Uses of e-assessment and a partially-automated approach

Peter Rowlett

Sheffield Hallam University

28 April 2022

Introduction

Introduction

- E-assessment (or computer-aided assessment or computer-assisted assessment) comes with advantages and limitations.
- So do other assessment methods like coursework assignments and written examinations!
- First, we'll discuss assessment methods.
- Second, a partially-automated approach is proposed.
- Third, some investigation into whether this method works.
- Fourth, some detail about how this method can be used.

Summative assessment

- I'm going to focus on summative assessment.
- Some of what I have to say applies less if the goal is purely formative.
- (To the extent summative and formative can be disentangled.)

In theory

Learning, teaching and assessment

- In learning and teaching, we should consider what students *do*, because that is where learning happens.
- We encode desired behaviour in learning objectives, and teach to encourage it.
- Then summative assessment should act as a signal of what students should focus on, and a measure of whether learning has happened.

Validity

- Assessment should assess what it is intended to assess.
- That is, it should be possible to derive marks that are meaningful in terms of the desired learning behaviours.
- For example, if a student can perform well without necessarily engaging in the desired learning behaviour this may be a problem with validity.
- An assessment with low validity might be:
  - too narrow: fails to include what it is intended to assess;
  - too broad: relates to aspects not being assessed.
- Determining validity is tricky, and uses judgement.

Reliability

- Assessment should be objective and repeatable, including showing no bias between assessors.
- An assessment with low reliability:
  - is not fair;
  - does not serve its purpose well because it does not necessarily provide an accurate view of a student's learning.

Efficiency and practicality

- This is in terms of resources and time/workload for staff *and* students taking part.
- This includes the time and effort needed to set, complete and mark a task.
- You should value your time and effort, and your students'.
- Assessment should be no more burden than is necessary to satisfy the goal of measuring learning.
- Indeed, the burden should be proportionate to the weighting of an assessment.

Examples

- You teach proof, but a student gets a good mark in the exam by memorising proofs and does not understand how to write novel proofs. This may be a problem with *validity*.
- A project report is read by two markers who interpret the marking criteria differently and come up with very different marks. This may be a problem with *reliability*.
- A technique is assessed on coursework 1, then again on coursework 2, then again in the exam. This may be a problem with *efficiency*.

Assessment

- Clearly we should aim for an assessment instrument that has high levels of reliability and validity while being efficient and practical to operate.

Assessment

- Clearly we should aim for an assessment instrument that has high levels of reliability and validity while being efficient and practical to operate.
- Easy, right?

Assessment

- It may be that in practice an assessment can reach a threshold for reliability and validity, but with variety in both measures.
- Indeed, measures to increase reliability can decrease validity, and vice versa.
- Attempts to increase either may decrease efficiency via additional burden.

Example: open-ended coursework

- Perhaps a loosely-specified problem or task with quite wide limits for submission format.
- Could encourage students to select and apply appropriate techniques and communicate findings, performing a valid function in terms of the skills development goals of a maths degree.
- However, can have issues with reliability and marker agreement due to less-tightly-specified mark scheme and issues with academic misconduct.

Academic misconduct

- For example, copying from another in a way that cannot be fixed by attribution, or cooperating with another person in a way that obscures the origin of a piece of work (Seaton, 2019).
- A lecturer interviewed by Thomlinson, Robinson and Challis (2010) said it is "not clear what the real benefit is" of coursework, due to the problem of copying.
- These affect reliability. For example, an invigilated assessment might reveal that a student did not know the topic so well as a take-home coursework suggested they did.
- Iannone and Simpson (2012) report some departments moving away from coursework due to concerns about copying.

Addressing academic misconduct #1: written examination

- Virtually guarantees the work is the students' own.
- May reduce validity, as:
  - exams tend to use a more tightly-specified series of tasks with closed-form questions, so cannot easily assess a broader skills base;
  - timed exam conditions tend to prioritise speed and memory, which may not be the learned objectives for the task.

