Development and evaluation of a partially-automated approach to the assessment of undergraduate mathematics

Peter Rowlett
Nottingham Trent University

This research explored assessment and e-assessment in undergraduate mathematics and proposed a novel, partially-automated approach, in which assessment is set via computer but completed and marked offline. This potentially offers: reduced efficiency of marking but increased validity compared with examination, via deeper and more open-ended questions; increased reliability compared with coursework, by reduction of plagiarism through individualised questions; increased efficiency for setting questions compared with e-assessment, as there is no need to second-guess the limitations of user input and automated marking. Implementation was in a final year module intended to develop students' graduate skills, including group work and real-world problem-solving. Individual work alongside a group project aimed to assess individual contribution to learning outcomes. The deeper, open-ended nature of the task did not suit timed examination conditions or automated marking, but the similarity of the individual and group tasks meant the risk of plagiarism was high. Evaluation took three forms: a second-marker experiment, to test reliability and assess validity; student feedback, to examine student views particularly about plagiarism and individualised assessment; and, comparison of marks, to investigate plagiarism. This paper will discuss the development and evaluation of this assessment approach in an undergraduate mathematics context.

Keywords: assessment, higher education, e-assessment.

Introduction

Coursework carries greater potential validity than other assessment methods, in part because of the ability to access greater depth through more open-ended questions (Cox, 2011; Thomlinson, Robinson and Challis, 2010). However, this is accompanied by concerns about plagiarism (Cox, 2011; Iannone and Simpson, 2012; Thomlinson, Robinson and Challis, 2010). Plagiarism might be simple copying (Beevers, 2006), collaborative working taken too far (Cooper, 2002) or impersonation (Beevers, Wild, McGuire, Fiddles and Youngson, 1999). E-assessment allows randomisation of question parameters, meaning individualised work can be set (Gwynllyw and Henderson, 2009), which may be useful for the avoidance of plagiarism (Hatt, 2007). However, writing "reliable, valid questions" for e-assessment is "a difficult task, requiring expertise" (Sangwin, 2012: 7), and input interfaces may add learning requirements (Lawson, 2002) and cognitive load (Mavrikis and Maciocia, 2003) unrelated to the assessment objectives. E-assessment may not be suitable for testing conceptual understanding (Robinson, Hernandez-Martinez and Broughton, 2012), extended work (Sangwin, 2012) or problem-solving (Beevers and Paterson, 2002). Therefore, a partially-automated approach is proposed, in which assessment is set via an automated assessment generator but printed for completion and marking offline as a traditional piece of coursework. The potential exists to maintain the validity of

coursework, while increasing the reliability via a reduction in plagiarism, at the cost of decreased marking efficiency. The main question of this research becomes whether there is a context in which this approach could be more useful than existing methods.

Teaching and learning context

Implementation was during a group project in a final year module (not at the author's current institution). Individual work was used alongside group work, partly to increase the amount of the module mark that reflected individual ability, following concern over groups carrying students as what MacBean, Graham and Sangwin call "passengers" (2001: 7).

The main project saw students spend three weeks answering a brief from a (fictional) client. Specifically, students were to investigate 'Art Gallery Problems', which are concerned with determining the minimum number of point 'guards' necessary for all points in a polygon (the 'art gallery') to be connected by a straight line (line of sight) to at least one guard (O'Rourke, 1987). The brief gave three art gallery floor plans and asked groups to propose the size of a staff which must be hired to guard each of these in a short report. The individual coursework gave a single art gallery floor plan and asked the same question.

The similarity of the individual and group tasks meant the risk of in-team plagiarism or collusion was high, suggesting a need for exam conditions or individualised work. The individual work required students to solve a problem and discuss its solution in the context of the real-world scenario. Challis, Houston and Stirling (2004) say that written examinations are "not useful for assessing extended investigations" (45) and this is beyond the limits of automated marking. In addition, a solution would involve drawing a diagram, which via computer input would introduce additional, irrelevant learning outcomes, such as use of a drawing package. The need to produce individualised work via randomisation, lack of suitability of automated marking and the need for students to be able to hand-write their answers suggests that the partially-automated approach suggested above may be appropriate.

