Teaching Mathematics and Its Applications (2015) **34**, 117–137
doi:10.1093/teamat/hrv012 Advance Access publication 30 July 2015

Effective computer-aided assessment of mathematics; principles, practice and results

MARTIN GREENHOW[†]

Department of Mathematical Sciences, Brunel University, Uxbridge, Middx, UB8 3PH, UK

[†]*Email: mastmmg@brunel.ac.uk*

[Submitted April 2015; accepted June 2015]

This article outlines some key issues for writing effective computer-aided assessment (CAA) questions in subjects with substantial mathematical or statistical content, especially the importance of control of random parameters and the encoding of wrong methods of solution (mal-rules) commonly used by students. The pros and cons of using CAA and different question types are discussed. Issues surrounding the selection and encoding of mal-rules are highlighted, especially for multi-choice and responsive numerical input questions. These generate mal-rule-specific feedback, the mal-rule used being deduced from the student's selection or input. Student answer file data from the use of over 800 questions and their embedding within an overall assessment regime is analysed and presented to show that this has had a very beneficial effect on the examination performance of a large cohort of first-year economics students in their mathematics module over the last 6 years. Question analysis of over 270,000 question attempts, identifying the most difficult/discriminating questions, shows that the questions are robust, valid and span an appropriate range of difficulties. The idea of underlying mal-rules is examined to see how far this explains this range.

1. Introduction

As predicted by Ridgeway *et al.* (2004), computer-aided assessment (CAA) now forms a significant part of many students' experience of higher education, especially for subjects with a significant mathematical/analytical content and especially at the lower levels (years 1 and 2) where assessments typically test more mechanistic skills and techniques. Several mature technologies now exist that can reach beyond their intended audience and discipline, meaning that CAA looks likely to be increasingly pervasive, for example in schools or for in-service training of numeracy skills for the general workforce. Such assessment should not only simply grade students, but also should promote learning, which means that effective questions will need to be based on sound pedagogic principles. This article discusses the processes of writing objective questions and exploiting the computer medium. Given its immediacy, it is clearly essential to write clear and full feedback. Moreover, questions should also include random parameters, so that many thousands or millions of realizations are generated by underlying encoded algorithms, the randoms being carried through to all parts of the question (stem, key, equations, figures, feedback). These parameters usually generate different numbers, and can also generate random words or question scenarios and are certainly not limited to the assessment of mathematical content. The question setter is then required to encapsulate the algebraic and pedagogic content of all questions of a certain class (or *question space*), a process quite unlike the setting of paper-based tests. This usually involves reverse-engineering the question from the answer's requirement to have certain characteristics, e.g., whole numbers and easy fractions. Algorithms for wrong answers (displayed in a multi-choice MC question, and behind the scenes for a responsive numerical input (RNI) question) should be encoded based on common mistakes or *mal-rules*.

Of course, even sound and robust CAA will only be effective if students engage with it. A practical strategy for encouraging this is discussed. Even with engaged students, the efficacy of CAA needs to be examined, both in terms of their perception of the tests and the actuality of their learning. Although the limitations of the *maths e.g.* CAA system used here are readily acknowledged, a strong case is presented for including CAA within the overall assessment strategy. A discussion of the question facility and discrimination is given based on a very large data set generated by a networked version of *maths e.g.* over many years. Although such evidence-based research is valuable in its own right, it begs the question 'what makes one question harder than another'. Surprisingly, the question types used in this study (mostly MC or RNI) have little influence on question facility, so this must therefore be due to the tested content itself. If we can understand, and hence predict, this, test setters (CAA or paper-based) might be able to set better assessments based on questions whose performance is known *a priori*. The paper ends by speculating on possible indicators of difficulty.

2. Writing effective questions for CAA

The process of writing effective questions for CAA is not entirely reductionist, but certainly a question author needs to be aware of very significant differences between paper-based questions that are designed to be marked by a human marker and those of CAA, and the possibilities that become available via CAA. See CAA (2002) for a general overview beyond assessment of mathematical topics. Simple translation of traditional questions into CAA will generally fail if one does not consider the following:

- A human marker can mark both objective and subjective questions, whereas a computer marking scheme can only handle objective questions. Thus, questions involving construction of a proof, or even a diagram, and those that ask for interpretation (along the lines of 'comment on the significance of your results') can only be mimicked by objective questions delivered by the *maths e.g.* CAA system at present: a proof might be laid out and students asked to identify where a mistake might lie (if any) or a multi-choice question (MCQ) might offer well-chosen (mis)interpretations for a student to select. This then is a rather passive form of assessment where one is asking if a student can recognize the correct response when he/she sees it rather than generating it him/herself. This is certainly a necessary skill, but it is not sufficient, and certainly not all we would aspire to in our students. It is worth noting that the STACK system does allow the testing of student-generated mathematical models in some simple cases and its wider use would be very useful, see Badger & Sangwin (2011).
- Human markers act very flexibly when faced with ill-posed or unanticipated student responses, whereas a computer cannot, at least not at runtime, but see Gwynllyw & Henderson (2012) for a discussion of the post-marking deliverable by the Dewis system. Thus, the setting of a CAA question will require much more clarity in its setting, including an absolutely clear specification of assumed and tested skills and the encoding of typical student mistakes or *mal-rules*. Such mal-rules are, at present, suggested by the structure of the question and the teaching experience of the

question author, and can be easily generated by evaluating the effect of one or more of the assumed skills being missing. These need not be very 'original' to be effective in fact. For example, a RNI question asking for the derivative of $\cos(ax)$ at $x = b$ should have at least the following answers encoded behind the scenes: $-a\ \sin(ax)$, $a\ \sin(ax)$, $-\sin(ax)$, $\sin(ax)$, $\cos(a)$, each option being evaluated with the calculator set to radians and degrees. The marking and feedback of this type of question is discussed below and such ideas feed into a measure of question facility proposed below.

- Synoptic questions that test multiple skills will generally need to be broken down into a series of smaller parts for CAA. The reason is simple: if a student gets a question correct, one can infer that he/she has skills A, B and C, but if he/she gets it wrong, then which skill is missing is hidden since the CAA system will have no access to the student's workings but record only the student's input(s). The necessary question atomization (also a feature of some examination board questions) is unfortunate since it tends to lead the student through a series of procedures in order, which tends to 'give the game away'.
- Coupled with the above is the need to know the purpose of the test. For mastery, a synoptic question response being correct or not might be all that is needed, but this would give little diagnostic information to the student or the teacher and would not be acceptable for formative or summative assessments where one would want to reward partial knowledge and encourage further learning with targeted feedback. Thus, an all-or-nothing mastery question would be valid, but the issue of fairness cannot be answered unless one knows the purpose of the test. I think it would be regarded as unfair by most students since an almost perfect solution with just a 'silly slip' would be marked the same as a completely wrong, or even a null, response.

An example *Consider the following question: 'Sketch the graphs of* $f(x) = x^2$ *and* $g(x) = x^{1/2}$. *Find the area of the lenticular shape between the graphs. Show that this area is twice that of the lenticular areas between f(x) and g(x) and the straight line* $h(x) = x$ *and explain why this is the case.'*

What skills are assumed here, and what skills are being tested? How many objective questions would this synoptic question form? How would you mark the graph sketching, including the identification of the intersection points of the functions? How would you test the student's recognition of the need to apply integration and formulate this problem correctly? How would you test their integration abilities and could you (would you want to) award partial credit? What percentage of the marks would you award to each part, including the significance of the fact that f(x) and g(x) are inverse functions?

Given the above limitations it is reasonable to ask 'Why bother at all with CAA?' A comprehensive and general overview of CAA, including guides and case studies, is given by JISC (2014). The main advantages that pertain to the assessment of mathematical content include:

- Providing students with immediate and very full feedback that comprises not only model solutions but also bespoke feedback, see question types below. This capitalizes on the students' current engagement with the task, especially if they get a question wrong. Our observations and questionnaire returns, see Gill & Greenhow (2006), show the unexpected result that students spend at least half of the time studying the feedback given to identify where exactly they went wrong. This contrasts strongly with the all too common experience of staff that even prompt (2–3 days' turnaround) written feedback on paper-based tests is both onerous and ineffective, even to the extent of being unread by students. This is probably the most significant effect of introducing CAA into a traditionally assessed course where it was apparent (also to the students themselves) that they were using this feedback as a learning resource in itself, rather than simply as a means to get marks. I had not anticipated this effect, naively assuming that teachers taught the material which could then be assessed. Of course what is missing here is any mention of the learning process, which I had assumed occurred in lectures. This certainly is not happening for the majority of students who only really learn when faced with a task that forces them to confront and rectify any weaknesses in their understanding. This is not to say they are entirely assessment-driven, but rather that mathematics (and possibly all other subjects) needs to be *done* to be *learned*.

