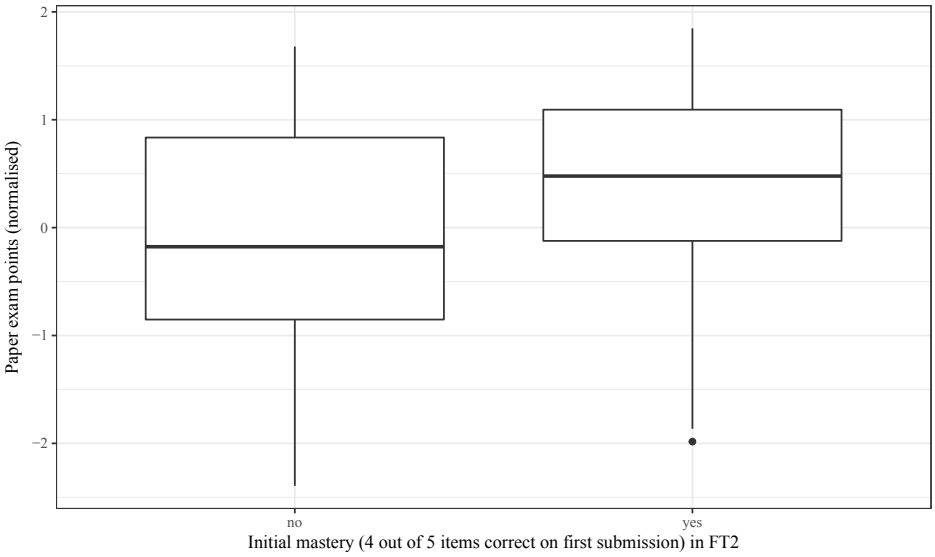

Fig. 3. Examination points compared between mastery and non-mastery students.

mastery on any attempt. The difference between initial and eventual mastery is the percentage of students who gained mastery after the first attempt. The mastery statistics are presented in Table 1. On average, 88% of quiz and 94% of formative test takers achieved mastery eventually. Initial mastery was achieved by 12% on quizzes and 45% on formative tests. As could be expected, the level of initial mastery was higher on the formative tests than on the quizzes. It was however significantly lower than the eventual mastery on quizzes would suggest.

When the pen-and-paper examination scores were compared against initial mastery on the second formative test (Fig. 3), a difference of 0.51 standard deviation in mean test scores was found. This would further suggest that eventual mastery is not entirely equivalent to initial mastery.

5.1. *Qualitative Questionnaires*

The MS-A0106 course feedback questionnaire included four likert-scale questions considering the ML model used on the course, and a summary of results is shown in Table 2. The feedback was mostly in favour of the model.

The "mastery bonus point scheme" (that is where no points are awarded below mastery) seemed to encourage (40% of respondents) more than discourage (14%) practice. 37% found the formative tests very useful, while only 5% found the formative tests not useful at all.

The videos that served as correctives on the formative tests were also found useful (91%) by those who had watched the videos (26%). It is unclear why so many chose not to watch the videos, but the figure should be nonetheless compared to the level of

Table 2
Mastery learning -related questions from the course feedback questionnaire (89 respondents)

	1	2	3	M	SD
Were the practice exams useful? 1) not at all 2) somewhat 3) very	5%	58%	37%	2.3	0.6
Were the practice exam related videos useful?* 1) not at all 2) somewhat 3) very	9%	41%	50%	2.4	0.7
Did you use accessories (calculators, books etc...) in the practice exam? 1) never 2) a few times 3) often	15%	71%	13%	2.0	0.5
Mastery bonus point scheme (0 points if less than 80% done) had mostly . . . to my practice 1) a negative effect 2) no effect 3) a positive effect	14%	47%	40%	2.3	0.7

* - including only those who reported watching the videos (26% of formative test takers)

initial mastery (50% on average in MS-A0106), as they were only watchable after the first attempt.

