Assessment item banks: an academic perspective

Dr Dick Bacon (r.bacon@surrey.ac.uk)
Senior Lecturer (HEA Consultant), University of Surrey

Background

Towards the end of 2006 the JISC CETIS Metadata & Digital Repository and Assessment SIGs, with support from the JISC Digital Repositories Programme, held a joint meeting to discuss the overlap of Assessment Item Banks (collections of question and test material) and Repositories. Following the meeting, and also with JISC DRP support, we commissioned two papers: one to describe technical aspects of item bank repositories, the other to describe the requirements from the organisational/user's point of view. We also wrote a short briefing paper on the topic.

- *What is assessment item banking?* A JISC CETIS briefing by Rowin Young and Phil Barker, Repository and Assessment domain coordinators
- *Assessment item banks and repositories* A JISC CETIS paper by Sarah Currier, Product Manager, Intrallect Ltd.
- ***Assessment item banks: an academic perspective*** A JISC CETIS paper by Dick Bacon, Senior Lecturer (HEA Consultant), University of Surrey

Contents

Introduction

The concept of an assessment item bank that can be used by academics to share assessment content within or across a range of institutions is not new, but technical developments now render such a resource far more attractive and realisable than ever before. Databanks or repositories are well established and can be accessed via a web interface for platform independence. The use of electronic assessment systems is becoming more widespread and acceptable by both staff and students. Finally, it is becoming possible to transfer questions between assessment systems with little or no change in functionality. This last factor is probably the most important in terms of the feasibility of an item bank for the sharing of questions.

This document first discusses the rationale of such systems and the usages of such electronic aids to teaching. Some aspects of the all important interoperability technology will then be described and discussed. The main assessment systems being used in HE will then be described with some details of the question types supported and how well they interoperate with other systems.

Electronic assessments

There are several electronic assessment systems being used in HE, some of which are built into the various Virtual Learning Environments (VLE) that institutions and departments are adopting. Most systems are general purpose with the expectation that most questions will be simple multiple choice or multiple selection containing text or pictures, and in practice this does indeed match many of the current requirements. Other supported questions types are listed and described below. Some other assessment systems are special purpose, particularly for mathematics and the sciences where algebraic input can be required, where equations and graphs need to be displayed, and numeric values need more functionality than is offered by most general purpose systems.

The major efficiency gain attributable to the use of an electronic assessment system is in the time saved during marking and the incorporation of those marks into student records. Other gains over conventional paper based work include factors such as the immediacy of any feedback provided, the ease of provision of supplementary materials and the randomisation of questions. This last feature can involve the random selection of equivalent questions to create an individual assessment or the randomisation of data (often numeric) within a question. When applied to assessed coursework this can be particularly beneficial, since collaboration between students can be encouraged (to discuss the generic problem) without compromising the individuality of each student's work. Such schemes clearly require a large selection of questions to choose from and this is one of the reasons why an item bank for the sharing of such questions is such a sensible idea.

Assessment item banks

Assessment item banks currently being used and developed tend to be either institutional involving several disciplines, or inter-institutional and discipline based. An institutional item bank can clearly be useful to the departments within an institution, allowing the build up of questions specific to their courses so that the need for creating new questions will eventually decline. Some sharing of questions between cognate disciplines might be possible, particularly where service teaching is prevalent, and sharing of questions by academics who would otherwise always generate their own becomes much more likely.

When creating such an institutional item bank the question of interoperability should not be ignored. It is always possible that the assessment delivery system in a given institution will be changed, for all sorts of reasons. In such circumstances the greater the investment in populating the item bank with questions, the greater the loss if the questions cannot easily be ported to another assessment system.

An institutional item bank may well have to address the problem of whether or not to include questions that are not appropriate to the institution's native assessment system. These would include questions normally to be set on paper (e.g. in examinations) but also questions requiring specialist assessment systems such as mathematical systems. Their inclusion complicates the design of the item bank, but has the advantage of keeping all assessment items accessible by a common mechanism.

It is quite obvious that the greatest benefits can accrue from having national or possibly international question item banks within specific disciplines. The creation of good questions is not a trivial task and the sharing of such investment helps reduce costs (although how the long term costs are truly to be shared within or between any national academia has not yet been addressed). Interoperability is of paramount importance, together with the quality of the searching/browsing facilities. Mechanisms will also need to be in place for turning a collection of questions from the item bank into a formative or summative assessment within any particular system.