- Exploiting the rich CAA medium that allows for graphics, video, audio and the ease with which students, especially those with special needs such as dyslexia, colour blindness and partial sightedness, can customize their screen settings of font sizes and colours, see the figures.
- Allowing teachers/lecturers to adopt a better role by removing the tedium of test setting, vetting and marking, feedback writing and delivery of marks/feedback to students. Moreover, for remaining paper-based tests or examinations, existing CAA questions can provide a start point for developing more synoptic questions.
- Simplifying the logistics of setting tests, particularly in-course or continuous assessment tests of large cohorts (up to several hundred in an HE context). Depending on the embedding scenario chosen (see below), CAA can remove many of the administration tasks such as timetabling, booking rooms, informing students, making arrangements for students with special needs, repeat tests for those unable to do the first test instance . . . and finally the whole process again for resit tests.

All of the above point very strongly to a blended assessment scenario where CAA complements, rather than replaces, more traditional forms of assessment (especially the traditional, timed, unseen examination). An analogy might be the case for running a car that offers flexible and effective transport, but only on a fairly small part of the 30% of the globe that is accessible by roads.

3. Authoring questions

The process of authoring one's own questions is time-consuming and difficult at both the pedagogic and technical levels. While it is true that packages such as Hot Potatoes and VLE quiz engines facilitate their setting greatly, the resulting questions are very limited since they contain no randomization within a question itself. The same comment applies to school-targeted packages, such as BBC BiteSize or MyMaths, where one simply draws on a large bank of fixed questions. This does not mean that learning via such packages will be ineffective, but rather that hard-wired questions seem to be a dead end in the long run. Fortunately, there are many much more advanced packages available, most of which span the school/university level: one could buy in to a publisher-based (commercial/closed) system, e.g., MyMathLab or Wiley Plus; use and/or extend a non-commercial (open/free) system, e.g., STACK, DEWIS, NUMBAS; or one could write one's own questions using a commercial CAA authoring system, e.g., Perception or Maple TA. Rather than go into the pros and cons of the above systems, I will outline desirable features that any advanced system should have.

Firstly, a CAA system should support the use of random parameters within a scripted question program. Thus one can generate, once and for all, all questions of its class authoring not simply a question, but a so-called *question space* that encapsulates the algebraic and pedagogic structure of the class. Thus one would, for example, ask for the derivative of $y = ax^b$ at $x = c$, where a, b and c are random parameters of the question. This would produce about 5000 question realizations with $-10 < a, b, c < 10$. Their choice cannot be left to chance in fact (obviously none should be 0 or 1) but rather this immediately forces the question setter to consider what skills are being tested and what are being assumed. Even at A-level, this could include a consideration of the student's handling of negative numbers and the subtraction of 1 when doing the derivative; for example, the derivative of $y = -5x^{-4}$ at $x = -2$ is significantly more prone to error than the derivative of $y = 5x^4$ at $x = 2$. It is even possible to quantify this by enumerating the number of mal-rule-generated accessible states for each part of the problem available to a student, multiplying them together and taking a logarithm to get the question's entropy—a concept well-known to physicists and discussed later. What is apparent here is that a generally written question should be cloned to produce questions that, while algebraically all equivalent, each produce pedagogically equivalent realizations at run time. A clone can be as simple as placing a standard formula on the screen, or not; clearly, the resulting questions are not the same, as the latter requires the realization that the formula is needed, its accurate recall and its application, while the former merely requires its application. We have observed that, usually after seeing a few realizations of the same question style, students realize that they have actually done the question before, i.e., they have abstracted the question to the higher level of mastery of an entire class of question. This is something that teachers want in their students and one that happens naturally with repeated practice, however delivered. It is worth mentioning that randomization is useful well beyond mathematics or numerical calculations in other disciplines and can include random words such as 'What is the molecular weight of <random chemical>?' or even 'Where are the mistakes in 'An affect is apparant here'?' (sic) where mistakes are randomly sprinkled into a sentence to be proof read.

A related point is that the values of the random parameters can significantly affect the difficulty of a question or even its solution method, e.g., integrate $1/(x^2 + 4x + c)$, where $c = 3$ (factorizable denominator) or 4 (perfect square denominator), which is part of an example by C. Sangwin (personal communication). One therefore generally needs to 'reverse engineer' the question from its solution. Teachers will be familiar with this when setting, e.g., questions on partial fractions starting with the desired factors of the denominator and small integer coefficients in the numerators. If not, a student is likely to be defeated by the sheer complexity of the fraction manipulations required, even though he/she knows very well what should be done to give the partial fractions. At a more advanced level, one clearly should not simply generate a 3 × 3 random matrix and ask for its eigenvalues and eigenvectors, since this would generally result in an unfactorizable characteristic equation and/or complex eigenvalues. Clearly, one needs to generate the matrix using single digit integers for all eigenvalues and all eigenvector elements.

An example *Set a question that requires the student to factorize a cubic starting with three linear factors $(ax + b)(cx + d)(ex + f)$. What range of parameters would you choose for each of the random letters a, b, c ... ? Taking your largest allowable (absolute) values, multiply out the cubic; is the question still reasonable? How many question realizations would this produce? Would you prohibit repeated factors? Given that negative coefficients are likely to give harder questions (do you agree?) how many cloned questions would you need and what would be their relative difficulties? Repeat the above with a linear factor and an irreducible (over the integers) quadratic factor, i.e., $(ax + b)(cx^2 + dx + e)$. How would you ensure the quadratic is irreducible?*

Secondly, a CAA system should be able to encode algorithms, not only just for the correct answer (the key) but also for incorrect answers (distracters) based on specific mistakes that are likely. These are called mal-rules (formerly called bugs in the development of intelligent-tutoring systems that were used to trigger a move down a particular branch in the system's responses). These mal-rules do, to a certain extent, interact with the choice of numbers (e.g., the square root of 99 is not 33 and a temperature change from -3.7° to 15.7° is not 12° although one can see why a student might produce such answers in these cases but not with different numbers). Mainly, however, mal-rules arise from a common error or errors often at a quite fundamental level, e.g., integrating $x^{-4/7}$ can trigger simple arithmetic errors when students cannot add 1 to $-4/7$. These are often dismissed as 'slips' but may say something more fundamental about the student's facility with, or even understanding of, the number line. Mal-rules are therefore essential in programming up distracters such as $(a + b)^2 = a^2 + b^2$ that would be used in expanding brackets questions like $(7x-3)^2$. Clearly, what is happening here is that students are incorrectly commuting the addition and squaring operations, so the mal-rule is easily categorized. However, categorizations spanning the whole range of likely mal-rules in, e.g., A level Mathematics are not sufficiently developed to give a useful taxonomy for analysing many hundreds or thousands of answer files, see Biggs & Collis (1982), Chick (1998), Hanson (2011), Haynes & Hermans (2007), Priem (2010) and Schechter (1994). The problem appears to be that for such taxonomies the generality of the categorization does not give enough detail to be useful for question designers, or, at the other extreme, their specificity is such that they cannot be applied to other questions, meaning that every question becomes a special case. Lacking an adequate structure, even for a limited range of mathematical levels or topics, means that while 'answers' undoubtedly lie in the vast amount of data considered below, we cannot yet ask sensible questions. For example, which overarching mal-rules are being applied across that level or topic and how do we understand, and eventually utilize, the underlying reasons for erroneous student responses?

Thirdly, the pros and cons of different question types are worth considering carefully:

- MCQs are the most obvious type of question—indeed some people think that CAA *is* MCQ. This has tended to give CAA a bad image because:

- MCQs merely ask students to spot the correct answer when they see it.
- They are very heavily 'scaffolded' with the key and distracters suggesting the form or size of the answer, e.g., not taking the square root of the variance would give a distracter that is unrealistically large for the standard deviation.
- Partial knowledge may eliminate some distracters and encourage guessing, for example by using symmetry (a good skill, but probably not the one intended to be tested by the question).
- Trial values such as $x = 0, 1, 2 \dots$ might be used until all but the correct answer is eliminated.
- The question might be done backwards, e.g., differentiating all the options rather than integrating the integrand in the question.
- And pattern spotting might be employed where the student looks for commonality in parts of the options. For example, what are the coordinates of a calculated point in space, with options (1,1,1), (1,1,0), (1,0,1) and (0,1,1) suggest that (1,1,1) is the correct answer. There are of course ways of combatting this by making two mistakes in some distracters.

Other very pertinent objections to the use of MCQs in the assessment of mathematical content are given by Sangwin (2013) while other authors (cited by Sangwin) have focused on possible bias against students with certain learning styles and confidence levels and even the possibility of a gender bias. These are certainly concerns that might affect the 'fairness' of assessments, especially high-stakes assessments, but here the main focus is on formative and low-stakes assessments. For this study, a more immediate question is whether the use of MCQs distorts the learning that is the main driver for setting the assessment in the first place. This is a difficult question that would require specific studies to answer; however, the data presented in Table 1 and discussed later show that the difficulty range of MCQs is mostly comparable to the other main type of question used, namely RNI (see below). In this sense, the well-designed MCQs used here, whatever their shortcomings, do test something and do not appear to be an easy option for the students.