As the formative tests were meant to be solved without the aid of calculators or learning materials, but were not invigilated, activities which might be considered as `cheating' caused some concern. A majority (84%) of students admitted using accessories like calculators and books during the formative test occasionally, although only 13% reported this often. Judging from the mass of erroneous answers even to the questions easily solvable with a CAS, it would seem that at least the first attempt was usually relatively sincere.

In the responses to the question "*Which things were good on the course? What promoted your learning?*" parts of the LFM model were commended. Almost all of the 74 responses mentioned exercises or exercise sessions in some way. 18 mentioned STACK exercises specifically and 8 the formative tests. Some examples (translated from Finnish to English by us) were:

> *STACK exercises and practice exams were a good addition. Altogether all kinds of extra homework helps, since in my case drilling the basics should be emphaised a bit more before moving on to applications.*
>
> *The middle exams gave a good sense of how well you have mastered the course content.*
>
> *The practice exams forced [me] to revise.*

Also the mastery-oriented bonus point scheme got mentioned:

> *A good thing on the course was that the STACK exercises were, in a way, mandatory.*

There was also a counterpart to the previous question (*Which things were bad / didn't work? What hindered your learning?*). The 68 responses were mostly focused on the

project assignments, lectures and lecture notes. Two students felt there were too many different types of activities on the course.

STACK exercises were mentioned to be both too difficult and not challenging enough.

> *... Also some of the STACK exercises were such that I couldn't find even a hint of a "basic exercise" in those. At least the lectures gave me no clue of solution models, and sometimes I didn't get it even after the teaching assistant had explained it.*

> *... There were all too many exercises and they all were unchallenging. I'd prefer three times less exercises but more challenging ones. Especially STACK exercises often felt like a waste of time.*

The formative tests were not criticised apart from unclear instructions.

5.2. *Predictive Modelling*

Predicting mastery on the formative tests based on prior performance on the quizzes proved to be more challenging than anticipated.

A notable ceiling effect was observed with the unpenalised quiz points. Simulated penalty was later applied with a formula

$$\text{penalised points} = \text{floor}\,(\text{raw points}) \cdot 0.7^{\text{reattempts}}, \tag{1}$$

where floor(x) rounds partial points down towards zero. The formula resulted in a less skewed distribution, shown in Fig. 4. The penalised points from quizzes 4–6 also had a

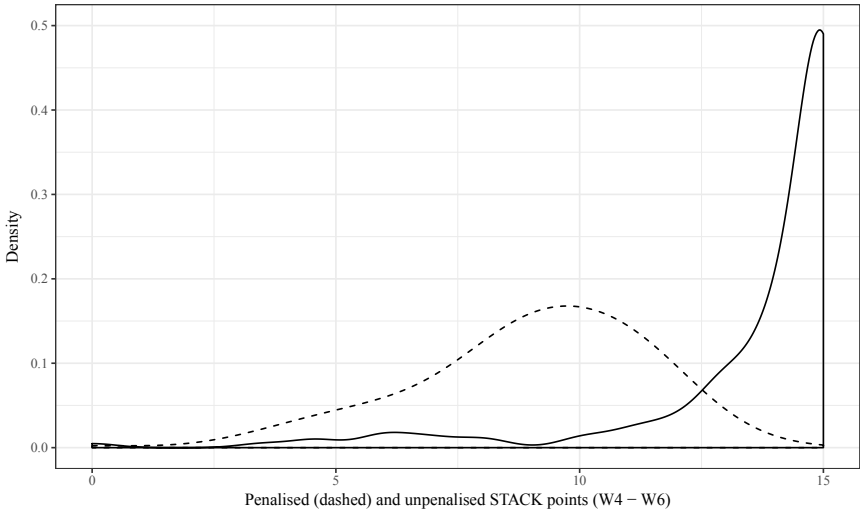

Fig. 4. Ceiling effect on unpenalised points from weeks 4 to 6.

higher Spearman correlation (0.51) with the paper exam points than the unpenalised raw points (0.40).

