Mapping University Mathematics Assessment Practices

Edited by

Paola Iannone
University of East Anglia

Adrian Simpson
Durham University

Chapter 19
Assessing Proofs in Pure Mathematics

Timothy J. Hetherington

Abstract Many mistakes made by students in coursework and exams arise from poor notation, poor expression of ideas or common misunderstandings. Previous coursework used to assess proof explored their comprehension, clarity of expression, and appreciation of the importance of rigour, but was very time-consuming to mark. Moreover, in the last three years, student numbers have doubled. These issues combined to mean that the assessment used in previous years was no longer viable. This report outlines a project which sought to facilitate the implementation and development of an interesting and innovative assessment on mathematical proof that reduced the marking burden, but that was still educationally rich. The result was a test on mathematical proof which began as a conventional multiple choice quiz, but has now evolved somewhat. This test has dramatically reduced marking time, whilst maintaining student engagement in, and learning from, the process of writing proofs.

19.1 Background and rationale

Since 2008, when I started lecturing, I have been the module leader for a first year module that teaches students about proof. To encourage students to develop their ability to write mathematics I set a piece of coursework (worth 30%) that required students to write a series of short proofs using a number of the standard techniques; direct proof, proof by contradiction, and induction. However, within the framework of each technique the students employed a variety of ideas (with varying success), which meant that each argument, however unorthodox, had to be carefully followed through. For a small group of 40 students this was not an overly onerous task, probably taking between 20 and 25 hours. However, since 2008 student numbers on the mathematics course at Nottingham Trent University have been rising steadily, and by 2011 student numbers were double those of 2008. This increase was the key

Timothy J. Hetherington, e-mail: `timothy.hetherington@ntu.ac.uk`
School of Science and Technology
Nottingham Trent University

driver to rethink the coursework assignment as the huge marking load had made the previous practice unsustainable.

During a Post-Graduate Certificate in Higher Education in 2009, I initially explored ideas on e-assessments, which I thought would naturally address the issue of marking load. These ideas initiated thoughts about how a multiple-choice test on mathematical proof could be developed so that marking the test was easy, but that the test itself was not too trivial. Having built up experience of the common mistakes made by students, it was possible to use these in the development of a multiple-choice test on mathematical proof by basing each incorrect option on a different common error. At the end of the test (worth 10% of the module) students were encouraged to provide feedback on the back of their test papers. There were many positive and supportive comments from the students about this novel mathematical assessment. One student said, "I really enjoyed this new type of test as I've never done anything like it before."

More recently the students sat a second version of the test, which again was well received. The main difference with the second test was that as well as choosing the correct answer, students also had to justify why they had rejected the other three options. This approach was used to eliminate the effect of guessing. After the second test, students were invited to complete a questionnaire and were also given the opportunity to be interviewed about the two tests.

19.2 Implementation

As mentioned, during the previous three years, the coursework assignment required students to write a short series of proofs. In this time I gained experience of common misconceptions and built up a library of false proofs, each of which contained some element of erroneous thinking, be it poor notation, poor expression of ideas, or some more fundamental misunderstanding. For example, students often assume the truth of an equation or inequality that they need to prove. They then work on this equation or inequality, sometimes on both sides at the same time, until they get something they know is true or the same expression on each side of the equals sign. Other popular mistakes include:

- omitting critical information such as what type of numbers are being used;
- misunderstanding concepts or notation such as thinking that 'divides' is the same as 'divided by';
- placing equals signs at the start of every line of working;
- not identifying the correct assumptions;
- not forming the contrapositive statement correctly;
- proving the converse of what was required;
- not covering all possible cases;
- proving the wrong base case;
- claiming that the induction hypothesis is for all natural numbers.

These errors were then used to create three false proofs for each question, where each false proof highlighted one common error (see figure 19.1 for an example). By basing the false proofs on students' answers from previous years it meant that the incorrect options would seem plausible to some of the students and the popular misconceptions would be highlighted. The provision of formative feedback enables students to learn what is expected of them (Yorke, 2003), and so once detailed feedback on the test was provided, students would be aware not only which option was a logical and well-written proof, but also what common traps to avoid. One aim was to get students to think about how they write mathematics and to appreciate the importance of notation and how they set work out. Many marks are lost in coursework and exams through poor notation, poor expression of ideas, and common misunderstandings. By making students aware in year one of how to write mathematics, it is hoped that improvements may be seen throughout the rest of their course.