As a general rule, MCQs are substantially improved by the consistent use of a 'None of these' option that should be the correct response at times. As well as guarding against question errors where an intended correct answer has an error, it also hinders attempts to do the question backwards. Paper-based tests show that this simple device also forces the student to check his/her working before committing to an answer, S. Hibberd (personal communication), and this develops the good habit of checking answers.

TABLE 1. *Summary of level 1 economics student activity at Brunel University from 2008–2009 to 2013–2014. In the last column, topics with a low number of MCQs are indicated*

Topic	Number of questions available	Question level	6-year total of tests	6-year total of questions	(R)NI facility range	MC facility range (number of questions)
Numbers	124	GCSE	3380	29,625	0.349–0.985	None
Algebra 1	108	AS	5241	42,345	0.478–0.944	0.324–1
Algebra 2	94	A	5009	31,938	0.053–0.978	0.253–0.974
Economics	53	subject-specific	4997	23,367	0.115–0.897	None
Differentiation	91	AS	3474	18,242	0.310–0.871	0.410–0.974
Mathematics of Finance	56	subject-specific	4835	32,905	0.145–0.933	0.729–0.967 (7)
Partial differentiation	67	Level 1 university	5295	27,941	0.132–0.876	None
Optimization	17	Level 1 university	3025	10,390	0.297–0.738	None
Integration 1	39	AS	2695	12,324	0.437–0.644	0.222–0.963
Integration 2	58	A	2639	11,546	0.439–0.939	0.474–0.866
Matrices 1	61	Further Maths A	2902	18,990	0.090–0.955	0.663–0.855 (4)
Matrices 2	41	Further Maths A	2754	13,230	0.349–0.833	None
Total	809		46,246	272,843		

From the authoring point of view, realistic mal-rule-based distracters for MCQs are generally difficult to create and need careful algebraic checking of the mal-rules to make sure that they produce unique options regardless of the choices of random parameters at run time (for example, 'What is a^2?' cannot have a distracter $2a$ if a can be 2 since there will then be two 4 s of the screen, one marked correct, the other not). What is more, a MCQ with four options, 'none of these' and 'I don't know' will actually require four distracters since they must all be displayed when 'none of these' is the correct response. If is often very difficult to come up with four realistic distracters that students will not be able to eliminate using any of the above methods.

For the assessment point of view, dichotomous marking of 1 for correct, else 0, means a non-zero average for the random guesser. Other marking schemes are possible, but a question author cannot decide that; it must be done by the test setter who will know the purpose of the test (mastery, summative, formative, diagnostic, etc.). It is notable that the DEWIS system allows this by the questions passing flag settings to any marking scheme specified by the tests setter.

Given the above, one might then ask 'Why bother with MCQs?'. Partly the answer is that they are robust and require only a mouse click to communicate what might otherwise be a long expression or a diagram (at present it is impossible to handle diagrammatic input from the student). Another 'soft', but very good, reason is to build students' confidence. How self-fulfilling it would be for a student who believes his/her mathematics is weak to get zero on a test comprising only numerical input (NI) questions? This is very unlikely with MCQs and will give students a sense that they do have some mathematics skills after all. This is especially valuable in early-stage diagnostic tests where the underlying mal-rules can, and should, trigger specific feedback (e.g., you have forgotten to use the chain rule) as well as the model solution. Such bespoke feedback is a major advantage of MCQs for formative assessment.

An example *Create an MCQ (stem, key and four distracters) for the derivative of cos(ax^2), with a random. How would your distracters change if the question were sin(ax^2) or tan(ax^2)?*

- NI questions avoid most of the problems associated with MCQs, and offer a much more acid test of the student's skills, including computational skills. However, they do have their own problems; first and foremost, it is impossible to offer bespoke feedback or partial credit—they either get the correct number (perhaps within a tolerance) or they do not. What is more, randoms can produce numbers so large or so small that calculators produce answers of insufficient or even unreliable accuracy, e.g., find the derivative of $\exp(x^2)$ when $x = 10$ to three decimal places. Even with calculable questions, one needs to check that a student has input a number and in acceptable form, e.g., two decimal places, three significant figures, in scientific notation and fully cancelled fraction. It is usual to issue a warning if a student inputs a correct answer in an incorrect format and then correct the format for him/her—unless the format (e.g., number of significant figures) *is* a skill being tested by the question.

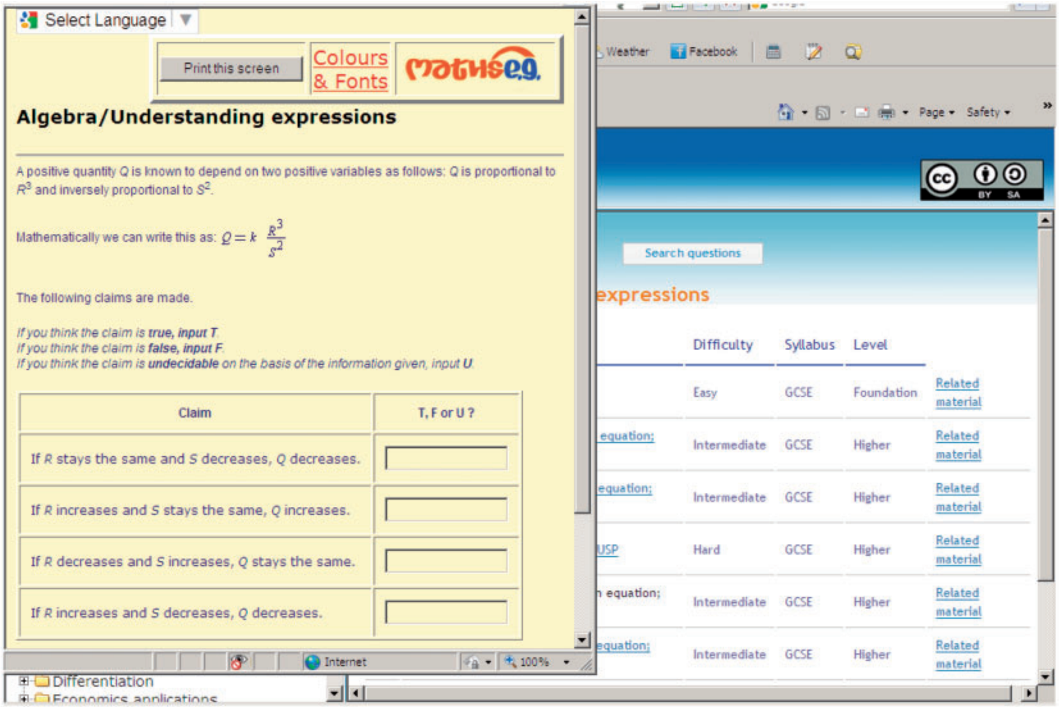
- NI questions should be superseded by RNI questions in my view. Such questions combine the better features of MCQ and NI questions by checking the student's answer, firstly against the correct answer, and then against the numbers generated by any number of distracters in turn until a match is found (issuing bespoke feedback) or no match is found (in which case the feedback can be 'I think you are guessing' which often surprises students into saying 'I was, but how did it know?'). These unscaffolded questions are an excellent precursor to more traditional forms of assessment, especially the end-of-module examinations.

Some systems offer free-form mathematical input which is verified before passing to a behind-the-scenes algebraic engine such as Macsyma or Maple, see Sangwin (2013). These supersede earlier, but quite robust string-evaluation questions used in the 1990s. While these systems require some student training and hence may be less useful in a school setting or for the casual user (e.g., a level 1 biology student), currently they are the gold standard for assessment of higher-level skills required by mathematics, physical sciences and engineering students. Setting of such questions is probably beyond the capacity of most occasional authors since one needs to constrain what is and is not an acceptable answer before giving it to the algebraic engine to check for algebraic equivalence with the question author's answer (usually generated by the symbolic engine itself).

- True/false/undecidable and assertion/reason questions are also worth considering. These can be randomized and Fig. 1 shows an example where the student is asked rather high-level questions about the meaning of an expression.
- Other question types also occur; any of the above questions can have a time limit; one can ask students how certain they are in their answers, with negative marks for deluded students who 'know' their incorrect answers are right! Questions can be multi-part, where the first input, even if wrong, can be used to calculate what the second or subsequent input(s) would have been, and partial credit awarded. Another question type is the revealed MCQ where students must accept or reject options shown one at a time, as in a 'game of sudden death' where a rejected option is never seen again. Finally, third-party resources, such as Excel or Geogebra can be made available in questions via buttons; this is especially useful for statistics questions where the length of the required calculations precludes hand calculation even with a calculator, see Fig. 2. This can be taken further to test students' facility with specific software, e.g., SPSS (see Gwynllyw *et al.*, 2015) or their skills in a variety of programming languages.

