ISSUES WITH SETTING ONLINE OBJECTIVE MATHEMATICS QUESTIONS AND TESTING THEIR EFFICACY

Nabamallika Baruah, Mundeep Gill and Martin Greenhow

Issues with Setting Online Objective Mathematics Questions and Testing their Efficacy

Nabamallika Baruah, Mundeep Gill and Martin Greenhow
Department of Mathematical Sciences
Brunel University
nabamallika.baruah@brunel.ac.uk

Abstract

The Mathletics database now comprises many mathematical topics from GCSE to level 2 undergraduate. The aim of this short paper is to document, explore and provide some solutions to the pedagogic issues we are facing whilst setting online objective questions across this range. Technical issues are described in the companion paper by Ellis, Greenhow and Hatt (2006). That paper refers to "*question styles to stress that we author according to the pedagogic and algebraic structure of the content of a question; random parameters are chosen at runtime ... This results in each style having thousands, or even millions, of realisations seen by the users.*" With this emphasis, and with new topics being included, new question types beyond the usual multi-choice (MC) etc have been developed to ask appropriate and challenging questions. We feel that their pedagogic structure (and underlying code) is widely applicable to testing beyond the scope of Mathematics. This paper describes three of the new question types: Word Input, Responsive Numerical Input and 4/True/False/Undecidable/Statement/Property. Of generic importance is the fact that each of these question types can include post-processing of submitted answers; sample Javascript coding that checks the validity of the input(s) before marking takes place is described. In common with most of the rest of the question style's content this could be exported to other CAA systems.

Ellis et al (2005) and Gill & Greenhow (2006) describe initial results of a trial of level 1 undergraduate mechanics questions. This academic year we have expanded the range of tests to foundation and level 1 undergraduate algebra and calculus, involving several hundred students. First and foremost we have underlined the value of Random Numerical Input (RNI) question types compared with traditional Numerical Input (NI) types for which answer files resulting from questions with randomised parameters are exceptionally difficult to interpret. Despite our current lack of a consistent and fully-meaningful way of encoding the mal-rules within the question outcome metadata, mal-rule-based question types (MC, RNI etc) have been analysed in terms of difficulty, discrimination and item analysis. In the case of multiple-choice questions any weaknesses are separately identified as skill-based or conceptual.

Introduction

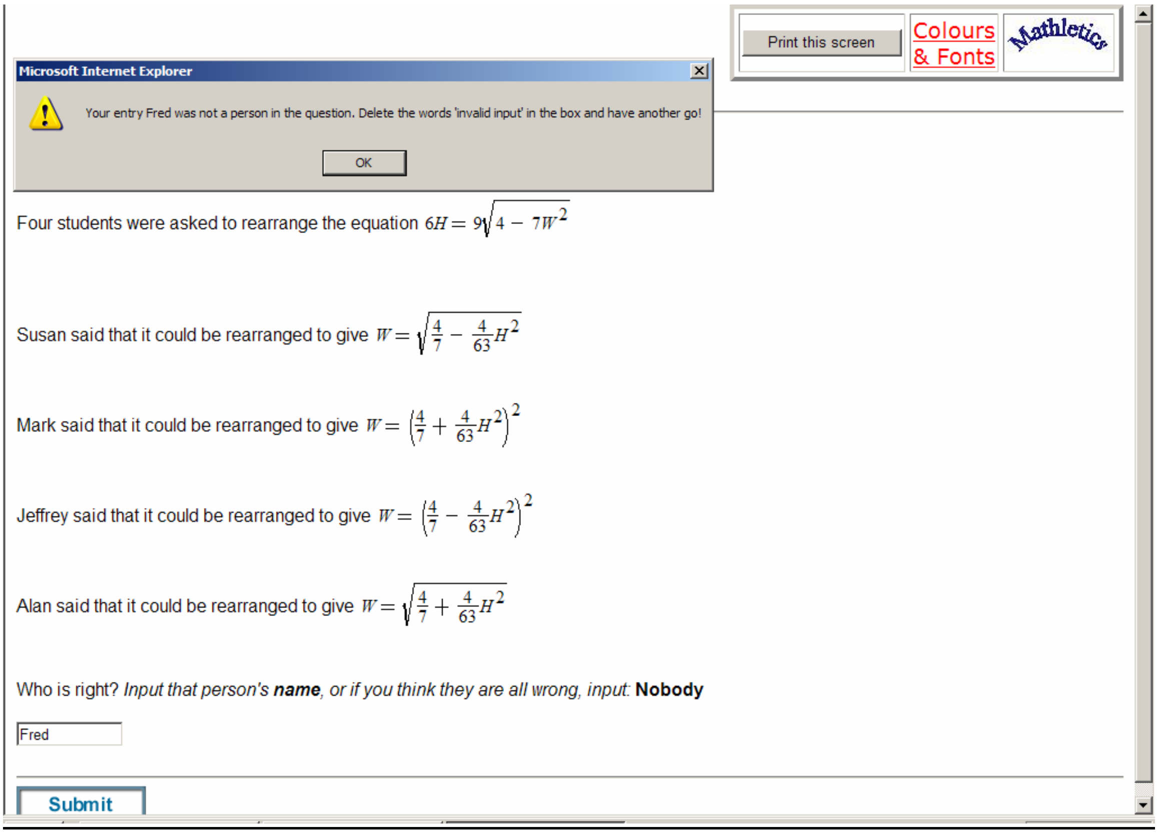
Multiple-choice questions are the most common types of questions used to set objective tests. Previous papers (Gill & Greenhow, 2006; Ellis *et al* 2005; Gill & Greenhow, 2005;) have discussed the methodology we have used at Brunel University to ensure that the options made available in multiple-choice questions are reliable and realistic. Past exam scripts in the areas of calculus and mechanics have been analysed to identify common mistakes that students make while answering certain types of questions. Similar work is also currently being carried out in the area of algebra. It is hoped that by identifying common mistakes and using these as distracters, the feedback will be more focused on individual errors and feedback to the lecturers will also highlight common mistakes that students are making.