1. If $\frac{b}{a} = p$ then $b = ap$ and if $\frac{c}{a} = q$ then $c = aq$. So
$$b + c = ap + aq = a(p+q) \qquad \text{and} \qquad b - c = ap - aq = a(p-q)$$
and it follows that $a|(b \pm c)$.
Incorrect: last bit doesn't follow unless $(p \pm q) \in \mathbb{Z}$.
2. If $a|b$ and $a|c$, then
$$\frac{b}{a} \pm \frac{c}{a} = \frac{b \pm c}{a}.$$
Therefore $a|(b \pm c)$.
Incorrect: is $(b \pm c)/a$ an integer?
3. If $a|b$ then $\exists m \in \mathbb{Z}$ such that $b = am$. Similarly, if $a|c$ then $\exists n \in \mathbb{Z}$ such that $c = an$. So
$$\begin{aligned} b \pm c &= am \pm an \\ &= a(m \pm n). \end{aligned}$$
Since $(m \pm n) \in \mathbb{Z}$, it follows that $a|(b \pm c)$.
Correct.
4. Let
$$\frac{a}{b} = x \qquad \text{and} \qquad \frac{a}{c} = y,$$
where $x, y \in \mathbb{Z}$. Then
$$\frac{a}{x} = b \qquad \text{and} \qquad \frac{a}{y} = c,$$
so
$$b \pm c = \frac{a}{x} \pm \frac{a}{y},$$
which implies that
$$\frac{a}{b \pm c} = x \pm y.$$
Hence $a|(b \pm c)$ since $x \pm y \in \mathbb{Z}$.
Incorrect: $a|b$ is not the same as a/b.

Fig. 19.1 Choices for proofs of the theorem 'if $a|b$ and $a|c$ then $a|(b \pm c)$', with model answers

Three weeks before the test, students were given a list of ten questions on proof. They were told that eight of these would be on their test paper, and that not every paper would be the same. It was hoped that by providing the questions prior to the test, students would spend time trying to write their own proofs, thereby engaging in exactly the same activity that was required for the assignment in previous years. Two examples were provided so that students were aware of how the test would be structured and the task that was required of them. In each example not only was the correct option highlighted, but reasons why the other options were wrong were also given.

The first time that I tried using this test, it was simply multiple-choice; there were eight questions (the easiest eight out of the ten provided), each with four options. Given that the answer for each question was a single letter, copying would be quite easy and so four different versions of the paper were created, each with the same questions, but in a different order. The answer sheet was attached to the back of the paper, again to make it more difficult to copy. Students were explicitly told not to detach their answer sheet as otherwise it would not be known which question paper they had had. With this being an unusual assessment, I was unsure of how long it would take students to complete, and so they were given 80 minutes for the test. However, nearly all had finished in 30 minutes. To mark this test the scripts were first sorted into four groups, each group containing the same version of the test so that each group should have the answers in the same order. It was then very simple to mark, and the whole process of sorting and marking took under two hours for 80 scripts, instead of the 20-25 hours the written coursework had taken. For the first test the mean mark was 67% with a standard deviation of 21%. The distribution of marks was highly skewed; almost half the class achieved a first class mark and three-quarters achieved at least an upper second.

Upon reflection there were some issues with the first test. Given the nature of the topic the options are much longer than one would usually find in a multiple-choice test and so there were not many questions on the paper, only eight in total. This meant that there were only eight marks available, one for each question. Consequently, a student's mark could be affected dramatically by getting one or two questions right or wrong. Given the nature of multiple-choice this means that it was possible that some students had achieved an excellent mark partially due to good luck, whereas other students had achieved a poorer mark due to bad luck.

For example, suppose that a student eliminates two of the options, but cannot decide between the remaining two. If they guess correctly they will achieve 100% for the question, but if they guess incorrectly they will be awarded nothing. However, in each case the student actually has only partial understanding and I wanted to ensure that this partial understanding was reflected in the mark for the multiple-choice question. The wrong answers were not more or less wrong than each other so it was not possible to have different weightings for different answers.