It is probable, however, that institutional item banks will be used more for facilitating just the re-use and localisation of questions, whereas the inter-institutional item banks will support the sharing of questions for re-use. It may well be that the two will be complementary, or even merge in time with the institutional item banks becoming the nodes of a distributed item bank. The rest of this document, however, deals mainly with considerations of the inter-institutional form.

The large number of questions that would be available in an inter-institutional bank could also mean that students could be allowed access to the questions for formative or diagnostic self-assessment. That is, students could be allowed to ascertain for themselves whether they have the required level of knowledge before they sit the summative exam, or indeed before they start studying the course material. Such use of assessment is often popular with students since it helps them to manage their time by allowing them to appraise which topics they should spend most time studying and provides them with opportunities to practice solving certain problem types. This is especially important for students who are learning in environments where other forms of feedback from teachers or peers may be limited, for students who are wary about exposing their lack of knowledge to teachers or other students, and for students who are strategic learners or need to be provided with explicit motivation.

There is little doubt that the inter-institutional version should indeed contain the other question types mentioned above, such as paper based or specialist questions designed for discipline based assessment systems. The more questions that are put into such an electronically mediated repository with its sophisticated searching and browsing features, the more useful it will be. If such an item bank is successful, it will become the first place for academics to look when considering what assessments to use with their students.

Interoperability

Interoperability is the most important single factor in the implementation and use of an inter-institutional question sharing scheme, but it involves more than just the technical aspects. Questions authored without re-use in mind are likely to contain features that effectively confine them to their original context. The fact that many assessment systems now support the exchange of questions with other systems, and that item banks are beginning to be set up, should encourage authors to produce better quality assessments that are less likely to be specific to a particular course or institution.

The quality of such questions from different authors is, however, unlikely to be consistent. For example, authors are unlikely to include high quality formative feedback for questions that they are preparing for summative use. It is important, however, that questions in an item bank are consistent in their provision of feedback and marks, so that they can be used effectively in different ways and contexts. It is also important that the metadata of each record is accurate and complete, so that searching produces consistent results. This means that questions to be shared will need editing and cataloguing by a subject expert, who can find or supply any missing data or question information.

Selecting and using other peoples' questions will involve similar problems to that of choosing a text book. Differences in nomenclature, the style and approach of the questions and the explanations in formative feedback, even the style of the language used, can all lead to the rejection of individual items. It might be hoped that questions will be less of a problem than text books because their granularity is smaller, but there is little doubt that most question will need to be read in their entirety before being accepted for use. It is therefore important that the question text is viewable during the selection process.
An academic's problem with sharing questions, therefore, will depend primarily upon the quality of the questions themselves, the searching, browsing and question reviewing system and the quality of the metadata upon which the searching is based. Most academics will only invest time in the search for questions if they can be assured:

1. of their trouble-free implementation within their own institution or own assessment system
2. that the questions are likely to be a close match to what they need
3. that the assessment can be deployed without their having to learn a new technology.

Under these circumstances it is obvious that the movement of questions to a new target system must be reliable, and it must result in the same question being posed to the students and the same mark and/or feedback resulting from user input as they had been led to expect from the database information. Academics will rapidly become disenchanted if the import of questions proves unreliable, and the questions do not provide their students with the learning or assessment experiences that were expected.

Interoperability specification

The main vehicle for this interoperability is the Question Test Interoperability (QTI) specification from the IMS Global Learning Consortium[1]. Whilst the most recent release is version 2.1, version 1.2 of this specification is currently the most widely implemented, with most commercial systems used in HE claiming compliance. It was released in 2002 and existing systems support it by allowing questions to be exported or imported in QTI format. A few systems (mostly academic) have chosen to use the QTI xml format as their own internal representation.

QTI version 1.2 has been very successful in that it has been implemented widely and has been shown to support the interchange of questions between different systems in a manner that has never before been possible.[2] Where problems arise with the interchange of questions it is most frequently to do with differences in the features supported by the different assessment systems. The only questions that can be moved between two systems with no change are those using features common to the systems involved and for which the features work the same way in the two systems. Many multiple choice questions fit this category, but most other question types will produce different interactions to a greater or lesser extent, particularly in feedback and marking. Details of the differences between systems are given later in this paper.

Some commercial systems that claim compliance with the QTI v1.2 specification fail to provide details of where the compliance breaks down. For example, if a question uses a feature (like randomised values) that is not supported by the specification, then it cannot be exported in QTI compliant form. In some cases, however, such questions are exported in a QTI-like form that cannot be interpreted by any other QTI compliant system, but without warning the user. Thus, it is sensible for question authors to be aware of what question features the specification does and does not support, unless they are confident that their questions will never need to be transferred to another system.