Addressing academic misconduct #2: e-assessment

- Individualised, automated assessment can mean each student is given a unique set of questions, either through
  - *random selection* from a question bank;
  - *random generation* – the use of pseudo-randomised parameters in a question template.
- Individualisation can reduce opportunities for a student who might consciously or unconsciously engage in copying and collusion.
- Students with similar questions discussing method - positive!
- Doesn't address other aspects of academic misconduct, including issues like contract cheating, when someone asks another person to do the work for them (Seaton, 2019)

E-assessment: question generation

- E-assessment also carries other disadvantages.
- Random selection may offer limited range of questions.
- Random generation questions can be difficult to write, requiring expertise unlike the setting of paper tests:
  - technically, because of the need to understand how an auto-marking system will handle a response;
  - pedagogically, because of the need to more tightly specify what is to be tested, to avoid introducing extra learning requirements, and to understand typical student mistakes;
  - it is also possible to generate mathematically impossible questions.
- These issues affect efficiency, as well as potentially affecting validity by reducing the range of what can be asked.

E-assessment: answer input

- Validity is also affected by difficulty communicating answers to the computer.
- Some questions provide responses for the student to select, including multiple choice, giving a hint or the opportunity for guesswork.
- Others require input of answers.
  - Numeric input is of limited use, and simple string matching is inadequate.
  - Better to have free-text input which is tested for algebraic equivalence using a computer algebra system.
- Such input requires practice for students to use correctly and may add additional requirements unrelated to the learning objectives.

E-assessment: automated marking

- Automated marking avoids human error and lack of objectivity, potentially improving reliability provided systematic marking errors can be avoided.
- The need to write questions a computer can mark tends to focus on procedural aspects, potentially reducing validity.
- There can be issues with partial-credit:
  - a single answer is not usually enough to establish partial-credit;
  - a question broken into parts may be giving hints or focusing more on procedural aspects.
- Students are reported as preferring human-marked work because humans can act flexibly.

Summary: E-assessment

- E-assessment is a powerful, objective and efficient assessment tool.
- Use it to assess procedural aspects, especially for large numbers.
- But be aware of the disadvantages too (just like all assessment methods!).
- Also, it may be better for formative or low-stakes assessment.

Proposal: a partially-automated method

- It is possible to largely unpick the advantages and limitations of summative e-assessment regarding copying and collusion.
- Advantages around copying and collusion come from individualisation, which results from *automated question generation*.
- Limitations, however, largely arise from *automated marking*:
  - focusing on procedural elements and not being able to assess complex, open-ended work;
  - difficulties setting questions which anticipate marking limitations;
  - difficulties inputting answers.
- If we separate automated question generation from automated marking, we might access the advantages of the former without the limitation of the latter.

Proposal: a partially-automated method

- Questions set via an automated question generator.
- Student work completed and marked by hand as if it were a non-automated piece of coursework.

Proposal: a partially-automated method

- Questions set via an automated question generator.
- Student work completed and marked by hand as if it were a non-automated piece of coursework.
- This could access the chief advantage of individualisation, while avoiding the major limitations.

Proposal: a partially-automated method

- Questions could be more open-ended because there are not automated marking limitations – setting is therefore comparable to traditional coursework.
- Though care would need to be taken to generate comparable tasks for each student.
- The advantage of algorithmic objectivity in marking would be lost, but should be no worse than traditional coursework.
- A fresh limitation is that individualised work will be less efficient to mark than traditional coursework, because each student has a different set of answers that cannot be memorised by the marker.

Proposal: a partially-automated method

- This potentially offers a third approach to decreasing the potential for academic misconduct in coursework with high validity and concerns about academic misconduct.
  - Exam:
    - greater potential reliability;
    - less potential validity, due to timed conditions and high-stress environment.
  - E-assessment:
    - randomisation of question parameters allows individualisation;
    - less potential validity due to restraints of automated-marking and user interface.
  - Partially-automated method may:
    - maintain validity of coursework;
    - add reliability with respect to academic misconduct from e-assessment;
    - at a cost of reduced efficiency.

Why?

- We don't just do things to students because we are enthusiastic about trying them (we shouldn't, anyway).
- Is there a context in which this approach could be more useful than existing methods?

Teaching and learning context

Module

- Final year module.
- Focused on project-based learning and skills development.

Project

- Group project: answering a brief from a fictional client.
- Aiming to assess learning objectives around:
  - problem-solving in a real-world context;
  - working in depth over an extended period;
  - communication via reports;
  - group working.
- Group project brief: three similar problems, asked for solution interpreted in a real-world context.
- Individual assignment: A fourth similar problem, same question.

Individual assignment

- Used individual work alongside group work to increase the amount of the module mark that reflected individual ability.
- Similar to group project
  - to assess the same learning outcomes (except group working);
  - students who contributed well to the group project would be advantaged in the individual work.