It should be noted that actual penalties, or limiting the number of attempts, are likely to have some effect on behaviour. High numbers of attempts were observed in some cases, suggesting that some students adopt a trial-and-error strategy when such behaviour goes unpenalised.

When comparing students' exercise point sums against initial mastery on the formative tests, it could be seen that the points provided poor separation between mastery and non-mastery. The difference in median points were highest when one reattempt was allowed, but the ceiling effect became apparent with further attempts.

As could be concluded from the data in Table 1, the eventual mastery in the quizzes did not translate into initial mastery in the formative tests, and the sum of points did not seem to separate mastery and non-mastery either (Fig. 5 and Fig. 6). Therefore, a more sophisticated model would be required to tell whether a student would be likely to achieve mastery in the following formative test. An attempt was made to construct a unit mastery classifier that could ultimately replace the formative tests.

We used various different methods found in the caret R package. Logistic regression performed comparably to some of the more advanced methods such as gradient-boosted trees and was chosen for the model. Logistic regression has the additional advantage of providing class probabilities, which allows us to optimise the classification threshold easily. In this case, the cost of a false positive (inadequate learning) could be considered greater than that of a false negative (waste of time).

Data from quizzes 4–6 were used to predict the initial mastery on the second formative test. In the end, the sum of penalised scores (equation 1) provided the best results. It should be noted that the number of complete observations (168 in the second learning unit) limits how many predictors can be used without overfitting, and might have been the reason why the individual question points and numbers of reattempts did not result in a more accurate model. Some pre-processing of the data was also needed, because the number of reattempts before success and giving up are measuring essentially different things. The exponential penalty scheme (equation 1) was chosen after some experimentation, as this provided a way of reducing points and number of reattempts into a single variable and did not suffer from a floor effect as would a linear model. The resulting model, predicting that a student would not achieve initial mastery on the second formative test, had an accuracy of 0.64 which is a small improvement over predicting that no student would achieve mastery (0.56).

Table 3
Confusion matrix of the classifier (10-fold cross validation with 3 repeats)

	Actual	
Prediction	non-mastery	mastery
non-mastery	33.7%	17.1%
mastery	18.7%	30.6%

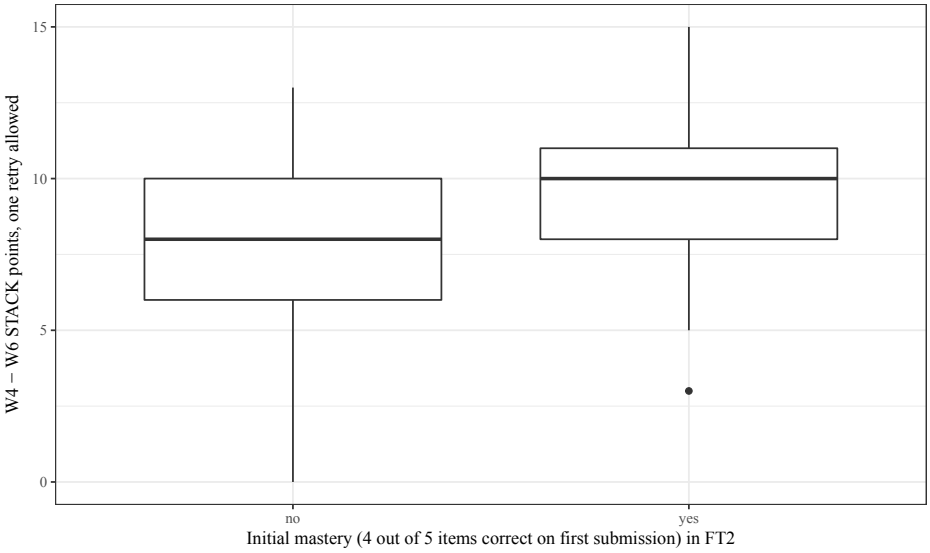

Fig. 5. When one reattempt per exercise problem was taken into account, exercise points between mastery and non-mastery students provided some separation.