4. Embedding CAA within a module

At school level, the embedding of CAA within a mathematics course for summative assessment will be subject to obvious constraints, such as the availability of computer laboratory time and the assessment regime set by external bodies, such as examination boards. For formative assessment these constraints



Algebra/Understanding expressions

A positive quantity Q is known to depend on two positive variables as follows: Q is proportional to R^3 and inversely proportional to S^2.

Mathematically we can write this as: $Q = k\ \frac{R^3}{S^2}$

The following claims are made.

If you think the claim is ***true, input T****.*
If you think the claim is ***false, input F****.*
If you think the claim is ***undecidable*** *on the basis of the information given, input* ***U****.*

Claim	T, F or U ?
If R stays the same and S decreases, Q decreases.	□
If R increases and S stays the same, Q increases.	□
If R decreases and S increases, Q stays the same.	□
If R increases and S decreases, Q decreases.	□

expressions

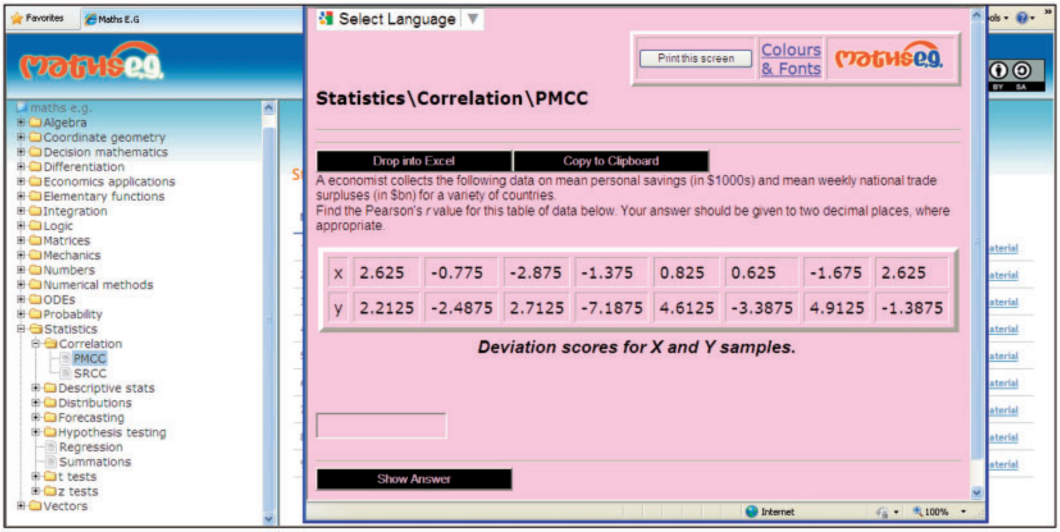
	Difficulty	Syllabus	Level	
	Easy	GCSE	Foundation	Related material
equation;	Intermediate	GCSE	Higher	Related material
equation;	Intermediate	GCSE	Higher	Related material
USP	Hard	GCSE	Higher	Related material
n equation;	Intermediate	GCSE	Higher	Related material
equation;	Intermediate	GCSE	Higher	Related material

FIG. 1. A true/false/undecidable question from the *maths e.g.* student interface. Note in the question window the Colours and Fonts option and Google's translator (to over 50 major languages) and the search facility and question tags in the main window. Note: randomized claims are built from random choices of subject/property (before/after the comma).

are much less binding and some systems allow guest login, allowing students to practice. In particular, *maths e.g.* has both student and teacher interfaces. Questions (available in either interface) have links to external web sites/resources. They are tagged to national syllabi and difficulty level (judged for questions of that topic, so that one can have hard fractions questions and easy matrices questions). These tags are used by the search engine. The teacher interface also allows for the creation of test by adding questions to a 'shopping trolley', see Fig. 3.

In an HE context, assessment is much more at the discretion of academics who may decide to embed summative CAA within a module. This can be done as invigilated tests, or mock tests followed by actual paper or CAA tests, or simply to allow 24/7 access to the tests and a limited number of attempts at each. We have found that an extremely effective scenario is to have staffed laboratory sessions where students do the tests and teachers (or postgraduate helpers) simply drift round and talk to them. This alters the teaching dynamic rather substantially; you are metaphorically and physically on the student's side in an 'us versus it' situation. This also exposes any conceptual misunderstandings in an effective way. However it is done, the rule I use is to take their best-ever mark from their first five attempts at any one of a suite of up to 12 tests per module. This gives students a real incentive seriously to attempt a very large number of questions—perhaps up to 600 per module.

Group work is actively encouraged so that weaker students can grow in confidence through discussions with their peers and during the necessary repeated practice when each of the group logs in and gets a similar, but certainly not identical, test. On the other hand, during uninvigilated tests, students



Statistics\Correlation\PMCC

A economist collects the following data on mean personal savings (in $1000s) and mean weekly national trade surpluses (in $bn) for a variety of countries.
Find the Pearson's *r* value for this table of data below. Your answer should be given to two decimal places, where appropriate.

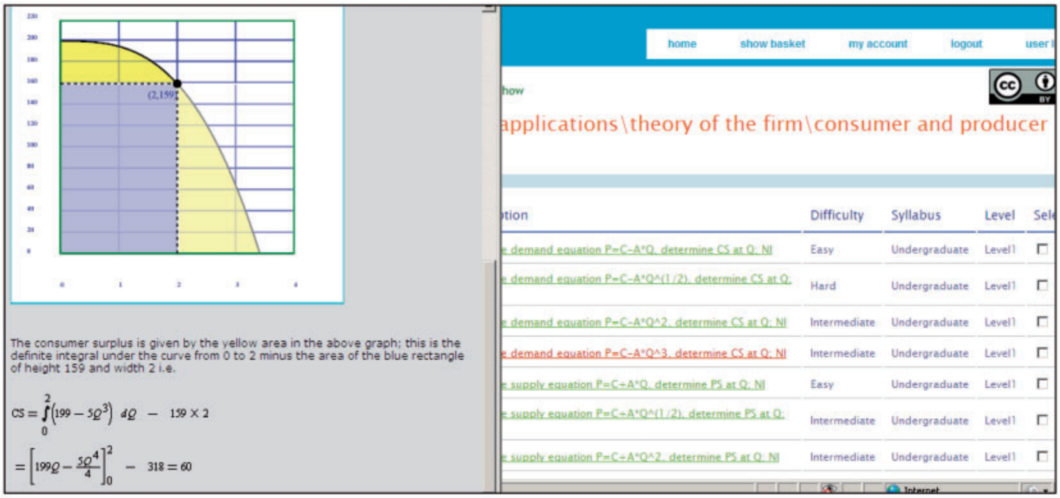
x	2.625	-0.775	-2.875	-1.375	0.825	0.625	-1.675	2.625
y	2.2125	-2.4875	2.7125	-7.1875	4.6125	-3.3875	4.9125	-1.3875

Deviation scores for X and Y samples.

FIG. 2. The *maths e.g.* student interface, showing the topic tree comprising circa 2000 questions and a question selected from Statistics showing Excel and Clipboard buttons.

can alias each other or even pay others to take the tests for them. This is definitely something that happens with other forms of assessment where copying is not rare and hiring a programmer from rent-a-coder web sites is not unknown, but I have no evidence that such unacceptable practices are happening for CAA (even by hearsay). The use of 'illegal' software is more worrying, given the power and free use of, e.g., Wolfram alpha and other symbolic engines. Fortunately, there is a clear-cut and fair way to provide a very strong disincentive for such cheating, namely they must pass the end-of-module examination with a genuine pass mark, rather than achieving a pass mark overall. This focuses the students on their learning right until the end of the teaching period so they do the tests to learn rather than simply accrue marks. Evidence that this is having the desired effect with a group of circa 200–300 Level 1 Economics students *p.a.* is given in Fig. 4, updated from Greenhow & Zaczek (2011). For the students, taking the tests is a significant part of the module and it is clear that one must replace, rather than duplicate, existing tests; over-assessment might encourage surface learning for the test rather than deep learning of the mathematics. Actually many students rather enjoy the process of building up a good set of marks, but they also realize that their mastery of the topics is rapidly improving as a result. What is more, one can even replace teaching of the rather more mundane or mechanistic aspects of a module by the CAA, provided it has sufficiently rich feedback screens.