Individual assignment

- Similarity to group work meant the risk of in-team copying or collusion was high — suggests need for exam conditions or individualised work.
- Nature of the task didn't suit exam conditions or e-assessment.
- Ideal for a partially-automated approach?

Evaluation method

Aims

- Partially-automated method was proposed as maintaining validity of an open-ended coursework while increasing reliability (with respect to academic misconduct).
- Should also check it behaves like an open-ended coursework in other ways, especially reliability (with respect to marker consistency).
- Evaluation aims:
  - Are the marks particularly sensitive to who is doing the marking?
  - Check validity, that the assessment is assessing the intended learning objectives.
  - Does the method contribute to reduction of copying and collusion?

Evaluation process

Phase	Details		Evaluation target
Second marker experiment	Reference	Written test, exam conditions Coursework, open-ended	Calibrating expectations re. marker consistency (ICC*)
	Main	Second-marking of individualised coursework	Marker consistency (ICC) Validity (professional judgement)
Student feedback	Questionnaire	Group A, who took the summative individualised coursework described	Student experience Risk of copying/collusion.
		Group B, who used the same individualised assessment technique for formative work at a different institution	Comparison group to detect innovator influence
Comparison of marks	Intra-group individual marks variation Raw group marks and individualised coursework marks		Occurrence of copying/collusion

*intraclass correlation coefficient

Results

Second-marker experiment (reliability) — summary

- Written examination (two markers): high level of agreement (ICC 0.992).
- Coursework reference experiment (two markers): lower level of agreement (ICC 0.586).
- Individualised coursework (four markers): between the two, closer to the coursework (ICC 0.635).

Second-marker experiment (validity) — summary

- All second-markers identified intended learning outcomes.
- No second-marker proposed additional outcomes.

Student feedback — summary

- "I disliked having different questions because I wanted to work together with another student on our answers": general disagreement;
- "I liked having different questions because it meant I could freely discuss the work with others with no risk of plagiarism": general agreement;
- "I liked having different questions because it meant that no one could copy from me": general agreement;
- "If we had been set identical questions, members of our group would have copied answers from other students: general agreement.
- Reference group: apparently similar.

Student feedback — summary

- "While at university, I have copied work from other students":
  - Group A: Yes, 22; No, 5;
  - Group B: Yes, 5; No, 11.
- "While at university, other students have copied work from me":
  - Group A: Yes, 35; No, 7;
  - Group B: Yes, 11; No, 5.
- Questions accompanied by a reminder that the survey was anonymous.

Comparison of marks

- The identified risk relates to intra-group copying or collusion.
- Individual marks within groups were examined:
  - variety of marks could indicate collusion is not a big problem;
  - lack of variety may be collusion, or normal close working and shared understanding.
- Individual marks within groups represent a range of 23–31 marks, standard deviations between 8.2–11.4.
- Also, comparison of individual and group marks:
  - similarity could indicate collusion.
- Group and individual marks and rankings do not correlate well ($\rho = 0.230$; $\tau = 0.229$).

Conclusions

- The partially-automated assessment was
  - of comparable reliability (with respect to who did the marking) to an open-ended piece of coursework;
  - valid, in the sense of assessing what it was intended to assess and no more.
- Copying and collusion were confirmed as risks and found to not have been a large problem, suggesting a reduction in academic misconduct.

Conclusions

- The partially-automated assessment appears capable of adapting a coursework assignment to make it less sensitive to copying and collusion while maintaining its validity, though leading to a reduced efficiency for the marker.
- This configuration of advantages and disadvantages is different to other approaches.

Implementations

Some special cases

- Data: give each student the same questions but an individualised data sample drawn from a larger data set.
- Computer algebra: give students the same questions which all involve some random parameter, and give each student a different value of the random parameter.