Many objective tests have been set up and used at Brunel University over the past academic year. These tests cover areas such as algebra, calculus, mechanics and statistics, mainly at level 0 and level 1. Some tests have been used purely for formative reasons while others have been used for summative purposes. Students are encouraged to use the questions for revision purposes to aid them in their learning process. From analysis of student answer files for calculus and mechanics it was found that a higher percentage of students were able to answer multiple-choice type questions correctly compared with numerical input (see table 2 below). Since final examinations do not generally contain multiple-choice questions, it was decided to develop other types of questions.

Some New Question Types

Word Input (WI)

Even in a tightly-specified setting requiring the input of only short phrases, marking algorithms in any objective system will find it difficult to equate the meaning of equivalent forms (e.g. *x is at least as large as y is equivalent to x is not smaller than y*). We have sought to facilitate the communication between user and marking scheme by casting questions in terms of the positions taken by protagonists. A very simple example is shown in figure 1, but this type could be used to require students to evaluate each of the protagonist's positions on a more complex or incompletely-specified "real-world" problem. Figure 1 shows a situation with five possible answers (note the use of *Nobody*), since here we need to link names with a mathematical expression; we have effectively created a multiple-choice question in another form. However, it would be entirely feasible to set up a much less constrained question stem with an arbitrary number of (unique) names, asking, for example, who's position is best supported by the evidence presented.



Your entry Fred was not a person in the question. Delete the words 'invalid input' in the box and have another go!

Four students were asked to rearrange the equation $6H = 9\sqrt{4 - 7W^2}$

Susan said that it could be rearranged to give $W = \sqrt{\frac{4}{7} - \frac{4}{63}H^2}$

Mark said that it could be rearranged to give $W = \left(\frac{4}{7} + \frac{4}{63}H^2\right)^2$

Jeffrey said that it could be rearranged to give $W = \left(\frac{4}{7} - \frac{4}{63}H^2\right)^2$

Alan said that it could be rearranged to give $W = \sqrt{\frac{4}{7} + \frac{4}{63}H^2}$

Who is right? *Input that person's* ***name****, or if you think they are all wrong, input:* **Nobody**

Fred

Figure 1 A

The variable names (H and W) are randomly chosen from a subset of upper/lower case alphabetical characters. All numbers are randomised with certain bounds determined by the pedagogy of the question (e.g. how difficult should the arithmetic be?). The protagonists' names are selected randomly from male/female datasets reflecting the 16-25 year old UK ethnic mix. This results in millions of (pedagogically and algebraically equivalent) realisations of this question style.