So, for the second trial of the multiple-choice test, to make it fairer it was decided that students should not only identify the correct answer, but they should justify why they had rejected the other three options. Each correct answer was worth one mark, as was each correct justification for rejecting another option. This gives a total of four marks per question and thirty-two marks for the paper. On this version of the test the proportion of marks that could be achieved by guessing has been significantly reduced. Furthermore, comprehension and deeper understanding are tested more rigorously as it is arguably more difficult to identify and articulate precisely what is wrong in an argument than to simply identify the correct argument. The students were given 45 minutes to complete this version of the test. As before, the scripts were sorted into four groups prior to marking, and the process of sorting and marking 80 scripts took about eight hours. For this second test the mean mark was 54% with a standard deviation of 14. The distribution of marks was much closer to a normal distribution, with the majority (58%) of the marks at second class level.

19.3 Evaluation

When evaluating and reflecting upon the delivery, content and assessment of a module it is important to obtain the views of students and to integrate this feedback into the continual cycle of module development (Harvey, 2001). Therefore, to evaluate the project, all students who took the tests were invited to complete a simple online questionnaire with ten questions. The response rate was just over 30% (25/80 students). However, as qualitative feedback is much more useful when trying to make improvements to a module (Harvey, 2001), twelve students were interviewed to establish further their thoughts, ideas, and approaches to the tests. Given that this group of students sat both versions of the test, their views, ideas and insights are highly valuable for evaluating and developing the test further. After all, although the aim was to reduce staff marking time it was not to be at the expense of an educationally rich task; if the assessment is not fit for purpose then the marking time is irrelevant as the most important aspect is the student experience. Since its conception the test has been developed to cater for the learning needs of the students, whilst aiming for a reduction in marking time, which will also lead to quicker feedback. This, in turn, will help to enhance the student experience because for effective learning to take place it is essential to provide quality feedback that is both constructive and timely (Huxham, 2007). Informal feedback after the first test fed into the development of the second test. The more formal feedback presented here, after the second test, will feed into the future development of this assessment.

The first question on the questionnaire asked “Have you ever done a mathematics test like this before?” The response was that 24/25 had never done a mathematics test like this. The results from the other questions are presented in figure 19.2. It can be seen that the vast majority agreed or strongly agreed with the statements presented.

The results from the questionnaire suggest that this was an innovative method of assessment that was enjoyed by the vast majority of the students. Despite the novel assessment method, students felt well-prepared for the test, citing the provision of questions and examples prior to the test as being particularly useful. Students felt that the assessment was a fair measure of their understanding, particularly the second test, and having learnt from the feedback they feel that the whole experience has made them more confident at writing mathematical proofs.

There was also a space for general comments, which most chose not to fill in, but there were some good points raised. Three students said that a detachable answer