The question types that QTI version 1 supports, and which of these are implemented in which assessment systems, is shown in the following table.

Question types	WebCT	Blackb	Moodle	QMP	TOIA	SToMP
1. Single selection from a list	Y	Y	Y	Y	Y	Y
2. Multiple selection from a list	Y	Y	Y	Y	Y	Y
3. Matrix (or pair matching)	Y	Y	Y	Y	Y	Y
4. Rank ordering		Y		Y		
5. Hot-spot questions (selecting position in an image)		Y	Y	Y		Y
6. Single or multiple text (word) entry	Y	Y	Y	Y	Y	Y
7. Single or multiple numeric entry	Y	Y	Y	Y		Y

Table A: Question types

- Question types 1 to 4 all support the optional randomisation of the order of options but also allowing individual items to be excepted from the randomisation.
- The specification is particularly flexible in its support of the checking of user responses, assigning marks and generating feedback, with arbitrary numbers of conditions and feedback items being allowed.
- Anywhere text is output to the screen it is possible to specify one or more images, sound files or video clips as well (e.g. in the question, the options or in the feedback)
- Hints are supported (but not well defined).

1. IMS Global (http://www.imsglobal.org/)
2. Pierre Gorissen, "Quickscan report", 2003 and 2006. (http://www.gorissen.info/Pierre/qti/)

- Marks can be integer or fractional and arithmetic operations are supported so that the final mark for a question can have contributions from several conditions. The final mark for each question can be given upper and lower limits.

This list of QTI features is not exhaustive, but it gives some idea of the problems that have to be addressed by a system when importing a QTI compliant question. If the system within which a question was authored supports a feature that is not supported by the target system, then a sensible interpretation of the feature into the target system cannot be assumed. Typical ways in which such situations are resolved include ignoring the feature, reporting an import or syntax error, or interpreting the feature in a way that is only sometimes appropriate. The major differences between systems that cause these sorts of problem are known, however, and systematic solutions can be incorporated into an item bank that can resolve these as question are extracted.

Another problem with version 1.2 of the QTI specification is that for some question types there is more than one way of representing what looks to the user to be the same question. Graphical hotspot questions, for example, can have the 'hot' areas defined in alternative ways within the question, requiring different conditions being used to test the user's response. Again, however, as far as a question bank is concerned, such vagaries can be resolved before questions reach users.

In 2004 work started on a completely new version of the QTI specification, designed to remove the ambiguities and to offer structure for the degree to which any particular implementation supports the specification. There have been two releases of this new version – 2.0 in 2005 and 2.1 in 2006 – and it is now up to commercial systems to take advantage of the new specification. There is some activity in this direction, but it is difficult to assess what progress has been made. It is certainly the case that at the moment interoperability can best be achieved across the widest range of platforms by using version 1.2.

Assessment systems

The general purpose assessment systems such as those that come with the Virtual Learning Environments, and others such as Questionmark Perception[3], are appropriate for use in most disciplines within UK Higher Education. Free text entry questions are supported, but unless they require only a few individual words to be recognised the text is passed on to an academic for hand marking. Some systems (e.g. QMP and SToMP[4]) have additional features that further render them appropriate for use with some specialist types of questions (e.g. in the sciences).

A recent survey of the physical sciences disciplines, for example, showed that the assessment systems most frequently used were Questionmark Perception (QMP), SToMP and Maple TA[5], together with the assessment systems from the Blackboard[6] (Bb) and WebCT[7] virtual learning environments.

Maple TA is particularly useful within mathematical tests. It can be used to set questions with randomised algebraic expressions, values and graphs. It can accept expressions entered by the user and can manipulate them using Maple's algebraic manipulator to test for equivalence with the result of a desired operation. It will not be discussed any further here, however, because it does not overlap in functionality with any of the other systems being discussed. Two further systems will be included in the discussion, however, and they are TOIA[8] and the assessment system within the Moodle[9] virtual learning environment. TOIA is included because it is a system developed by a JISC funded project and is free for the HE sector, and Moodle is included because it is open source and free, and is beginning to be taken up by some institutions.

All the assessment systems being considered here support multiple choice, multiple selection, matching, and true/false types of question. All support free text entry, although some are aimed at a single word entry (perhaps several instances in one question) and others are aimed at short text, paragraph or even essay type answers. All but TOIA support numerical questions and most support other question types that are often peculiar to their system.