Python

```
f.write("\\[\\mathbf{P} = "+"{}\\]\n\n".format(latex(P,round)))

f.write("(c) Eigenvalues and eigenvectors:\n\n")
Peig = P.eigenvects()

# eigenvalues
f.write("\\[ \\lambda =
"+"\\left({},{}\\right)\\]\n\n".format(Peig[0][0].round(4),Pei
g[1][0].round(4)))
# eigenvectors
f.write("\\[\mathbf{v_1}"+"={}\\]\n\n".format(latex(Peig[0][2]
[0].evalf(3))))
f.write("\\[\mathbf{v_2}"+"={}\\]\n\n".format(latex(Peig[1][2]
[0].evalf(3))))

f.write("(d) 100 customers after 3 steps:\n\n")

s0 = Matrix([[75],[25]])
s3 = P**3*s0

f.write("\\[\mathbf{P}^3\mathbf{s}_0"+"={}\\]\n\n".format(late
x(s3.evalf(3))))

f.write("(e) Long term:\n\n")

if Peig[1][0] == 1:
    sn = 100*Matrix([[Peig[1][2][0][0]/
    (Peig[1][2][0][0]+Peig[1][2][0][1])],[Peig[1][2][0][1]/
    (Peig[1][2][0][0]+Peig[1][2][0][1])]])
elif Peig[0][0] == 1:
    sn = 100*Matrix([[Peig[0][2][0][0]/
    (Peig[0][2][0][0]+Peig[0][2][0][1])],[Peig[0][2][0][1]/
    (Peig[0][2][0][0]+Peig[0][2][0][1])]])
```

1 Question 2 - uses data from Customer 1 for jam

(a) Transition counts from data:

		Now Chapman's Strawberry Jam (£2.60)	Kolmogorov's Raspberry Jam (£2.90)
Next	Chapman's Strawberry Jam	32	18
	Kolmogorov's Raspberry Jam	18	35

(b) Resultant transition matrix:

$$\mathbf{P} = \begin{bmatrix} 0.64 & 0.3396 \\ 0.36 & 0.6604 \end{bmatrix}$$

(c) Eigenvalues and eigenvectors:

$$\lambda = (0.3004, 1.00000000000000)$$

$$\mathbf{v_1} = \begin{bmatrix} 0.707 \\ -0.707 \end{bmatrix}$$

$$\mathbf{v_2} = \begin{bmatrix} 0.687 \\ 0.728 \end{bmatrix}$$

(d) 100 customers after 3 steps:

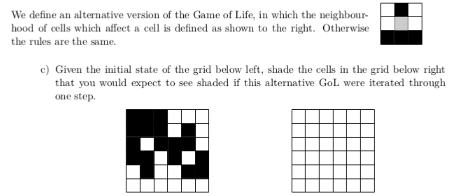
$$\mathbf{P}^3\mathbf{s}_0 = \begin{bmatrix} 49.3 \\ 50.7 \end{bmatrix}$$

(e) Long term:

$$\mathbf{s}_n = \begin{bmatrix} 48.5 \\ 51.5 \end{bmatrix}$$

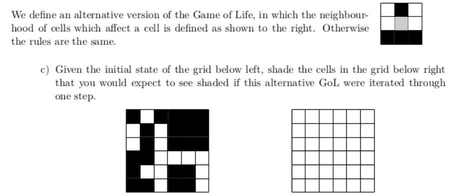
LaTeX

- A colleague used this approach, creating separate TikZ graphics for each student — here is an example question.



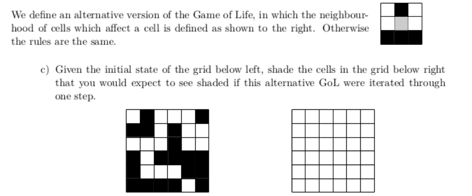
We define an alternative version of the Game of Life, in which the neighbourhood of cells which affect a cell is defined as shown to the right. Otherwise the rules are the same.

c) Given the initial state of the grid below left, shade the cells in the grid below right that you would expect to see shaded if this alternative GoL were iterated through one step.



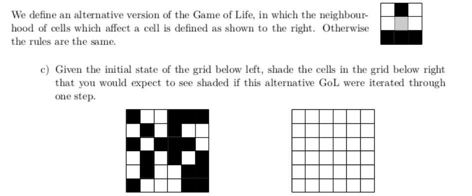
We define an alternative version of the Game of Life, in which the neighbourhood of cells which affect a cell is defined as shown to the right. Otherwise the rules are the same.

c) Given the initial state of the grid below left, shade the cells in the grid below right that you would expect to see shaded if this alternative GoL were iterated through one step.



We define an alternative version of the Game of Life, in which the neighbourhood of cells which affect a cell is defined as shown to the right. Otherwise the rules are the same.

c) Given the initial state of the grid below left, shade the cells in the grid below right that you would expect to see shaded if this alternative GoL were iterated through one step.



We define an alternative version of the Game of Life, in which the neighbourhood of cells which affect a cell is defined as shown to the right. Otherwise the rules are the same.

c) Given the initial state of the grid below left, shade the cells in the grid below right that you would expect to see shaded if this alternative GoL were iterated through one step.