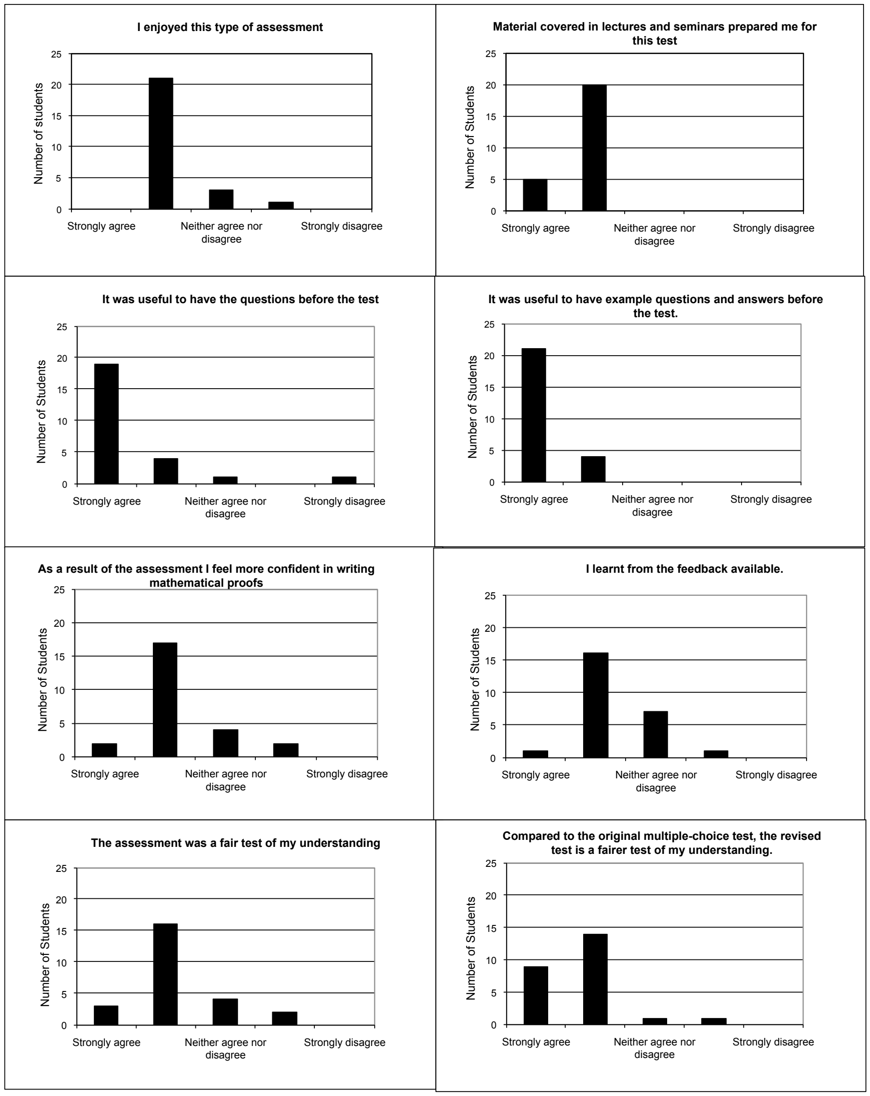

Fig. 19.2 Responses to the questionnaire

sheet would make completing the test easier as currently it is difficult having to flick back and forth to fill in the answers. Two students mentioned that the second test was a better test saying, “Although the first test was easier because one didn’t have to explain any reasoning, the second test is a better assessment of pupil’s abilities” and “The revised version was better at testing our understanding of proof than the original version.” One student commented that they did not think each question should be worth the same mark, another said, “It was quite unclear how we were supposed to write the reason for our choices; were we to write next to every option why it was/wasn't that, or just chose one and justify that one." To pick up on this last comment, it is felt that the rubric on the test paper was perfectly clear; "Stating the correct option will be worth 1 mark AND for each of the three incorrect options a brief justification for why it is incorrect will be worth 1 mark."

The interviews were extremely useful as all of the students who volunteered were quite talkative and had obviously reflected upon their experience.

They were asked to comment on:

- their general thoughts about the test;
- what approach they took;
- whether they would use the same approach again;
- what they thought of the feedback and whether they had learnt from it;
- whether the test raised their awareness of how or how not to write a proof;
- how easy or difficult they found it to put their reasons into words and
- whether they had any suggestions for improving the test.

The general consensus was that the second test, although more difficult, was "fairer than the first one" because "it stopped you from guessing answers" and was "a better way of testing proof". They thought that "having to say why an answer was wrong was a better idea" because "you need to understand to choose the right answer" and they liked the fact that "if you've got half an idea you can still get some marks".

Many of the students had worked in groups to prepare for the test. All of them essentially said that they had "worked through the questions as if it wasn't multiple-choice". They had looked at examples of proof, some making links with other modules, and had tried "to get an understanding of the structure". Some students had tried to think of mistakes that they could make by "working out what the answers might be". It was felt that writing out one's own version of the proof was "good practice". One student said that in the test he "used a highlighter to find differences". All students said that they would use the same approach again.

After the first test each student was given their script, which was attached to the question paper. For each incorrect answer the correct option had been written on their answer sheet. After the second test a mark scheme of model answers was provided so that students could identify exactly why certain options were wrong (as seen in Figure 19.1). The advantages of using model answers are that such feedback can be distributed quickly, avoids being overly negative, and requires the student to actively reflect upon how their own work compares to the model answer (Huxham, 2007). All agreed that the feedback provided was good and that they could not think of anything else that would be useful; "you gave everything we needed". Most students said they had looked at all of the available feedback, and learnt from it. The mark scheme was felt to be most helpful because it gives reasons why options are incorrect; "I learnt why rather than 'it is just wrong'. Now I can pinpoint why."