Individualised worksheets were generated using the system Numbas, principally a mathematically-aware e-assessment system (Foster, Perfect and Youd, 2012) that can also provide printable question sheets and corresponding answer sheets (where answers can be generated). Producing this was much like writing questions for an e-assessment system, without the requirement to comply with the limits of automated marking. (Systems other than Numbas could presumably be adapted for a similar approach.) When marking, answers could not be learned and student submissions needed to be matched to an appropriate answer sheet using an ID number, which added to the time taken for marking, though not substantially.

Evaluation method

The partially-automated approach was proposed as having potential to maintain the validity of a piece of coursework while increasing reliability via reduction in plagiarism. In general, an assessment method must not be unduly sensitive to who is doing the marking (Cox, 2011). It is important, therefore, to check that reliability, with respect to who is doing the marking, and validity, specifically what is being assessed, are not adversely affected by the use of this method, and to examine its contribution to reducing plagiarism.

Discussing practicalities of evaluation, Moore (2011) suggests that it is important that evaluation is proportional to the activity being evaluated. This has an

effect on the workload incurred by participants, in this case students giving feedback and volunteer second-markers. Data already compiled was used for a comparison of marks to reduce the overall resource need.

Second-marker experiment

Asking a second person to mark an assessment is a straightforward way to test the objectivity and accuracy of multiple markers. We should not expect complete agreement between multiple markers for this deeper, more open-ended form of assessment (Cox, 2011). Also, Bloxham (2009) criticises the inherent assumptions that higher education work can be awarded an accurate and reliable mark and that academics share common views regarding academic standards. Therefore, conclusions about whether the level of agreement found between multiple markers is reasonable or not require context. In order to calibrate expectations and provide reference information, the level of agreement for multiple markers of two more established assessment methods was examined. This used: a class test under examination conditions, a method of assessment recognised as being highly reliable (Cox, 2011); and, an open-ended piece of coursework, a method reported as having problems with consistency of marking (Iannone and Simpson, 2012). Comment on the differences in the marks and the intraclass correlation coefficient (ICC) are presented.

A simple test of validity was to ask the second-markers what they thought the coursework was assessing. They were given enough information to mark student work, but were not told the intended learning outcomes.

Written examination reference experiment

The work arose from an open-book test under examination conditions, during a basic mathematical methods module for first year mathematics students. The test comprised five well-focused, short problem questions for which 50 marks were available. A 10% sample of all scripts was checked by a moderator, with reference to the original marks, as part of the usual departmental process. The moderator agreed with the marks awarded in all cases.

I marked a sample of ten scripts without reference to the marks assigned by the original marker but using the same mark scheme (blind second-marking). The mark scheme was a set of worked solutions with individual marks indicated for components of answers and for working. The original marker was working at the same university as me so was used to marking work from similar students.

Coursework reference experiment

The work for this reference experiment arose from a task to write an 800-1000 word review of a popular book or textbook on mathematics or the history of mathematics. The marking criteria specified those pieces of information that each review should contain, as well as some general subjective criteria around the quality of the writing and level of critical understanding. Marks were a simple percentage. A sample of work had previously been approved via a departmental moderation procedure, conducted with reference to the original marks.

I marked a sample of 14 scripts via blind second-marking. The original marker was working at a different university with a similar entry requirement to my own.

Second marking of the individualised coursework

Three second-marker volunteers each had experience of marking work at university; one as a senior academic, one as a junior academic and one as a PhD student. One was from a university with a similar entry requirement to where the work was submitted, one had a lower entry requirement and one a higher entry requirement.

A 10% sample of student work was anonymised (5 pieces from 44 submitted). This was provided along with grade descriptions, a mark scheme and a sample piece of marked work (written to be correct on the non-subjective parts of the mark scheme) as a reference piece since the second-markers were not familiar with the topic.