It might seem at first sight that teacher/student contact is being diminished by the CAA (although one would have to establish how much real contact would occur in a traditionally delivered module with up to 300 students!). Surprisingly, students say that it actually increases contact with the lecturer by providing a link; students often email me or other teachers querying the validity of the marking of a question and/or bring up such queries at tutorials. They are almost always wrong in that the error is due to misreading the question or some misconception they have (not understanding the difference between AND and OR in logic and sets is common) but the teacher's response can clear this up, sometimes by return email, or, more usually, by additional teaching. Such teaching is also informed by analysis of the answer files, both on an individual and whole-cohort basis, where particularly difficult topics or especially discriminating questions are easily flagged up during the course delivery,



The consumer surplus is given by the yellow area in the above graph; this is the definite integral under the curve from 0 to 2 minus the area of the blue rectangle of height 159 and width 2 i.e.

$$CS = \int_0^2 \left(199 - 5Q^3\right)\, dQ \;-\; 159 \times 2$$

$$= \left[199Q - \frac{5Q^4}{4}\right]_0^2 \;-\; 318 = 60$$

applications\theory of the firm\consumer and producer

tion	Difficulty	Syllabus	Level	Sel
e demand equation P=C–A*Q, determine CS at Q: NI	Easy	Undergraduate	Level1	□
e demand equation P=C–A*Q^(1/2), determine CS at Q:	Hard	Undergraduate	Level1	□
e demand equation P=C–A*Q^2, determine CS at Q: NI	Intermediate	Undergraduate	Level1	□
e demand equation P=C–A*Q^3, determine CS at Q: NI	Intermediate	Undergraduate	Level1	□
e supply equation P=C+A*Q, determine PS at Q: NI	Easy	Undergraduate	Level1	□
e supply equation P=C+A*Q^(1/2), determine PS at Q:	Intermediate	Undergraduate	Level1	□
e supply equation P=C+A*Q^2, determine PS at Q: NI	Intermediate	Undergraduate	Level1	□

Internet

FIG. 3. The *maths e.g.* teacher interface showing the feedback for a NI question and test creation options.

allowing action to be taken, rather than being discovered only when examinations are marked, when it is too late of course.

5. Analysis of answer files

At Brunel University, the first-year Mathematics for Economics module produced some 12,000 answer files for each of the 6 years from 2008–2009 to 2013–2014 (no data for the current year is yet available). One can focus on individual student performance, for example how often they engage with a test and how this affects their CAA marks, questionnaire responses, examination marks, etc., see Gill (2007) and Zaczek (2015), but no corresponding study has been carried out for the module considered here. Here, we focus instead on the questions themselves. Since each is taken 100–500 times, depending on how many questions within each topic are available to randomly choose from, we can regard the measures discussed below as being typical for this cohort and not an artefact of small numbers of data points. Table 1 summarizes the data: it is immediately apparent that such high levels of student activity can only be supported via a CAA system and this provides a robust evidence base on which to develop our teaching and assessment regimes.

Figure 5 displays two measures of the question performance for selected topics, namely facility (0 to 1) and the product–moment correlation coefficient between the student's performance on the question and on the test overall (-1 to 1). It is seen that the difficulty ranges from about 0.2 (a hard question) to almost 1 (almost everyone got it correct), and that for the majority of questions the correlation (a measure of question discrimination or how well that questions stands as a proxy for the whole test) is uniformly high for all but the very easiest questions. This provides very robust evidence that the questions were well designed, since low or negative correlation usually indicates an incorrect or badly designed question, for example an obviously correct option in an MC question. The easiest questions, with facility close to one, provide no measure of the overall test performance and hence have low correlations, despite being correctly designed. It could be argued that they should be omitted but then one is again forced to consider the purpose of the test. I argue that it is no bad thing for some easy questions in the most basic topics to be included so that weaker or more maths-phobic students are able to build their confidence.

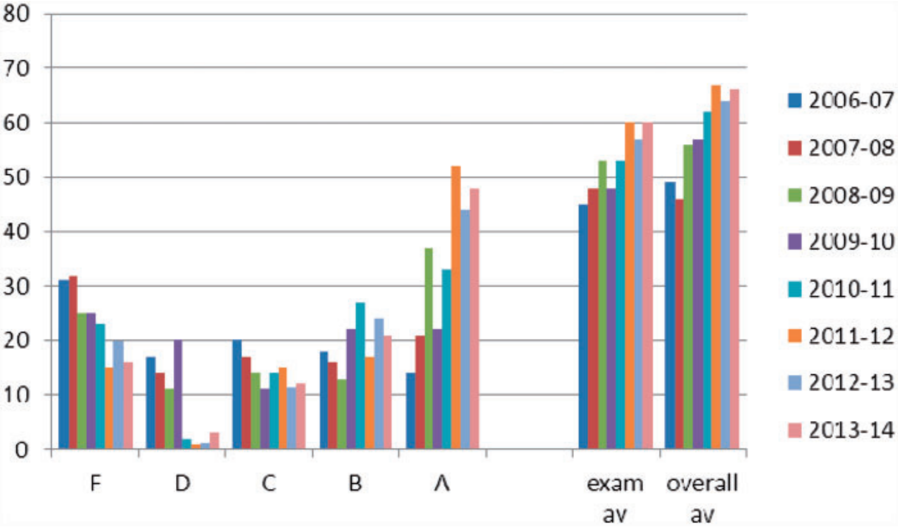


FIG. 4. Year-on-year overall grades (% of students), and examinations and overall averages (% marks). CAA was introduced in the 2008–2009 academic year. The reason for the performance increase in 2011/2012 is unknown; nothing had changed with the module and neither had the admissions criteria (see Appendix).

These data could be used to address many questions, for example the effect of question type certainly could play a role. However, somewhat surprisingly, no strong message comes out of the data comparisons of facility between MC and (R)NI questions, see Table 1. Notice that some topics, by their nature, had only NI questions while others had very few MC questions so comparison of the performance of question types is only possible for the other topics where numbers of questions are roughly comparable.

6. Possible measures of difficulty

Only in a few clearly identifiable circumstances is the facility an artefact of the question type or marking rules (for example, zero mark is given if any of the submitted answers is wrong in a question requiring several inputs or selections). An example occurs in Numbers, where the hardest question requires five inputs so is discounted below. Thus, we are forced to consider the actual content of the questions in order to understand in any way their facilities. This is obviously a long procedure unless some sort of systematic and hence possibly automatic measure of difficulty is found. For now, we consider only the hardest/easiest questions in each topic; spanning revision topics at GCSE level to completely new topics at university level, this gives a good test bed for ascertaining the extent to which the difficulty measures are generally applicable across a range of mathematical topics.

The most obvious is simply the operations count widely used to deduce the order of an algorithm in programming. In mathematics, however, this does not seem justifiable; for example, if a student could add three numbers together, then adding a further two should not make it harder in any significant way (apart from 'silly slips'). It would also be quite difficult to identify how many operations are involved in rather more advanced mathematics, for example, finding the value of an integral, so we do not pursue this further here.

As mentioned above, the concept of entropy might provide a suitable measure if one can identify the number of accessible states. This is explained in detail for Numbers, but it appears to provide only a way