All students agreed that the test had raised their awareness of how to write mathematics by encouraging them to think about what they are writing; "it made me realise that you have to be pretty accurate with maths proofs" and "it makes you look at notation rather than just the working out". "The way you write it is important. It made us think about why it is wrong. Hopefully we won't make the same mistakes."

Students seemed to "feel more confident about writing proofs now", some adding, "especially in induction questions". Interestingly one student said, "I have more of an idea than if you'd tested in a different way". However, the words of one student provide a reminder about the importance of the continual reinforcement of ideas, "some mistakes are quite obvious, but I think I'll forget".

Most students claimed "it was hard to write reasons" for rejecting certain options; "it is in your head but it is difficult to write down". It was felt that "some were harder than others", with the induction questions being easiest in this sense as explanations such as 'wrong base case' were simple to state.

Ideas for further development were: detaching the answer sheet, making the test open book, giving more time (for the second test), splitting each option into steps to help students pinpoint which step is wrong, writing the answers in the booklet at the end of each question, and giving clearer instructions about what response to give for the correct option.

To summarise the interviews, the students thought the second test was better and fairer as it reduced the element of guesswork and justly rewarded different levels of understanding. In preparation for the tests, students practised writing proofs and tried to think critically about what they were writing. The feedback was useful, particularly the reasons given in the mark scheme, as this was the aspect of the test that students found most difficult.

19.4 Discussion, learning and impact

The outcome of the project is a novel multiple-choice based assessment on mathematical proof that tests comprehension, the ability to identify assumptions in an argument, and raises the appreciation of the importance of rigour. Students felt that this innovative assessment was a fair test of their understanding, but remained a challenging assignment on a topic that many students find difficult. The test is quick and easy to mark, but has maintained student engagement in, and learning from, the process of writing proofs. Moreover, it has raised awareness of common misconceptions and mistakes in mathematical writing. Therefore I feel that the project has not only achieved the intended outcomes, but that after a few modifications (which are discussed later) the assessment will have surpassed my initial expectations.

One of the strengths of the test is that it was based on previous students' work. This means that in the incorrect options the illogical assumptions and mistakes in notation and techniques are exactly the sort of mistakes that students make. Given that the students prepared for the test by writing out their own versions of the proofs, this assignment may have deeper educational merit than simply asking students to write some proofs, as in previous years. In fact, one student commented that "I have more of an idea than if you'd tested in a different way". For this test to work best, the provision of questions and examples prior to the test is critical, which was highlighted in the questionnaires as being particularly useful. The other crucial thing to provide is detailed feedback.

From the questionnaires and interviews it is clear that the students found the test an interesting, enjoyable and rewarding learning experience. The test has helped to raise awareness of the importance of rigour in mathematics, particularly when writing proofs, which should lead to improvements across the whole curriculum. Students mentioned that the whole experience has made them more confident at writing mathematical proofs.

The marks for the (more difficult) second test were encouraging; the mean mark was 54% with a standard deviation of 14. It should be noted that the second test took place three months after the work on proof had been completed. Also, the students suspected that their mark would not count much, if at all, towards their mark in the module, and it was discovered that the test took place on the same day as another test. Therefore, taking these things in account, it is hoped that in future the marks for this test may be a little higher.

Initially it was intended that the test would be an e-assessment so that both marking and feedback would be immediate. However, it has been noted that "e-learning systems are poorly adapted to mathematics" (Smith and Ferguson, 2005: 1). The trouble is that mathematics has its own language, and virtual learning environments are unable to adequately support the necessary mathematical notation and diagrams (ibid.); this was certainly my experience. The University has a virtual learning environment called NOW, which provides access to a variety of tools to enable the creation of e-learning tools. However, when trying to use the e-assessment tool, which can be used to create multiple-choice tests, it was found to be useless for mathematics. Within NOW it is impossible to input any symbols other than those found on the keyboard, and the options for formatting text are very limited. Therefore, in the virtual learning environment, it was not possible to directly type up a well-structured mathematical proof.