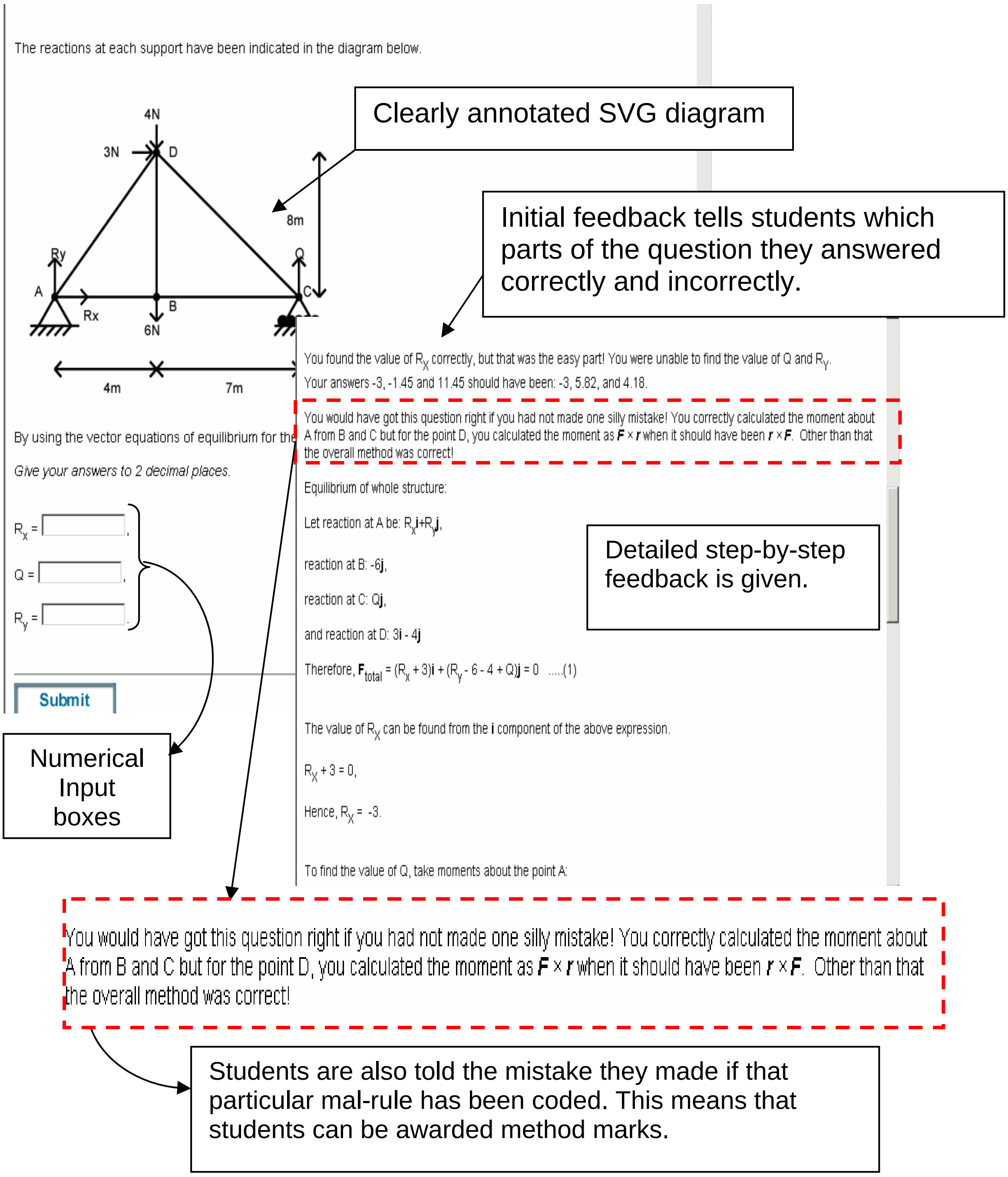
Although seemingly straightforward to mark, a degree of post-processing of user input is now required. By comparison with each of the n entries in the question's protagonist list (person[]), we firstly check that an entry is a valid name (not a misspelling or in the wrong case) and issue an appropriate warning (as shown in figure 1) if necessary. Next, for the sake of correct grammar, proper nouns are automatically capitalised for the student if they have not used them before marking comparison takes place. We believe that something like the following code will generally be needed for robust handling of word input:

```
//If input did not begin with an upper case, then this will be automatically updated for them
okinput=0
for (k = 0; k <=n-1; k++){
if (document.forms[0].elements[item].value.toUpperCase() == person[k].toUpperCase()){okinput=okinput+1}
}

if (document.forms[0].elements[item].value.toUpperCase() == "nobody".toUpperCase()){okinput=okinput+1};

//If input was not a person in the question, then a alert message is prompted saying so
if (okinput == 0) {
		alert("Your entry "+document.forms[0].elements[item].value+" was not a person in the question. Delete the
words 'invalid input' in the box and have another go!");
document.forms[0].elements[item].value="invalid input";}else{

strlength = document.forms[0].elements[item].value.length;
part =  document.forms[0].elements[item].value.substring(0,1).toUpperCase();
rest =  document.forms[0].elements[item].value.substring(1,strlength).toLowerCase();

document.forms[0].elements[item].value = part+rest}
```

Responsive Numerical Input (RNI)