Student feedback

Students were asked via a questionnaire to express their views, anonymously, on the role of individualised work and how this affected interaction with other students, as well as questions about plagiarism in this assignment and other work. This was completed by the cohort taking the assessment task described above (group A) and by a group at a different university (group B) to provide input from an independent cohort of students with which I had not interacted. The lecturer for group B had also used the technique developed for this project via Numbas for an individualised formative in-class question sheet in a final year digital signal processing module. For both groups, questionnaires were administered via Google Docs. For group A, this was six weeks after the group project had been submitted. For group B, this was at the end of the session in which the individualised assessment was used.

Comparison of marks

The risk was around intra-group plagiarism, since group members were working together on similar problems, so individual marks from within groups were examined. Wide variety of individual marks might indicate that intra-group plagiarism is not a large problem. A lack of variety, however, could indicate plagiarism or perhaps just that students had been learning the topic together and so have similar understanding. If group members colluded on the individual work, we might expect to see similarity between individual and group marks, since they certainly colluded on the latter.

The correlation of raw group project marks and rankings (prior to scaling due to peer assessment of contribution) with the individualised coursework is presented via Pearson's ρ and Kendall's τ. The dispersion of marks for the coursework is examined via the range and standard deviation of the marks within each group.

Results

Second-marker experiment

Written examination reference experiment: There were five discrepancies of one or two marks (2% or 4% of the total) in ten scripts. The ICC for the two sets of marks is 0.992. This value is regarded by Landis and Koch (1977) as an "almost perfect" level of agreement (165).

Coursework reference experiment: There were differences in all fourteen pieces of work. Six were differences of around 5% or less, a further six were around 10% and two were greater differences. The ICC for the two sets of marks is 0.586. This value is regarded by Landis and Koch as a "moderate" level of agreement (165).

Second marking of the individualised coursework: The marks are given in table 1. The ICC for the four sets of marks is 0.635. This value is regarded by Landis and Koch as a "substantial" level of agreement (165).

Student	PR	Second-marker A	Second-marker B	Second-marker C
1	56	31	38	49
2	74	64	59	72
3	67	72	74	77
4	67	46	51	51
5	74	59	54	69

Table 1: Original and second marks for five pieces of work submitted for the individual coursework.

Comments on learning outcomes: Second-markers A and B suggested wording very similar to the three intended outcomes (problem-solving, working in depth and communicating results). Second-marker C, with less experience, suggested two of the three but did not identify communication skills. No marker proposed additional outcomes.

Student feedback

Students were asked to indicate their level of agreement with each of four statements, listed with numbers of responses in table 2. Also in table 2 are the p-values obtained for each Likert-type question when comparing the two groups via Fisher's Exact Test. In each case, there is no evidence at the 5% level to reject the null hypothesis that the distribution of answers is independent of the group. Responses to two questions about copying, which were accompanied by a reminder that the questionnaire was anonymous, are given in table 3. Again, p-values from Fisher's Exact Test are listed in table 3 and do not give evidence at the 5% level to reject the same null hypothesis.

Group	'Strongly disagree' 1	2	3	4	5 'Strongly agree'	p-value
"I disliked having different questions because I wanted to work together with another student on our answers."						
A	12	16	13	1	0	0.08851
B	3	8	2	3	0	
"I liked having different questions because it meant I could freely discuss the work with others with no risk of plagiarism."						
A	0	1	10	22	9	0.6193
B	0	0	4	6	6	
"I liked having different questions because it meant that no one could copy from me."						
A	0	4	14	17	7	0.1366
B	2	2	5	3	4	
"If we had been set identical questions, (members of our group [group A]/some students [group B]) would have copied answers from other students."						

A	2	5	11	15	9	0.3132
B	2	1	6	2	5	

Table 2: Number of students indicating level of agreement with four statements about individualised work.