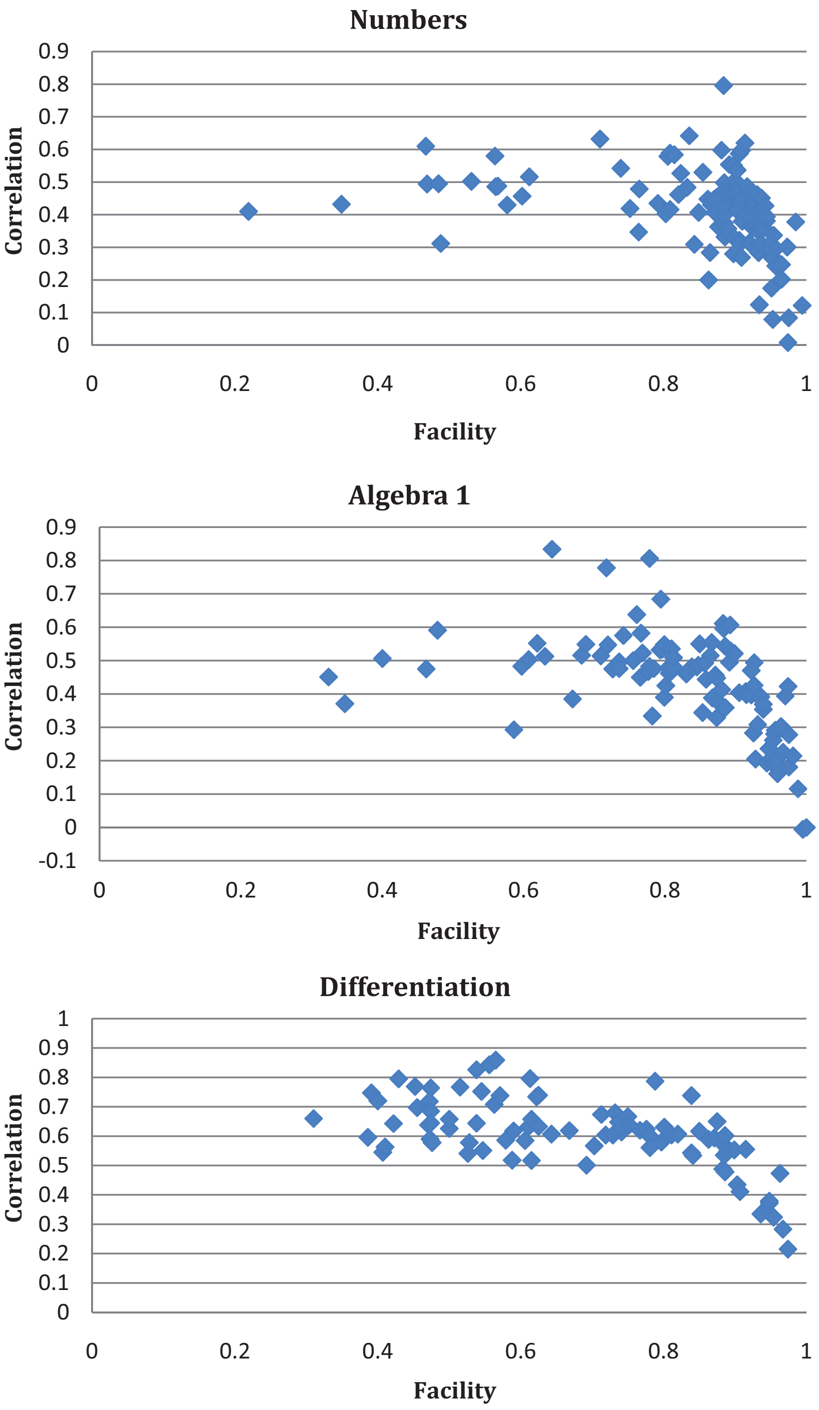


Fig. 5. Correlation versus facility for selected topic areas. Note that some markers overlap; the higher number of question in each topic is given in Table 1.

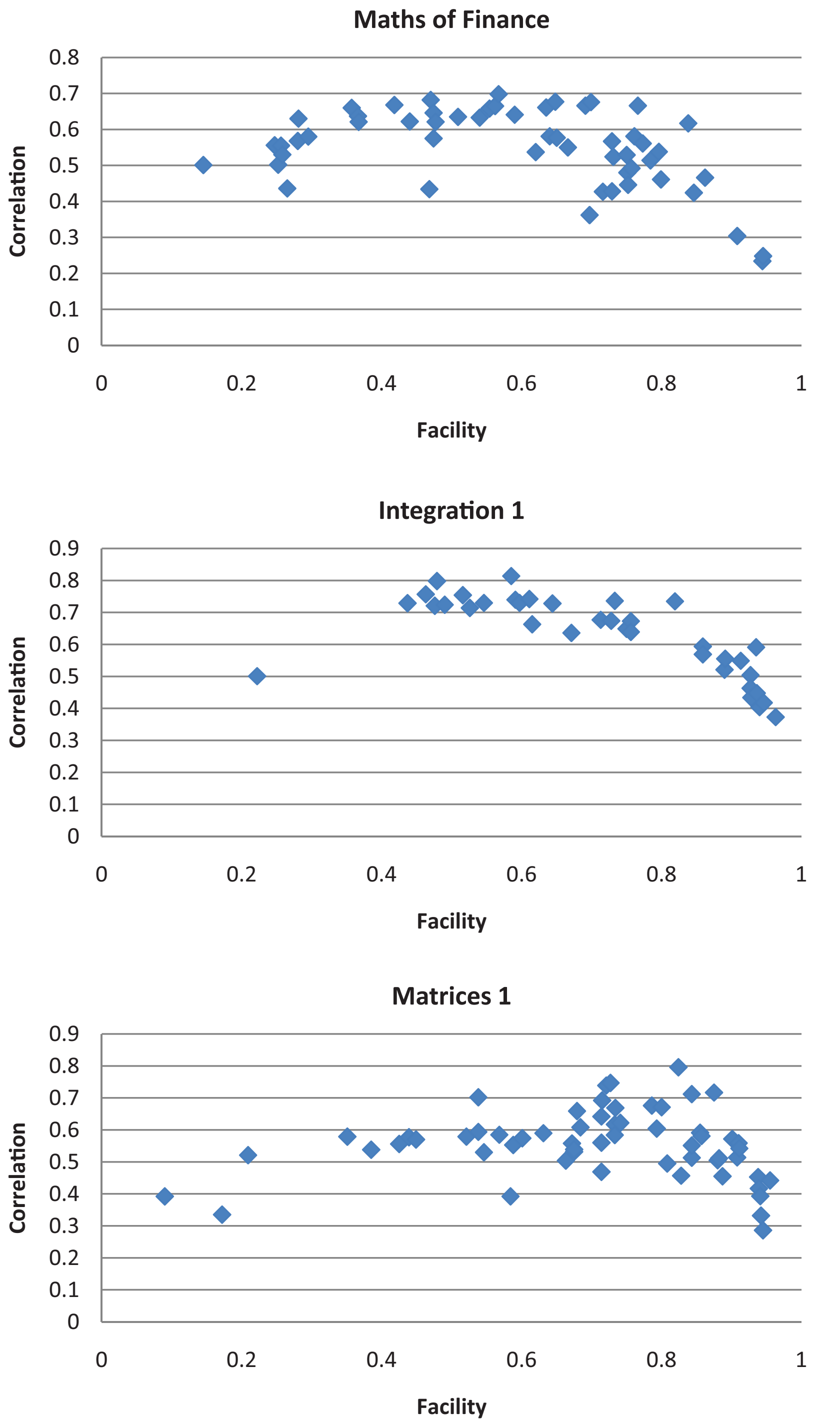


FIG. 5. Continued.

of rank ordering the facilities, not of providing their (even approximate) values. Another possible rank-ordering measure is to consider the concepts needed to answer a question, with some sort of agreement on the rank ordering of the concepts and probably some weighting if several concepts are used. It is very unclear how to do this systematically and certainly begs the question of what underlying structure makes one concept harder than another. Moreover, the notion of understanding a concept may rely on the student's prior knowledge (or *cognitive domain*) and hence be specific to the student rather than being a feature of the question. Nonetheless, it is seen below that both the number of accessible states and the used concepts play a role in the questions' facilities, whose ranges are given in Table 1.

Taking the above topics in turn we have:
Numbers hardest question: (Facility 0.219)
Evaluate the product of $-2.3391918 \text{ x } 10^{-6}$ *and* $-5.88260208 \text{ x } 10^{5}$
Input your answer in scientific notation, with minus signs where necessary. The first input should be correct to four significant figures.

Clearly, there are a number of ways of going wrong, each triggered by the application of a mal-rule, each of which gives two accessible states (correct and mal-rule generated): for example, we might consider (2 for the overall sign) x (4 for the exponent) x (2 states for the significand) x (2 for sig.figs. rather than decimal places) = 32 states.

Numbers easiest question: (Facility 0.994)
Winifred said when $x=4$, $7x^2$ *is 784.*
Ivy said when $x=4$, $7x^2$ *is 112.*
Tia said when $x=4$, $7x^2$ *is 196.*
Victoria said when $x=4$, $7x^2$ *is 121.*
Who is right?

In a sense, the number of states here is just four, i.e., one of the named people. Thus, the concept may fail for Word Input (essentially MC) questions and a comparison with the hardest question may not be valid.

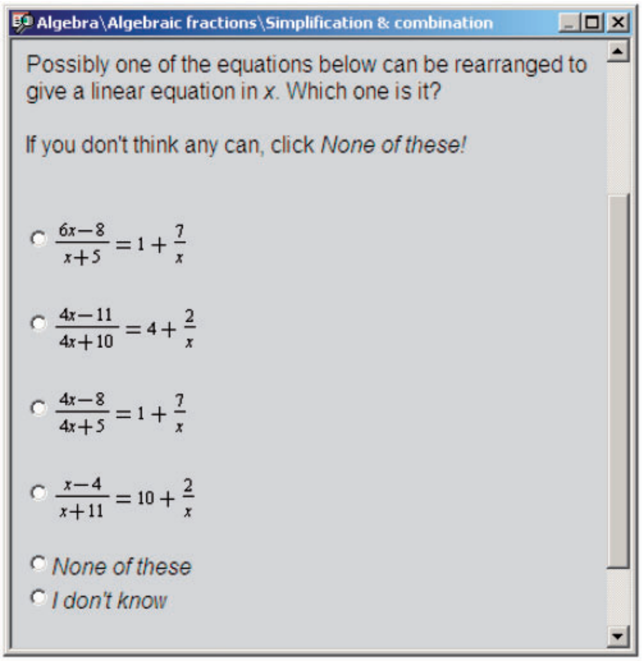
Numbers second easiest question: Facility 0.985

A jar of coffee is usually sold in 150.00 g containers. During a special promotion an extra 25% is added to each jar.
What will the new weight be?
Input your answer correct to 2 decimal places.

Here, a comparison is valid and we have perhaps the following: 2 choices for the multiplier, i.e., 25 or 25/100 (although the first would yield an 'obviously' wrong answer) and 2 for failing to add the extra amount to 150 g. It is difficult to see what else could go wrong and it is quite likely that students getting this wrong simply mistyped their answers.