A weakness of basic numerical input type questions is that the answer inputted by students is marked either correct or incorrect. Therefore the feedback provided can only indicate whether students answered the question correctly or not and provide the standard worked solution. These types of questions do not provide directed feedback, as multiple-choice do, and hence are not seen to be as effective. However, we have developed a new question type known as *Responsive Numerical Input*. This type of question is very similar to multiple-choice but differs in that (an arbitrary number of) distracters are coded in the background and are not presented to students as in multiple-choice questions. This means that if a student makes a particular mistake that has been coded as a mal-rule, then the feedback can be similar to that of a multiple-choice question, correcting specifically the mistake they have made in their working; for example, a student may have interpreted (a+b)/c as a+b/c. Partial credit can be awarded if appropriate. However, in contrast to multiple-choice questions, students will be unable to eliminate the correct answer from a list of options. Feedback to lecturers will be more informative and students will be faced with a more realistic form of testing, i.e. similar to that of exams.

Responsive numerical input type questions can also be extended to *Sequential Responsive Numerical Input* types. This type of question is used for questions that contain more than one part and the different parts are connected. For example, students may need their answer to the first part to answer the second part. The advantage of using a sequential responsive numerical input type is that not only will feedback be directed (as in responsive numerical input) but students can also be told whether the method they attempted is correct or not (given their answer to the previous part of the question was incorrect). Figure 2 shows an example of a sequential responsive numerical input type question.



The reactions at each support have been indicated in the diagram below.

By using the vector equations of equilibrium for the

Give your answers to 2 decimal places.

R_x = ,

Q = ,

R_y = .

You found the value of R_X correctly, but that was the easy part! You were unable to find the value of Q and R_Y.

Your answers -3, -1.45 and 11.45 should have been: -3, 5.82, and 4.18.

You would have got this question right if you had not made one silly mistake! You correctly calculated the moment about A from B and C but for the point D, you calculated the moment as $\boldsymbol{F} \times \boldsymbol{r}$ when it should have been $\boldsymbol{r} \times \boldsymbol{F}$. Other than that the overall method was correct!

Equilibrium of whole structure:

Let reaction at A be: $R_x\mathbf{i}+R_y\mathbf{j}$,

reaction at B: $-6\mathbf{j}$,

reaction at C: $Q\mathbf{j}$,

and reaction at D: $3\mathbf{i} - 4\mathbf{j}$

Therefore, $\mathbf{F}_{total} = (R_x + 3)\mathbf{i} + (R_y - 6 - 4 + Q)\mathbf{j} = 0$ .....(1)

The value of R_X can be found from the **i** component of the above expression.

$R_X + 3 = 0$,

Hence, $R_X = -3$.

To find the value of Q, take moments about the point A:

You would have got this question right if you had not made one silly mistake! You correctly calculated the moment about A from B and C but for the point D, you calculated the moment as $\boldsymbol{F} \times \boldsymbol{r}$ when it should have been $\boldsymbol{r} \times \boldsymbol{F}$. Other than that the overall method was correct!

Clearly annotated SVG diagram

Initial feedback tells students which parts of the question they answered correctly and incorrectly.

Detailed step-by-step feedback is given.