Group	Yes	No	p-value
"While at university, I have copied work from other students"			
A	22	19	0.1513
B	5	11	
"While at university, other students have copied work from me"			
A	35	7	0.2811
B	11	5	

Table 3: Number of students answering yes and no to two questions about copying.

Comparison of marks

The raw group project marks and rankings do not correlate well with the marks and rankings for the individualised coursework (ρ=0.230; τ=0.229). The range and standard deviation of the individual marks within each group are presented in table 4. Individual marks for each group represent a range of at least 23 marks and up to 31 marks, and have a standard deviation of at least 8.216 and up to 11.411.

Group	Individualised coursework marks range for group members	Individualised coursework standard deviation for group members (3 d.p.)
A	31	11.411
B	30	10.706
C	23	8.216
D	28	9.584
E	30	9.513

Table 4: Marks range and standard deviation for the individualised coursework within each group.

Conclusions and discussion

This research proposed a novel partially-automated approach, in which the tools of e-assessment are used to set an individualised assessment that is taken and marked offline. The evaluation focused on whether a reliable and valid assessment had been set and attempted to examine to what extent this addressed the issue of plagiarism via comparison of marks between students working in the same group.

Second-marker reference experiments showed a high level of agreement for an open-book written examination and a moderate level of agreement for an open-ended piece of coursework. For the individualised coursework, a group of four markers showed a level of agreement that was between the two reference experiments, and close to the open-ended piece of work. This suggests a conclusion that the coursework was, despite its unusual status as individualised work, not unduly sensitive to who was doing the marking.

The three second-markers identified the learning outcomes with a fair degree of accuracy and did not recognise unintended learning outcomes being assessed. The conclusion, based on this, is that the assignment was assessing what it was intended to assess, and no more.

Some sources question whether concern about plagiarism is overblown (e.g., Cox, 2011). Among my 42 students, 22 confessed copying work from another student at university and 35 said another student had copied from them at university. Students generally appreciated being able to discuss individualised work with no risk of plagiarism and reported concerns about copying, including that if identical work had been set then some students would have copied from others. The responses from an independent reference group of students at another university are apparently similar.

Individual marks were not well correlated with group marks and dispersion of individual marks in each group was high. We may conclude, therefore, that plagiarism was not a big problem.

Since student feedback indicated a high risk of plagiarism and none was detected, we may conclude that the individualised nature of the coursework did contribute to a reduction in plagiarism. One of the interviewees of Thomlinson, Robinson and Challis (2010) said that it is "not clear what the real benefit is" of coursework, given that copying is a particular problem among weaker students, and Iannone and Simpson (2012) report some departments moving away from coursework towards in-class tests. The partially-automated approach proposed here appears to be capable of adapting a coursework assignment to make it less sensitive to plagiarism while maintaining its reliability and validity, though it lead to a reduced efficiency for the marker. By contrast, converting the coursework to a written examination or e-assessment in order to reduce the risk of plagiarism can result in reduced validity.

Acknowledgements

I am grateful to Christian Perfect, School of Mathematics and Statistics, Newcastle University, for adapting the Numbas e-assessment system for my experiment. The 'worksheet' theme he created for this remains part of the free system at numbas.mathcentre.ac.uk. I am also grateful for the anonymous contributions to this work by the students who gave feedback, the second-markers and the other lecturer who administered the survey with group B.

References

Beevers, C. (2006) IT was twenty years ago today... *Maths-CAA Series, January*. Retrieved from: www.mathstore.ac.uk/repository/mathscaa_jan2006.pdf

Beevers, C. & Paterson, J. (2002) Assessment in mathematics. In Kahn, P. & Kyle, J. (Eds.) *Effective Teaching and Learning in Mathematics & its Applications* (pp. 49-61). London, U.K.: Kogan Page.

Beevers, C.E., Wild, D.G., McGuire, G.R., Fiddles, D.J. & Youngson, M.A. (1999) Issues of partial credit in mathematical assessment by computer. *ALT-J,* 7(1), 26-32.