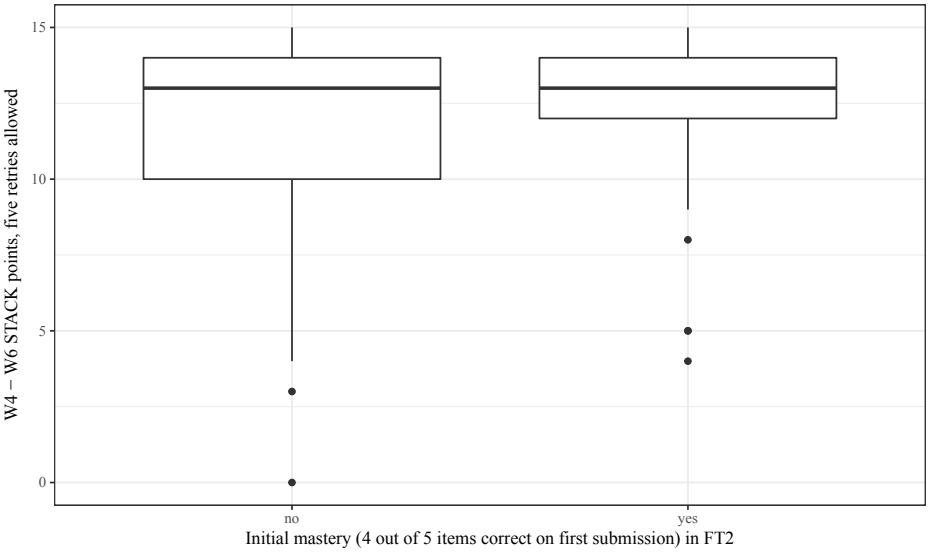

Fig. 6. After five reattempts there is no more Difference in median points.

6. Conclusion

Is STACK suitable for implementing mastery learning?
The STACK system is able to assess most of the learning objectives of Calculus I, and as such is in theory suited for implementing ML on the course. From a technical perspective STACK has many advantages over other similar online assessment systems, particularly in the potential to create sophisticated feedback. However, we believe any online assessment system accepting algebraic answers as students' answers is likely to generate similar overall results.

The implementation was also proven to work in practice, since on each formative test and weekly quiz a considerable portion of students' achievement was raised from non-mastery to mastery (69–83% on the quizzes and 47–52% on the formative tests).

Based on the course feedback, students generally approved of the model. The formative tests were seen as useful and the mastery-oriented bonus point scheme encouraged the students as was intended. However, some concern is caused by the fact that eventual mastery on the weekly quizzes did not translate to initial mastery on the formative tests, and that those who had achieved initial mastery on the second formative test also did better on the paper examination. This could be due to a difference between *exercise* and *test proficiency*.

Solving an exercise problem might be considerably easier than solving the same problem in a test situation for a number of reasons. The student may get help from a peer or a teacher, does not have to rely only on memorised facts, can check his answer, reattempt and may also be more inclined to use a calculator. Similarly, reattempts of a test may also be fundamentally different from the first attempt.

Even so, there is no definite answer to which one of these is the desired level of proficiency. The formative tests however do seem to reveal something the exercises alone cannot, and thus could be beneficial to learning in any case. The difference between initial and eventual mastery could also be blamed on the ineffectiveness of the correctives or the fact that eventual mastery may have been achieved with the aid of a calculator. Paper-based examinations have been refined for many centuries, but using online assessment effectively is in its infancy and further cycles will be necessary to develop a full theoretical understanding of all the related issues.

Can mastery according to formative tests be predicted from STACK exercise data?
Performance on the exercises predicted mastery according to the formative tests to a small extent, and in this case does not warrant using a predictive model as a replacement for the formative tests. However, the result was still positive and could possibly be further improved with more observations, different independent variables and fine-tuning of the model. Some of the considerations from the previous section also applies here. Invigilation of the formative tests could make the model training data more reliable. Our results also suggest that current STACK based examinations are not a completely realistic substitute for pen and paper examinations. This is a rather significant result which deserves further attention to investigate whether the problems are fundamental to online assessment or if the problems are technical and can be addressed by improved software design.