Numerical Input boxes

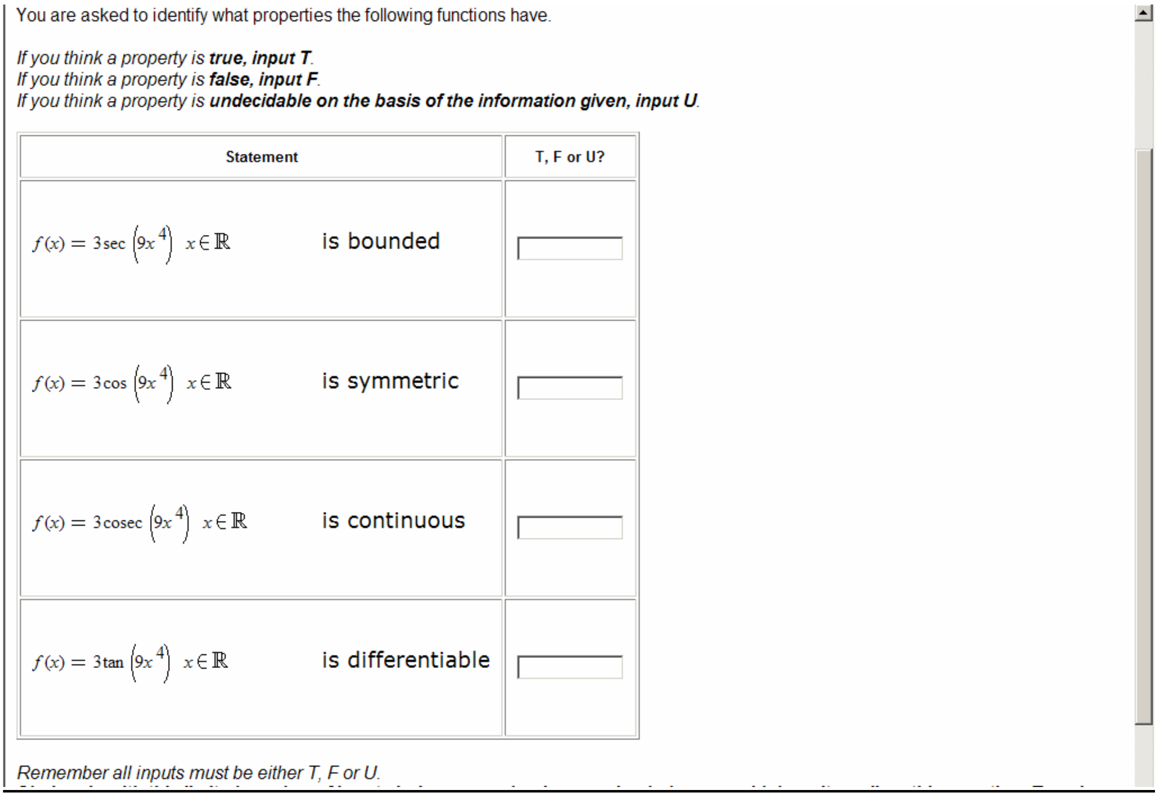
Students are also told the mistake they made if that particular mal-rule has been coded. This means that students can be awarded method marks.

Figure 2: Example of a Sequential Responsive Numerical Input type question

The feedback that is provided to students not only indicates the parts of the questions that students answered correctly and incorrectly, but it also tells students where an error in their working has been made. This type of feedback is useful for questions where the method students are required to use is lengthy and students may spend a long time attempting such questions. The amount of coding required for a question such as that shown in Figure 2 is extensive, but it is hoped that students find such questions worthwhile and more challenging than multiple-choice type questions.

4 True, False or Undecidable; Statement and Property (4TFUSP)

Figure 3 shows a realisation of this type of question. Not only are the statement parameters (choice of trig function and coefficients) randomised, but the properties of the propositions (bounded, symmetric etc) are also randomised. This considerably expands the number of realisations available in the question style. By adding four parts to the question an expansive almost exam-like question is generated that could challenge many students. Variants having either statement of property choice fixed, are useful for determining a students' knowledge of a function (e.g. sine having properties such as continuity, antisymmetry etc) or a property (e.g. which of the randomly-chosen functions are symmetric).



You are asked to identify what properties the following functions have.

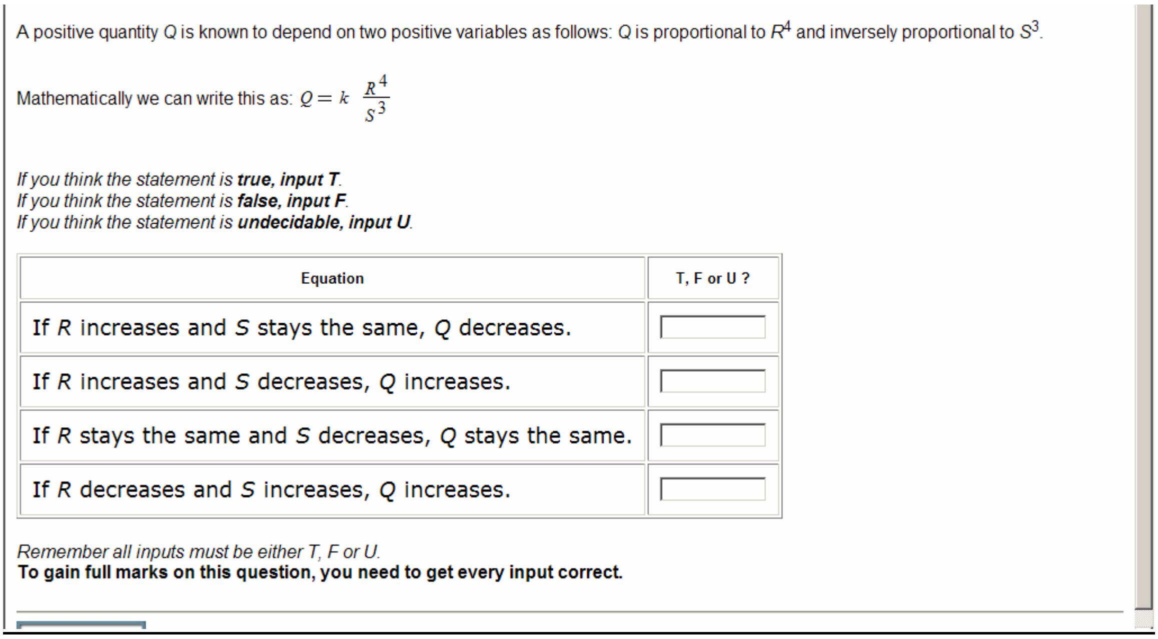
*If you think a property is **true, input T**.*
*If you think a property is **false, input F**.*
*If you think a property is **undecidable on the basis of the information given, input U**.*