For Algebra 1 the hardest question (Facility 0.324) is shown in Fig. 6. As a MC question it should have five accessible states and hence not be hard at all. The hardest NI question (Facility 0.478) is shown in Fig. 7. Here, the number of accessible states might be considered as: (2 × 2 for squaring up) × (2 for division by 36) × (2 × 2 for rearranging terms to give $4G^2$) × (2 for division of both terms by 4) × (2 for numerator/denominator confusion) × (2 for lowest fraction) = 256. Note that warnings are given for the last two possibilities here and that in this case (but not others) the resulting fraction (25/9) is already in its cancelled form.

The easiest NI question (Facility 0.944) is of the form $(bx+c)/d+e=f$, where the letters are positive or negative numbers; a similar count of the number of accessible states yields 8. The easiest MC



Possibly one of the equations below can be rearranged to give a linear equation in *x*. Which one is it?

If you don't think any can, click *None of these!*

○ $\frac{6x-8}{x+5}=1+\frac{7}{x}$

○ $\frac{4x-11}{4x+10}=4+\frac{2}{x}$

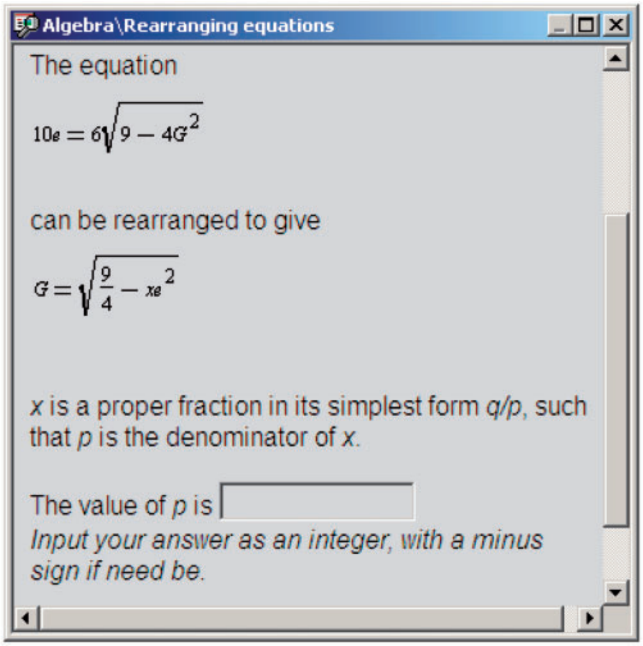
○ $\frac{4x-8}{4x+5}=1+\frac{7}{x}$

○ $\frac{x-4}{x+11}=10+\frac{2}{x}$

○ *None of these*

○ *I don't know*

FIG. 6. A hard MCQ.



The equation

$$10g=6\sqrt{9-4G^{2}}$$

can be rearranged to give

$$G=\sqrt{\frac{9}{4}-xg^{2}}$$

x is a proper fraction in its simplest form *q/p*, such that *p* is the denominator of *x*.

The value of *p* is

Input your answer as an integer, with a minus sign if need be.

FIG. 7. A hard NI question.

question (Facility 1) asks students to expand $a(bx + c)$ with a, b, c positive, having a count of accessible states as 2.

For Algebra 2 both the hardest and easiest NI questions were on proportionality. A realization of the hardest question (Facility 0.053) was 'The average age of a group of teachers and professors is 39 years. If the teachers average 35 years and the professors 52 years, then what is the ratio of the number of teachers to professors?', while the easiest (Facility 0.978) was 'If 13 pears cost £2.73, how much would 12 pears cost?' The number of accessible states of the hardest question could be estimated at 8 while for the easiest one it is 2—hardly a big enough difference to explain the difference in difficulty. It is likely that students simply do not know how to model the hardest question, but do understand what operations are needed for the easiest one.

For Economics the NI hardest question (Facility 0.115) on calculating the marginal product of labour Q_L and marginal product of capital Q_K for a Cobb–Douglas function again has a very low number of accessible states but arguably involves the harder concept of partial differentiation (and students would need to know that this was what was required). The easiest question (Facility 0.897) on finding the average and marginal costs from a quadratic total cost function actually has a similar number of required operations and accessible states, but clearly involves more familiar concepts.

For Differentiation, the hardest NI question (Facility 0.310) was to find the value of the derivative at a given value of x for an expression of the form [polynomial]^m·[polynomial]^n. This involves relatively difficult concepts of both the chain and product rules and has a very large number of accessible states (of the order of 10,000) in contrast to the easiest NI question (Facility 0.871) to find the gradient of a straight line of the form $ax + by = c$ with four accessible states.

For Mathematics of Finance, the hardest NI question (Facility 0.352) involved finding an annual percentage reduction in consumption so that reserves never run out. This involved the concept of summing a GP to infinity and solving the resulting equation, so despite the number of accessible states not being high (estimated to be of the order of 100), the students were required mathematically to model the information given in the rather wordy question, which clearly defeated many. The easiest NI question (Facility 0.933) simply involved a single percentage having four accessible states.

For Partial Differentiation, the hardest/easiest NI questions (Facilities 0.132 and 0.876) involved evaluating the second/first partial derivatives of an exponential/algebraic function in x and y namely of the forms $\exp(x^m y^n)$ and $x^{p/q}y^{r/s}$. The hardest question involved both chain and product rules and hence had more accessible states (of the order of 100) than the easier one (of the order of 10).

For Optimization, the hardest NI question (Facility 0.297) involved a quadratic objective function subject to a linear constraint and asks for optimal values of x and y and the approximate change in optimum when the r.h.s. of the constraint equation changes by a small amount. Given that differentiating the Lagrangian gives a 3×3 system of equations that can be solved in many different ways, it is difficult to estimate the number of accessible states (which is at least several hundred). An additional factor is that students must understand, or at least recall, the underlying theory for the approximate change, although its implementation is trivial; if they do not, the question is impossible. The easiest NI question (Facility 0.738) involves modelling the profit from the demand curves of two goods and maximizing the profit; this involves quite a lot of operations and hence probability of the order of 100 accessible states, but by the time of the test, students were generally very familiar with the concepts and hence able to apply a fairly algorithmic approach successfully.

For Integration 1, the hardest NI question (Facility 0.437) asked students to evaluate a definite integral of a binomial divided by a power of x; although not hard, students may have failed to simplify the integrand first and tried some sort of a substitution method, or been confused that the upper limit of integration was smaller than the lower limit and/or both were negative numbers. This substantially increases the number of accessible states to well over 100. The easiest 21 questions were all multi--choice indefinite integrals where students could have differentiated all the options (these were included in the test noting comments about building confidence given above); the easiest NI question involved definite integration of a power plus an exponential term, with positive limits and integrating from left to right. Thus, although the integration was not substantially easier than the hardest question, the constraints on the limits substantially reduced the number of accessible states to order 10.

For Integration 2, the hardest NI question (Facility 0.439) involved a rational function and integration by substitution with the same limit constraints as in the hardest question for Integration 1 and hence a similar number of accessible states. The easiest NI question (Facility 0.939) was a definite integral of a binomial raised to a negative integer power, to be done by an obvious substitution. Again the choice of limits as in Integration 1 substantially reduced the number of accessible states to order 10.

For Matrices 1, the hardest NI question (Facility 0.090) was to evaluate the long-term market share for a Markov chain for three competing companies. This is quite complicated and students did not understand the concepts either so is not discussed further here. For the second-hardest NI question (Facility 0.172), the market share after 2 or 3 years was required, requiring students to understand that they need to set up the transition matrix T from the rather wordy question specification and performing two or three matrix multiplications (or finding T^2). So the concepts here are quite hard and there are certainly hundreds of accessible states. In contrast, the easiest NI question (Facility 0.955) involves scalar multiplication and addition of two 2×2 matrices and hence is conceptually easy and has of the order of 10 accessible states.

For Matrices 2, the hardest NI question (Facility 0.349) involved answering the final step of an extensive calculation on when a 3×3 system was underspecified (setting $\det(A) = 0$ and solving an equation), what value would be required for part of the r.h.s. and finally expressing x or y in terms of z in that case. The number of accessible states would be extremely high in this case (of the order of thousands) and the concepts are hard. The easiest NI question (Facility 0.833) involved finding a single element of an inverse 2×2 matrix—conceptually straightforward with a clear calculation algorithm and perhaps as few as 16 accessible states.

For the diversity of topics listed above, entropy measure does agree with the data for the hardest and easiest questions regardless of topic. As pointed out by a reviewer of this article, the above methodology is probably too simplistic is be applied universally. For example, by neglecting the fact that there may be several (equally valid) ways of attempting some questions, especially in the more advanced topics, the attribution of an entropy value to a question could be problematic. This is certainly correct and it is likely that further studies will require a theoretical model of difficulty that combines (at least) entropy and the inclusion of the used concepts' difficulties. The large answer file data set considered here could provide a benchmark against which possible theoretical models could be tested, for example by considering rank correlation coefficients. This might then provide an empirical, rather than theoretical, way of discriminating between possible methodologies for assessing question difficulty. Similar comparisons with question discrimination seem even more difficult to tackle, but one could speculate that concepts might be more important that the more mechanistic procedures required by entropy (which do not distinguish between diligent and able students). The initial aim of finding methodologies that work might then trigger an understanding of why they work.