However, as it is possible to import pictures into NOW, I created a question and its four options as separate .pdf files, which were then converted into .jpeg and imported. However, this process is cumbersome and time-consuming, and although pictures could be added, it led to issues with pagination and alignment. It was obvious that the technology available was unsuitable for the development of the multiple-choice test on mathematical proof.

One obvious practical issue with the implementation of e-assessment as a summative piece of work for a large class is that more than one computer room would be required to run the test. Also, most people, such as the student who used his highlighter to find differences, would probably prefer to have a paper copy of the test as it is much easier to spot mistakes reading from a printout rather than from the screen.

Now that the assessment has evolved from a simple multiple-choice test it makes implementation through e-assessment impossible: a computer cannot judge whether a reason is correct or not unless it matches pre-programmed permissible phrases.

The development of this novel assessment has been a rewarding experience, and despite the initial cost in terms of time for development and implementation, that time will quickly be recovered in subsequent years due to the huge reduction in marking time. Moreover, the time for implementation and development has been incredibly worthwhile since the product is an excellent assignment that will help with the development of mathematics students for years to come.

19.5 Further development and sustainability

The students mentioned a variety of ideas for improvement and development, many of which coincided with my own. Some of these ideas will be easy to implement, such as slightly changing the rubric to clarify that a correct answer does not require justification. Another simple change is to allow slightly more time for the test. Perhaps somewhere between 50 to 60 minutes would be sufficient, which would mean that the test would still fit into a single lecture slot. It will also be easy to number each line to help students to pinpoint where illogical statements occur, although obviously a valid reason will still need to be given.

One student commented that they did not think each question should be worth the same mark and another suggested writing the answers in the booklet at the end of each question. However, for ease of marking it is felt that neither of these ideas is suitable for implementation.

Another easy improvement would be to place an identifier on each answer sheet to allow them to be detached from the question paper. This will also make the process of sorting the papers into groups easier. However, this sorting process will be unnecessary after the implementation of the following idea.

Students highlighted the mark scheme as invaluable feedback as it included reasons why certain options were incorrect. To ensure that all students engage with this feedback the proposal is to make the assessment into self-assessment, with the answers and marking process forming a discussion within seminars. Not only will this reduce the marking time to a few hours, but it will engage all of the students in a period of reflection by exploring in detail why certain answers were incorrect. It will give students the opportunity to clarify their thoughts, and staff the opportunity to highlight further the popular misconceptions.

In summary, the test will be developed by:

- adding clarification to the rubric;
- numbering each line to aid precise reasoning;
- allowing slightly more time;
- adding an identifier to each answer sheet to allow detachment;
- using self-assessment and discussion within seminars.

As e-learning tools can enhance the learning for those students for which e-learning is a positive experience (O'Regan, 2003), when e-learning systems are available that are well-adapted to mathematics, it would be beneficial to develop multiple-choice tests that concentrate on writing mathematics, but through different subject material. These tests (with no reasoning required) would be well-suited for inclusion in a series of formative e-assessments; formative rather than summative so that the issues with the original multiple-choice test on proof, such as fairness and to some extent the effect of guessing, do not apply. They would be tools for the students to use to guide and inform their own learning and development by reminding and encouraging them to think critically about what they are writing so that they do not slip back into bad habits and are not given the opportunity to forget.

Acknowledgements

I wish to thank Oana Radu for her help with marking the second test and with transcribing some of the student interviews.

References

Harvey, L. (2001) *Student Feedback: a report to the Higher Education Funding Council for England*. Accessed at `www0.bcu.ac.uk/crq/publications/studentfeedback.pdf`

Huxham, M. (2007) Fast and effective feedback: are model answers the answer? *Assessment & Evaluation in Higher Education*. 32(6), 601-611.

O'Regan, K. (2003) Emotion and e-learning. *Journal for Asynchronous Learning Networks*. 7(3), 78-92.

Smith, G. G. & Ferguson, D. (2005) Student attrition in mathematics e-learning. *Australasian Journal of Educational Technology*. 21(3), 323-334.

Yorke, M. (2003) Formative assessment in higher education: moves towards theory and the enhancement of pedagogic practice. *Higher Education*. 45, 477-501.

Mapping University Mathematics Assessment Practices
Published 2012.
University of East Anglia
ISBN 978-1-870284-01-1