Statement	T, F or U?
$f(x) = 3\sec\left(9x^4\right) \quad x \in \mathbb{R}$ is bounded	
$f(x) = 3\cos\left(9x^4\right) \quad x \in \mathbb{R}$ is symmetric	
$f(x) = 3\operatorname{cosec}\left(9x^4\right) \quad x \in \mathbb{R}$ is continuous	
$f(x) = 3\tan\left(9x^4\right) \quad x \in \mathbb{R}$ is differentiable	

Remember all inputs must be either T, F or U.

Figure 3. A 4 True, False or Undecidable; Statement and Property (4TFUSP) question type.

Another example is shown in figure 4. Obviously the question stem could be altered to describe a "real-world" scenario with the input boxes stating plausible conclusions or recommendations that might, or might not, follow from the scenario. Indeed it is planned to utilise this type of question (and word input questions) to test students' understanding of statistical inference and transferable skills, such as critical thinking. Notice again that the validity of student input must be checked, with lower case t, f, u inputs being changed to capitals. All other inputs triggering an invalid input message similar to that shown in figure 1 must be addressed.



A positive quantity Q is known to depend on two positive variables as follows: Q is proportional to R^4 and inversely proportional to S^3.

Mathematically we can write this as: $Q = k\ \frac{R^4}{S^3}$

If you think the statement is ***true, input T****.*
If you think the statement is ***false, input F****.*
If you think the statement is ***undecidable, input U****.*

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

Remember all inputs must be either T, F or U.
To gain full marks on this question, you need to get every input correct.

Figure 4. A 4 True, False or Undecidable; Statement and Property (4TFUSP) question type testing interpretation of a mathematical expression.

Methods Used to Evaluate the Feedback Provided and the Overall Question Efficacy

For all questions that have been produced much time and effort has been dedicated to the feedback being provided to the students. Within the Brunel group there was much debate over the amount of feedback that should be provided: some members thought that students would simply ignore the feedback if too much was provided, while others thought that students would benefit from the detailed feedback. We therefore decided to investigate how effective the feedback provided actually was. Initial results, mainly specific to the topic area of mechanics, were reported in Gill & Greenhow (2006); we now have more data to report.

Over the past two academic years we have incorporated mechanics lab sessions into the level 1 mechanics module at Brunel University (a core module for Mathematics students). These sessions ran on a weekly basis and though not compulsory, the students were encouraged to attend. Students completed a different assessment at each session, and were able to make use of any resources they wanted. Answer files for all assessments attempted were also recorded. We used the Assessment Experience Questionnaire (AEQ), from the Formative Assessment in Science Teaching (FAST) project group (FAST 2004), to get very positive feedback from the students about the questions, see Gill and Greenhow (2006). That paper also identifies the longer-term effects of participation in the lab sessions on students' approach to tackling questions on the end-of-module exam.

Student Retention Periods: Recorded Answer Files

It was hoped that although the feedback provided was extensive, students would be able to retain and make use of it after a delayed time period. Some students repeated the assessments more than once, either within the same lab session or after a period of time. By analysing these student answer files we aimed to see if students could retain the feedback and make use of it in their subsequent attempts. Table 1 shows the results obtained from the analysis of student answer files for mechanics topics: no similar data is yet available for calculus or algebra topics. It lists each assessment that students repeated and the periods of time students were able to retain the feedback. These have been grouped into either short time periods (1 day to 4 weeks) or long time periods (5 weeks to 7 weeks).

	Retention Period			
Assessment	Retain Feedback Immediately	Short Period 1 day to 4 weeks	Long Period 5 weeks to 7 weeks	Unable to retain feedback for any period of time longer than immediate use
Forces & Vectors	6	1	2	5
Forces & Vectors 1	5	1	2	3
Resolving Forces	3		1	2
Resolving Forces (Tension)	3	4		6
Resolving Forces (Equilibrium)	4	1	3	2
Resolving Forces (Inclined Plane)	5	1	1	4
Revision of Resolving Forces	2			1
Trusses & Loaded Beams	3	1		4
Trusses	2			4
TOTAL	**33**	**9**	**9**	**31**

Table 1: Retention of feedback as identified by correct answers recorded for subsequent test(s) for each of the topic areas; from Gill and Greenhow (2006).