7. Conclusions

While acknowledging the limitations of CAA, this article makes the case for using CAA as part of a blended formative and summative assessment of any mathematics module at level 1 of a university degree. Here, the mathematical skills to be tested are more mechanistic and hence lend themselves to testing via objective questions, particularly those of multi-choice and NI types. Key to the design of these questions is the use of random parameters and encoded mal-rules that trigger highly specific and very detailed feedback. The article offers compelling evidence from the questions' facilities and correlations that the resulting questions are robust and fit-for-purpose and can be successfully embedded within the module using a clear and practice regime. This has resulted in substantially enhanced performance compared with previous cohorts of Level 1 Economics students at Brunel University. The administrative benefits of implementing this regime are highlighted too.

Given the very large data set of 6 years' worth of answer files (comprising over 270,000 question attempts) and the relative ease with which these can be analysed, the author is then able to speculate on the question of what makes one question harder than another. It is shown that at least for some of the easier topics, the number of likely mal-rules that can be applied to a question is directly related to the number of ways a student can go wrong (the number of accessible states) and hence it is reasonable to suppose that this affects the question's difficulty. By looking at the hardest/easiest questions, some evidence is presented to show that this could be a useful predictor of question facility, although it is also clear that familiarity with the concepts needed also plays a role and this may dominate when students presumably cannot even get started on the hardest question.

References

Badger, M. & Sangwin, C. J. (2011) My equations are the same as yours!: Computer aided assessment using a Gröbner basis approach. *Teaching Mathematics Online: Emergent Technologies and Methodologies* (A. A. Juan, M. A. Huertas & C. Steegmann eds). IGI Global, pp. 259–273, Hershey, PA 17033, USA.

BBC. BiteSize. Available online at: http://www.bbc.co.uk/schools/gcsebitesize/ (accessed 28 April 2015).

Biggs, J. & Collis, K. (1982) *Evaluating the Quality of Learning: The SOLO Taxonomy.* New York: Academic Press.

CAA (2002) Computer-aided assessment. Available online at: www.caacentre.ac.uk (accessed 28 April 2015).

Chick, H. (1998) Cognition in the formal modes: Research mathematics and the SOLO taxonomy. *Math. Educ. Res. J.*, 10, 4–26.

DEWIS. University of the West of England. Available online at: http://www.cems.uwe.ac.uk/caa/welcome/index.html (accessed 28 April 2015).

Gill, M. (2007) Development and evaluation of CAA as a Formative Assessment Tool for University-level Mechanics. *Ph.D. Thesis*, Mathematical Sciences, Brunel University.

Gill, M. & Greenhow, M. (2006) *Computer-Aided Assessment in Mechanics: Question Design and Test Evaluation.* Teaching Mathematics and its Application, MEE, Oxford, OX2 6DP, UK.

Greenhow, M. and Zaczek, K. (2011) Embedding computer-aided assessment of mathematics and statistics for first year economics students. *Proceedings of the ICTMT10 Conference*, Portsmouth. Available online at: http://www.dm.uniba.it/ictmt11/download/ICTMT10_Proceedings.pdf (accessed 28 April 2015).

Gwynllyw, R. and Henderson, K. (2012) Intelligent marking in summative e-assessments. *Proceedings of the HE STEM Conference*, Available online at http://journals.heacademy.ac.uk/doi/book/10.11120/stem.hea.2012 (accessed 28 April 2015).

Gwynllyw, R., Weir, I. and Henderson, K. (2015) Using DEWIS and R for multi-staged statistics e-Assessments (to appear).

Hanson, J. (2011) A determination and classification of student errors in lower-level calculus through computer-aided assessment and analysis. *M.Phil. Thesis*, Mathematical Sciences, Brunel University. Available online at: http://bura.brunel.ac.uk/handle/2438/6050 (accessed 28 April 2015).

Haynes, L. and Hermans D. F. M. (2007) An investigation into common student errors in first year Calculus and Algebra. *MSOR Unpublished Report*.

Hot Potatoes. Available online at: http://hotpot.uvic.ca/ (accessed 28 April 2015).

JISC (2014) Available online at: http://www.jisc.ac.uk/rd/projects/assessment-and-feedback (accessed 23 June 2015).

Maple TA. MapleSoft. Available online at: http://www.maplesoft.com/products/mapleta/ (accessed 28 April 2015).

maths e.g. Brunel University. Available online at: http://www.mathcentre.ac.uk:8081/mathseg/ and http://www.mathcentre.ac.uk:8081/mathsegteacher/ (accessed 28 April 2015).

MyMathLab. Pearson Education. Available online at: http://www.mymathlab.com/ (accessed 28 April 2015).

MyMaths. Available online at: http://www.mymaths.co.uk/ (accessed 28 April 2015).

NUMBAS. Newcastle University. Available online at: http://www.ncl.ac.uk/maths/numbas/ (accessed 28 April 2015).

Perception: Question Mark Computing. Available online at: http://www.questionmark.com/uk/ (accessed 28 April 2015).

Priem, J. (2010) Fail better: Toward a taxonomy of e-learning error. *J. Educ. Comput. Res.*, 43, 377–397.

Ridgeway, J., McCusker, S. and Pead, D. (2004) *Literature Review of E-assessment*. Futurelab Series 10, Futurelab. ISBN: 0-9544695-8-5. Available online at: http://www.academia.edu/2037079/Literature_review_of_e-assessment (accessed 23 June 2015).

Sangwin, C. (2013) *Computer Aided Assessment of Mathematics*, 1st edn edn.. Oxford: Oxford University Press (accessed 28 April 2015).

Schechter, E. (1994) *The Most Common Errors in Undergraduate Mathematics*. Available online at: http://www.math.vanderbilt.edu/~schectex/commerrs/.

STACK: Birmingham University. Available online at: http://www.stack.bham.ac.uk/ (accessed 28 April 2015).

Wiley Plus: Wiley. Available online at: https://www.wileyplus.com/WileyCDA/ (accessed 28 April 2015).

Zaczek, K. (2015) Development and evaluation of computer-aided assessment in discrete and decision mathematics. *Ph.D. Thesis*, Mathematical Sciences, Brunel University.

Appendix. EC1005 Mathematics for Economics at Brunel University

In 2006–2007, the Department of Mathematical Sciences was asked by our Economics Department to share the teaching of their first-year mathematics and statistics/research methods modules (EC1005 and EC1006). At that time I was engaged in writing the CAA material for mathematics for economics at this level, see http://www.metalproject.co.uk/ from where the assessments, and much else, can be downloaded, so I welcomed the chance to teach the group without A-level mathematics (typically about half the cohort of about 340 students aged 18–20 years, 55% male). I anticipated using the CAA to replace the two class tests. In fact this did not happen until 2008–2009 (for the both cohorts); in addition to saving me about 8 solid days marking, the examinations and overall performance of the students then increased substantially, see Fig. 4. During these 3 years the performance of the two groups (with/without A level) was indistinguishable at the end of the year, and it was felt that the weaker students benefitted relatively more from the CAA. In 2009–2010, I took over the entire cohort and the admissions policy changed to require at least AS level mathematics at grade C. In fact, about half of the cohort had the full A-level at grade B or above, and the qualifications for this year (2010–2011) comprise grades A*(3%), A(12%), B(50%), C(21%), D(4%) with the rest having AS grades A(3%), B(4%) and C(3%). It is gratifying that very few students who took AS level mathematics failed to continue to the full A-level in year 13. Given the enhanced mathematical preparedness of the students since 2009–2010, the low examination average was somewhat disappointing. It can be explained by the fact that I set the examination for the first time and overestimated the speed with which students could do the questions (corrected in 2010/2011). So although the overall average slightly increased, some 70 students had to resit the examination in September 2010 since a genuine pass in the examination component was introduced as a module requirement (in addition to the examination being weighted at 70%). Of the 2009/2010 resitters about 50 passed. Results for 2010/2011 and 2011/2012 are encouraging across all grades: in 2011 of 55 resitters, 50 passed and in 2012 of 23 resitters, all passed (helped no doubt by the continued use of the CAA over the summer for revision purposes, but this was not recorded in 2011 or in 2012 when they could also use *maths e.g.* from home). For 2013, 18 out of 19 resitters passed.

An obvious confounding influence is that the admissions criteria evolved over the course of this study. These are given below, but it is worth nothing that even if not a prerequisite, at least half of the class always had a good A level in mathematics (A or B), so the effect may be less marked than that suggested below:

2006–2008 entry: GCE A-level BBC no Maths or Statistics prerequisite (280 points).
2009 entry: GCE A-level BBC including at least AS level Maths or Statistics at grade C (280 points).
2010–2013 entry: GCE A-level BBC including at least AS level Maths or Statistics at grade C (320 points from 3 A levels + 1AS).