Acknowledgements

The authors acknowledge the financial support of their departments in for this work.

Disclosure statement. There is no conflict of interest reported by the authors.

References

Mäkelä, A.-M., Ali-Löytty, S., Humaloja, J.-P., Joutsenlahti, J., Kauhanen, J., Kaarakka, T. (2016). STACK assignments in university mathematics education. In: *Proceedings of 44th SEFI Conference, Tampere, Finland*.

Adams, R.A., Essex, C. (2013). *Calculus: A Complete Course*, 8th ed. Prentice Hall, Toronto, Canada.

Alpers, B. (2013). *A Framework for Mathematics Curricula in Engineering Education: A Report of the Mathematics Working Group*. Tech. rep., SEFI Mathematics Working Group, Brussels, Belgium.

Anderson, J.R., Corbett, A.T., Koedinger, K.R., Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences*, 4(2), 167–207.

Barbas, H., Schramm, T. (2016). The Hamburg Online Math test MINTFIT for prospective students of STEM degree programmes. In: *Proceedings of SEFI, Tampere, Finland*.

Barry, M.D.J., Steele, N.C. (1992). *A Core Curriculum in Mathematics for the European Engineer*. Tech. Rep. 92.1, Société Européenne pour la Formation des Ingenieurs (SEFI), Plymouth, UK.

Beevers, C.E., Cherry, B.S.G., Foster, M.G., McGuire, G.R.M. (1991). *Software Tools for Computer Aided Learning in Mathematics*. Avebury Technical.

Bellhäuser, H., Lösch, R., Winter, C., Schmitz, B. (October 2016). Applying a web-based training to foster self-regulated learning – effects of an intervention for large numbers of participants. *Internet and Higher Education*, 31, 87–100.

Bloom, B.S. (1984). The 2 Sigma Problem. *Educational Researcher*, 13(6), 4–16.

Brooks, V. (2004). Double marking revisited. *British Journal of Educational Studies*, 52(1), 29–46.

Burkhardt, H., Swan, M. (October 2007). Problem solving in the United Kingdom. *ZDM Mathematics Education*, 39, 395–403.

Coletta, V.P. (2010). *Physics Fundamentals*, 2nd ed. Physics Curriculum and Instruction Inc., Lakeville, Minessota.

Collective, T.D.B.R. (2003). Design-based research: an emerging paradigm for educational inquiry. *Educational Researcher*, 32(1), 5–8.

Foster, C. (2013). Mathematical études: embedding opportunities for developing procedural fluency within rich mathematical contexts. *International Journal of Mathematical Education in Science and Technology*, 55(5), 765–774.

Harjula, M., Malinen, J., Rasila, A. (2017). STACK with state. *MSOR Connections*, 15(2), 60–69.

Hattie, J. (2012). *Visible Learning for Teachers Maximizing Impact on Learning*. Routledge, Oxford, UK.

Hodges, C.B., Murphy, P.F. (June 2009). Sources of self-efficacy beliefs of students in a technology-intensive asynchronous college algebra course. *Internet and Higher Education*, 12(2), 93–97.

Iannone, P., Simpson, A. (2012). *Mapping University Mathematics Assessment Practices*. University of East Anglia, Norwick, UK.

Inglis, M., Attridge, N. (2017). *Does Mathematical Study Develop Logical Thinking?: Testing the Theory of Formal Discipline*. World Scientific Publishing Company.

Kilpatrick, J., Swafford, J., Findell, B. (2001). *Adding it up: Helping Children Learn Mathematics*. National Academy Press, Washinton D.C.