On analysing student answer files it was found that all students were able to retain the feedback long enough to make use of it within the same day. However, many students were unable to retain the feedback for any longer other than immediate use. Some students were able to retain the feedback for a period of 7 weeks, which may imply that these students have mastered the material that was being tested. These results are positive and imply that students are able to retain the feedback provided to them. Observations made during the lab sessions indicated that many students were using the questions as a learning tool rather than an assessment. There was evidence of randomly selecting options and inputting random numbers just to get to the feedback screen. This was surprising since it was thought that students would be more concerned with what *mark* they received and would therefore make use of other resources to help them answer the questions. In actual fact students made use of the questions by reading through the feedback and then reattempting them.

Item Analysis

Mechanics assessments

Throughout all the mechanics assessments there were 2 main question types: Multiple-choice and Numerical Input. The numerical input questions ranged from 1 numerical input to 4. Some questions were sequential and/or responsive. So far we have only analysed the results in terms of students answering the different types correctly and incorrectly. Individual question item analysis has yet to be done where common student mistakes can be identified and reported on. Table 2 shows the percentage of students that answered the different question types correctly and incorrectly.

Question Type	Correct	Distracters	Other (Don't know or only parts correct)	Wrong	Random Input for Feedback
Multiple-Choice	58%	21%	9%	12%	-
1 Numerical Input	38%	-	-	62%	-
2 Numerical Input	39%	-	18%	43%	-
3 Numerical Input	20%	4%	35%	24%	17%
4 Numerical Input	3%	11%	11%	50%	25%

Table 2: Summary of ways students answer different question types.

Table 2 shows that a higher percentage of students answer multiple-choice questions correctly compared with the other types of questions. One possible reason for this may be due to the fact that 4 numerical options are presented to select the answer from (although n*one of these* could be the correct answer). Students have the opportunity to work through a number of different methods until they have a numerical answer that is identical or at least similar to one that is presented to them. In a sense this makes multiple-choice questions 'easier' to attempt compared with Numerical Input types and hence strengthens the need to use question types such as Responsive Numerical Input.

Roughly the same percentage of students answer 1 Numerical Input and 2 Numerical Input types correctly. Many students did not even attempt to answer 3 Numerical and 4 Numerical input type questions but used them only for the purpose of reading through the feedback.

Foundation level assessments

The item facility index is one of the most useful, and most frequently reported, item analysis statistics. The facility index of an item indicates what percentage of students know the answer. For this reason it is frequently called the *p-value*.

Table 3 shows a small selection of questions that were used to test 170 foundation students on differentiation and integration. The table indicates the concept being tested, facility of the question and the discrimination.

Question Type	**Concept**	**Facility**	**Discrimination**
Multiple-Choice	Differentiation: Chain rule	0.629	0.815
	Differentiation: Product rule	0.551	0.554
	Integration: Polynomial	0.71	0.669
	Differentiation: Polynomial	0.667	0.702
RNI	Integration: Rational form	0.363	0.447
	Integration: Polynomial form	0.34	0.753
	Integration: Powers	0.273	0.805
NI	Integration: Logarithmic form	0.056	0.472
	Differentiation chain rule	0.417	0.789
Hot line	Differentiation chain rule	0.407	0.615

Table 3: A selection of questions that were used in the foundation differentiation and integration test.

The facility of the multiple choice questions range from 0.551 to 0.71. This indicates that students did not find these particular questions difficult or challenging. In comparison, students found responsive numerical input questions difficult since the facility ranged from 0.273 to 0.363. This is much lower than the facilities obtained for the multiple choice questions. Similarly, numerical input questions were also perceived to be difficult since the facility levels ranged from 0.056 to 0.417. This indicates that numerical input type and responsive numerical input types are comparatively harder than multiple choice questions.

Discrimination measures how performance on an item correlates to performance in the test as a whole. There should always be some correlation between item and test performance, however, it is expected that discrimination will fall in a range between 0.5 and 1.0. Figure 5 shows the relationship between discrimination and facility for the results obtained from the integration test.