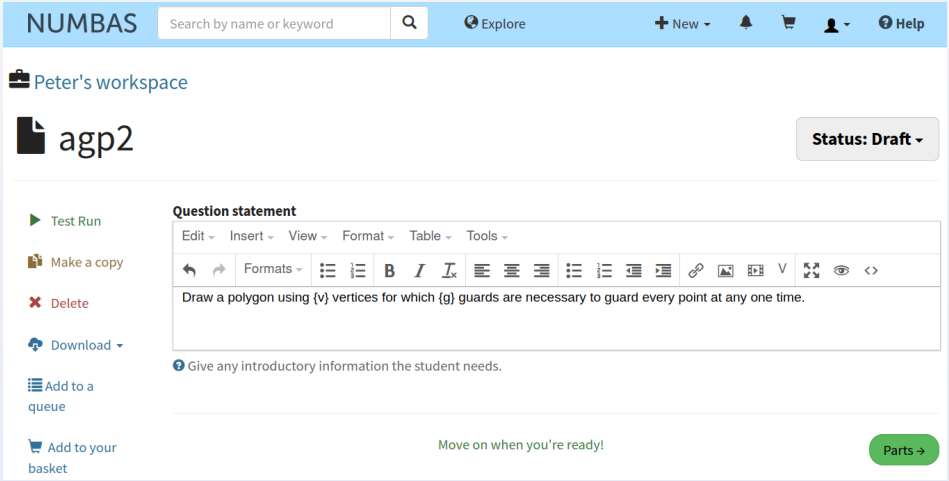
Assessment-setting system

- Used Numbas, a mathematically-aware e-assessment system developed at Newcastle University.
- Modified by its developer[†] to provide printable question and answer sheets.

[†] Christian Lawson-Perfect; for which, thanks.

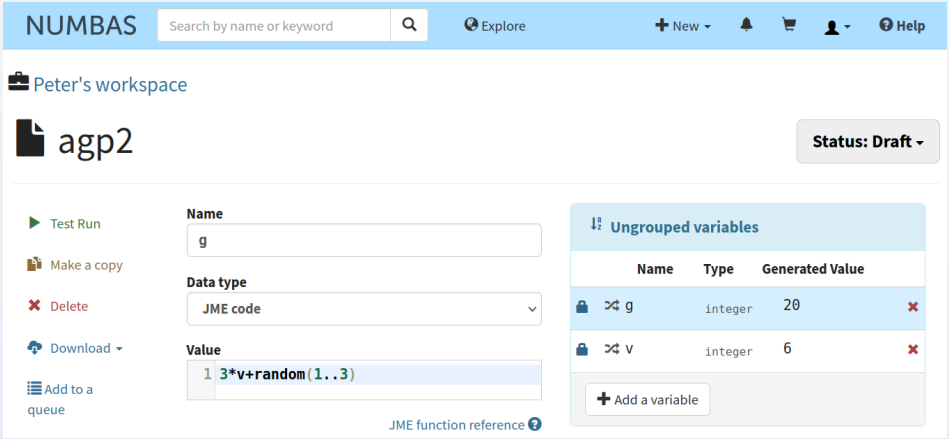
Numbas

numbas.mathcentre.ac.uk



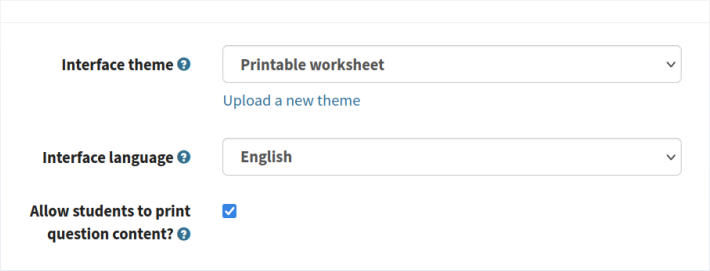
Numbas

numbas.mathcentre.ac.uk



Numbas

numbas.mathcentre.ac.uk



Numbas

numbas.mathcentre.ac.uk

PREVIEW of agp in Peter's workspace © All rights reserved Edit Download Share

Generate 1 exam, please (starting with ID no. 0).

1. Done

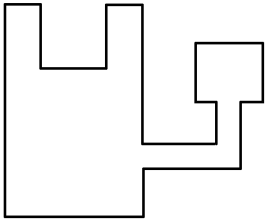
Now print

◉Question sheets

○Answer sheets

ID: 0

Question 1



a) Show, by triangulating and three-colouring the polygon, how many guards are necessary to guard every point in the museum shown above at any one time.