Majander, H., Rasila, A. (2011). *Tutkimus Suuntaamassa 2010-Luvun Matemaattisten Aineiden Opetusta*. Tampereen yliopistopaino Oy – Juvenes Print, Ch. Experiences of continuous formative assessment in engineering mathematics, pp. 197–214.

Michalewicz, Z., Michalewicz, M. (2008). *Puzzle-Based Learning: Introduction to Critical Thinking, Mathematics, and Problem Solving*. Hybrid Publishers.

Mustoe, L., Lawson, D. (March 2002). *Mathematics for the European Engineer: A Curriculum for the Twenty-First Century*. Tech. rep., SEFI Mathematics Working Group.

Noyes, A., Wake, G., Drake, P., Murphy, R. (2011). *Evaluating Mathematics Pathways: Final Report.* DfE Research Report 143, Department for Education, London, UK.

Paiva, R.C., Ferreira, M.S., Mendes, A.G., Eusébio, A.M.J. (2015). Interactive and multimedia contents associated with a system for computer-aided assessment. *Journal of Educational Computing Research*, 52(2), 224–256.

Pointon, A., Sangwin, C.J. (September 2003). An analysis of undergraduate core material in the light of hand held computer algebra systems. *International Journal of Mathematical Education in Science and Technology*, 34(5), 671–686.

Polya, G. (1962). *Mathematical Discovery: On Understanding, Learning, and Teaching Problem Solving*. Wiley, London, UK.

Rasila, A. (July 2016). E-Assessment Material Bank ABACUS. In: *Proceedings of EDULEARN16, 8th Annual International Conference on Education and New Learning Technologies*.

Rasila, A., Havola, L., Majander, H., Malinen, J. (2010). Automatic assessment in engineering mathematics: evaluation of the impact. In: *ReekTori 2010: Symposium of Engineering Education*. Aalto University, Finland, Teaching and Learning Development Unit, `http://www.dipoli.tkk.fi/ok`.

Rasila, A., Malinen, J. (September 2016). MOOCs in First Year Engineering: Mathematics Experiences and Future Aims. In: *Proceedings of 44th SEFI Conference, Tampere, Finland*.

Rasila, A., Malinen, J., Tiitu, H. (2015). Automatic assement and conceptual understanding. *Teaching Mathematics and its Applications*, 34(3), 149–159.

Sangwin, C.J. (2013). *Computer Aided Assessment of Mathematics*. Oxford University Press, Oxford, United Kingdom.

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.

Schoenfeld, A.H. (1985). *Mathematical Problem Solving*. Academic, Orlando, USA.

Skemp, R.R. (1971). *The Psychology of Learning Mathematics*. Penguin.

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

Swan, M., Burkhardt, H. (2012). Designing assessment of performance in mathematics. *Educational Designer*, 2(5), 1–41.

Tallman, M.A., Carlson, M.P., Bressoud, D.M., Pearson, M. (April 2016). A Characterization of Calculus I Final Exams in U.S. Colleges and Universities. *International Journal of Research in Undergraduate Mathematics Education*, 2(1), 105–133.

Watson, A., Ohtani, M. (Eds.) (2015). *Task Design In Mathematics Education: an ICMI study*. Vol. 22 of *New ICMI Study Series*. Springer International Publishing, Switzerland.

T. Pelkola is a MSc student in Mathematics Education who recently worked as a research assistant in Dr. Rasila's research team at Aalto University. His research interests include blended learning, automatic assessment and learning analytics.

A. Rasila is an Associate Professor at Guangdong Technion – Israel Institute of Technology. Previously, he worked at Aalto University Department of Mathematics and Systems Analysis, where he has acted as the leader of the computer aided mathematics teaching research group MatTa since 2006. His other academic interests include complex analysis, partial differential equations and computer aided methods in mathematical analysis.

C. Sangwin is Professor of Mathematics Education at Edinburgh University. His learning and teaching interests include (i) automatic assessment of mathematics using computer algebra, and (ii) problem solving using Moore method and similar student-centred approaches. In 2006 he was awarded a National Teaching Fellowship.