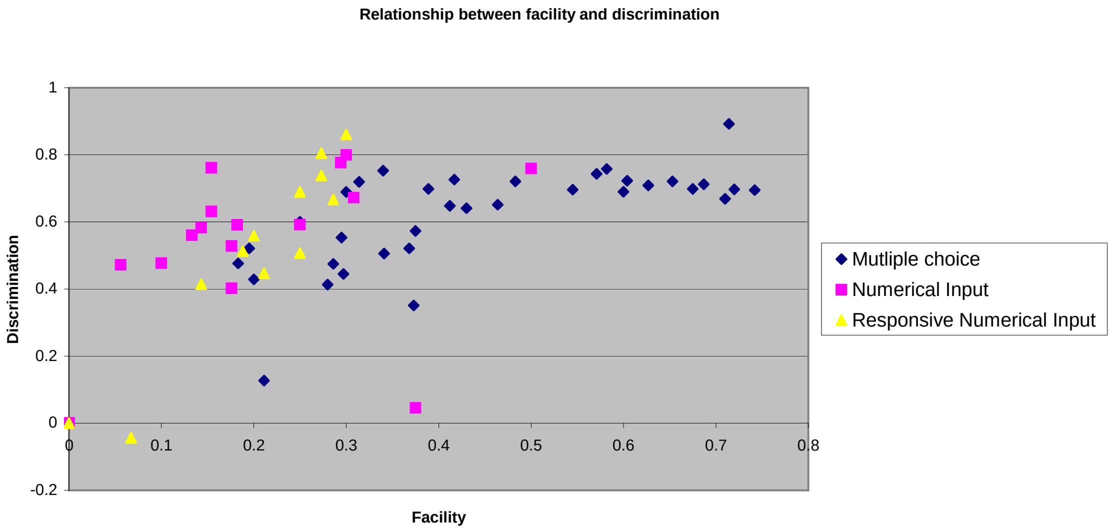


Figure 5: A scatter diagram of the relationship between facility and discrimination for questions in the foundation integration test

From Figure 5 it can be seen that differing facilities between the question types is apparent. The facility for numerical and responsive numerical input type questions is small whereas the mean for the multiple choice questions is much larger. For the majority of the questions the discrimination level is above 0.4, which indicates that most of the questions discriminate well, and ensured the efficacy of the test. The items lying above discrimination level of 0.5 indicate that these questions are highly discriminating.

The items showing negative correlation indicates that a higher proportion of the low scoring group answered the question correctly than that from the high scoring group and conversely. Such type of questions should be examined for finding the possible reason(s) for the reverse difference between the high and low scoring groups.

In the case of multiple-choice questions, responsive numeric input and hot line questions the weaknesses can be separately identified as skill-based or concept based. The structured mal rules record the difficulties of the students in the answer file. Before setting the questions, their objectives are determined (whether skill based or concept based). The skill level and the concept level questions of the foundation level calculus test has been analysed according to the mean facility and the discrimination index.

Levels	Mean facility	Mean discrimination index
Skill	0.48	0.48
Concept	0.475	0.467

Table 4: Table showing mean facility and mean discrimination index for skill and concept questions.

It has been observed that the mean facility and discrimination of the two levels i.e. skill and concept are nearly equal. The lower difference of facility and discrimination of both the skill based and concept based question indicate that the questions are of moderate difficulties with acceptable discrimination.

Conclusions

Our results so far show considerable variability of success rate for different question types across a range of mathematical topics. Students certainly engage with the questions and make extensive use of the feedback provided; they regard this as a valuable learning resource and appreciate the directed feedback offered in response to wrong choices made for multiple-choice questions. Therefore, as part of a formative assessment, multiple-choice questions are very valuable in building knowledge and confidence. However, comparison with other question types, such as numerical input, show the limitations of multiple-choice questions when used summatively or for testing topic mastery. This implies that a variety of question types, including the new ones described here, should be used to give a more sophisticated measure of the student's profile of skills and abilities. In particular we recommend that responsive numerical input types should displace traditional numerical input questions, and multi-stage questions should be authored as sequential (responsive) numerical input if possible.

References

Ellis, E., Baruah, N., Gill, M., Greenhow, M. 2005 Recent developments in setting objective tests in mathematics using QM Perception Proc 9th CAA Conference, Loughborough, July http://www.caaconference.com

E Ellis, M Greenhow, Hatt, J. 2006 Exportable technologies: MathML and SVG objects for CAA and web content Proc 10th CAA Conf, Loughborough, July. http://www.caaconference.com/

FAST – Formative Assessment in Science Teaching 2005 http://www.open.ac.uk/science/fdtl

Gill, M. & Greenhow, M. 2004, Setting objective tests in mathematics using QM Perception Proc 8th CAA Conference, Loughborough, July http://www.caaconference.com

Gill, M. & Greenhow, M. 2005, Learning via online mechanics tests Proc Science Learning and Teaching Conference, Warwick, June

Gill, M. & Greenhow, M. 2006, Computer-Aided Assessment in Mechanics: what can we do; what can we learn; how far can we go? Proc IMA Conf Mathematical Education of Engineers, Loughborough, April.