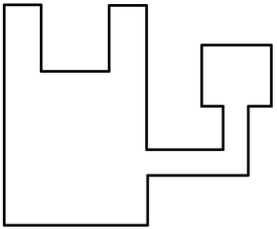
b) In reality, given a staff of 9 guards, could you arrange for every point in the museum shown to be guarded 24 hours a day and seven days a week?
If so, how would you arrange this?

Numbas

numbas.mathcentre.ac.uk

ID: 1

Question 1



a) Show, by triangulating and three-colouring the polygon, how many guards are necessary to guard every point in the museum shown above at any one time.

b) In reality, given a staff of 9 guards, could you arrange for every point in the museum shown to be guarded 24 hours a day and seven days a week?
If so, how would you arrange this?
If not, why not and how many staff would you require?

Question 2

Draw a polygon using 16 vertices for which 5 guards are necessary to guard every point at any one time.

Print — 1 sheet of paper

Destination: Save to PDF

Orientation: Portrait | Landscape

Pages: All

Colour mode: Colour

Fewer settings

Paper size: A4

Scale: Fit to page width / Scale 100

Pages per sheet

More details

- **Partially-automated individualized assessment of higher education mathematics**
- *International Journal of Mathematical Education in Science and Technology*
- https://doi.org/10.1080/0020739X.2020.1822554

INTERNATIONAL JOURNAL OF MATHEMATICAL EDUCATION IN SCIENCE AND TECHNOLOGY
https://doi.org/10.1080/0020739X.2020.1822554

Taylor & Francis
Taylor & Francis Group

Check for updates

Partially-automated individualized assessment of higher education mathematics

Peter Rowlett

Department of Engineering and Mathematics, Sheffield Hallam University, Sheffield, UK

ABSTRACT

A partially-automated method of assessment is proposed, in which automated question setting is used to generate individualized versions of a coursework assignment, which is completed by students and marked by hand. This is designed to be (a) comparable to a traditional written coursework assignment in validity, in that complex and open-ended tasks can be set with diverse submission formats that would not be suitable for written examination or automated marking; and, (b) comparable to e-assessment in terms of reduction of academic misconduct, with individualization acting as a barrier to copying and collusion. This method of assessment is implemented in practice. Evaluation focuses on expert second-marking, student feedback and analysis of marks, and aims to establish that the partially-automated method can be useful in practice. The partially-automated method proposed appears to be capable of adapting a coursework assignment to make it less sensitive to copying and collusion (and therefore more reliable) while maintaining its validity, though leading to reduced efficiency for the marker. This paper therefore contributes the introduction of a novel approach to assessment which offers a way to bring automated individualization to the assessment of higher order skills in higher education mathematics.

ARTICLE HISTORY
Received 15 November 2019

KEYWORDS
Partially-automated assessment; assessment; e-assessment; computer-aided assessment; skills

2010 MATHEMATICS SUBJECT CLASSIFICATIONS
97D60; 97U50

1. Introduction

There are a number of assessment methods available for higher education mathematics. One that has been popular in recent decades is automated assessment, called e-assessment, computer-aided assessment or computer-assisted assessment. There are advantages in assessing higher education mathematics via automated methods, but also limitations, which is a matter of debate in the professional and research literature (discussed later). Non-automated methods, such as coursework assignments and written examinations, also come with advantages and limitations. A model is proposed for viewing assessment methods as offering a balance of advantages and limitations suitable for different assessment circumstances, particularly in relation to validity, reliability and efficiency.

CONTACT Peter Rowlett p.rowlett@shu.ac.uk http://twitter.com/peterrowlett
https://www.linkedin.com/in/peterrowlett

References

Iannone, P. & Simpson, A. (2012). A survey of current assessment practices. In P. Iannone & A. Simpson (Eds.), *Mapping university mathematics assessment practices* (pp. 3–15). Norwich, U.K.: University of East Anglia.

Seaton, K. A. (2019). Laying groundwork for an understanding of academic integrity in mathematics tasks. *International Journal of Mathematical Education in Science and Technology*, *50*(7), 1063-1072. doi: 10.1080/0020739X.2019.1640399

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

Thanks for listening!

Any questions?

Uses of e-assessment and a partially-automated approach

Peter Rowlett

28 April 2022