Bloxham, S. (2009) Marking and moderation in the UK: false assumptions and wasted resources. *Assessment & Evaluation in Higher Education,* 34(2), 209-220.

Challis, N., Houston, K. and Stirling, D., 2004. *Supporting Good Practice in Assessment*. Birmingham, U.K.: Mathematics, Statistics and OR Network.

Cooper, D. (2002) A do-it-yourself approach to Computer-Aided Assessment. *Maths-CAA Series, August*. Retrieved from: www.mathstore.ac.uk/repository/mathscaa_aug2002.pdf

Cox, B. (2011) *Teaching Mathematics in Higher Education – the basics and beyond*. Birmingham, U.K.: Mathematics, Statistics and OR (MSOR) Network.

Foster, B., Perfect, C. & Youd, A. (2012) A completely client-side approach to e-assessment and e-learning of mathematics and statistics. *International Journal of e-Assessment,* 2(2). Retrieved from: journals.sfu.ca/ijea/index.php/journal/article/viewFile/35/37

Gwynllyw, R. & Henderson, K. (2009) DEWIS - a computer aided assessment system for mathematics and statistics. In: Green, D. (Ed.) *Proceedings of the CETL-MSOR Conference, Lancaster University, 8th-9th September 2008* (pp. 38-44). Birmingham, U.K.: Mathematics, Statistics and OR Network.

Hatt, J. (2007) Computer-Aided Assessment and Learning in Decision-Based Mathematics. In Nunes, M.B. & McPherson, M. (Eds.), *Proceedings of the IADIS International Conference on e-Learning, Lisbon, Portugal 6th-8th July 2007* (pp. 382-385). Lisbon, Portugal: International Association for Development of the Information Society.

Iannone, P. & Simpson, A. (2012) A Survey of Current Assessment Practices. In Iannone, P. & Simpson, A. (Eds.) *Mapping University Mathematics Assessment Practices* (pp. 3-15). Norwich, U.K.: University of East Anglia.

Landis, J.R. & Koch, G.G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics,* 33(1), 159-174.

Lawson, D. (2002) Computer-aided assessment in mathematics: Panacea or propaganda? *CAL-laborate*, 9(1). Retrieved from: ojs-prod.library.usyd.edu.au/index.php/CAL/article/download/6095/6745

MacBean, J., Graham, T. & Sangwin, C. (2001) *Guidelines for Introducing Groupwork in Undergraduate Mathematics*. Birmingham, U.K.: Mathematics, Statistics and OR Network.

Mavrikis, M. & Maciocia, A. (2003) Incorporating Assessment into an Interactive Learning Environment for Mathematics. *Maths-CAA Series, June*. Retrieved from: www.mathstore.ac.uk/repository/mathscaa_jun2003.pdf

Moore, I. (2011) *Evaluating your Teaching Innovation*. Birmingham, U.K.: National HE STEM Programme.

O'Rourke, J. (1987) *Art gallery theorems and algorithms*. New York, U.S.A.: Oxford University Press.

Robinson, C.L., Hernandez-Martinez, P. & Broughton, S. (2012) Mathematics Lecturers' Practice and Perception of Computer-Aided Assessment. In: Iannone, P. & Simpson, A. (Eds.) *Mapping University Mathematics Assessment Practices* (pp. 105-117). Norwich, U.K.: University of East Anglia.

Sangwin, C. (2012) Computer Aided Assessment of Mathematics Using STACK. *Proceedings of 12th International Congress on Mathematical Education, 8th-15th July, 2012, COEX, Seoul, Korea*. Gangnae-myeon, South Korea: Korea National University of Education. Retrieved from: www.icme12.org/upload/submission/1886_F.pdf

Thomlinson, M.M., Robinson, M. & Challis, N.V. (2010) Coursework, what should be its nature and assessment weight? In Robinson, M., Challis, N. & Thomlinson, M. (Eds.), *Maths at University: Reflections on experience, practice and provision* (pp. 122-126). Birmingham, U.K.: More Maths Grads.