---

3. Questionmark (http://www.questionmark.com)
4. SToMP (http://www.stomp.ac.uk)
5. Maple TA (http://www.maplesoft.com)
6. Blackboard (http://www.blackboard.com)
7. WebCT (http://www.webct.com)
8. TOIA (http://www.toia.ac.uk)
9. Moodle (http://moodle.org)

Unless otherwise stated, the following analysis refers to Blackboard v7, Questionmark Perception 4, SToMP v1, TOIA v1, WebCT 6 or Vista 4.

Marking and feedback

A factor that is liable to lead to much work being required at an individual question level as questions are moved from one system to another, is to do with the manner in which the questions types are supported within each system, specifically the manner in which the marks and the feedback are related to the user responses. The worth of feedback when questions are used to aid learning can depend critically upon how it is organised and how different student misconceptions can be associated with each response. The following summarises the way in which the systems handle marking and feedback for the different question types.

True and false questions will not be dealt with separately because they are a subset of simple multiple choice questions.

1. Multiple choice questions
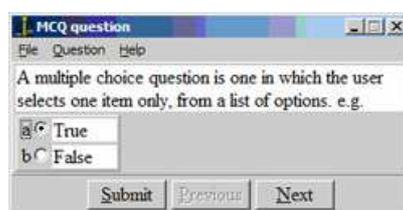
Figure 1: Multiple choice question

Question types	WebCT	Blackb	Moodle	QMP	TOIA	SToMP
Individual mark for each option	Y	Y	Y	Y	Y	Y
Individual feedback message for each option	Y	Y	Y	Y	Y	Y
Possible to have "always displayed" feedback	Y	-	-	Y	-	Y

Table B: Multiple choice question types

2. Multiple selection
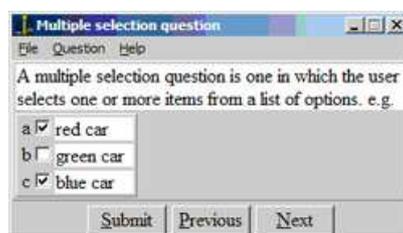
Figure 2: Multiple selection question

Question types	WebCT	Blackb	Moodle	QMP	TOIA	SToMP
Individual cumulative marks for each option	Y	-	Y	Y	Y	Y
Individual feedback message for each option	Y	-	Y	Y	Y	Y
Possible to have "always displayed" feedback	-	-	-	Y	Y	Y
Individual feedback message for each option		Y				
Possible to have "always displayed" feedback		Y				
Individual feedback message for each option	Y	Y	Y	-	-	Y

Table C: Multiple selection question types

3. Pairing questions

A pairing question is one in which the user may link one or more items in one list to items in a second list. There is a variety of ways in which different systems render such questions (see the section on *Question rendering* later).

Question types	WebCT	Blackb	Moodle	QMP	TOIA	SToMP
Individual cumulative marks for each pairing	-	-	-	Y	Y	Y
Incremental +ve and -ve marking	Y	-	-	Y	Y	Y
Incremental +ve marking	Y	Y	Y	Y	Y	Y
All or nothing marking	Y	Y				Y
Marks shared automatically between "right" options	Y	Y	Y			
Individual feedback message for each pairing	-	-	-	-	Y	Y
Only one "right" and one "wrong" feedback message		Y				
Only one "always displayed" feedback message	Y			Y		
Can limit final value when using negative marking	Y	Y	Y	-	-	Y
Can include distracters (i.e. not for selection)	-	-	-	Y	-	Y

Table D: Pairing question types

4. Rank ordering
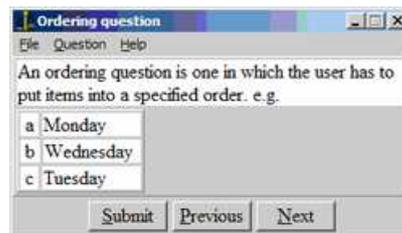
Figure 3: Rank ordering question

Question types	WebCT	Blackb	Moodle	QMP	TOIA	SToMP
Can test for more than one ordering	-	-	-	Y	-	Y
Can test for a subset of an ordering	-	-	-	-	-	Y
Only one "correct" order can be checked		Y				
Individual cumulative marks for each alternative order	-	-	-	Y	-	Y
Marks allocated proportional to number of items in correct order	Y					
All or nothing marking (alternative)		Y		Y		Y
Individual feedback message for each option	-	-	-	Y	-	Y

Table E: Rank ordering question types

5. Hotspot

Figure 4: Hotspot question

Question types	WebCT	Blackb	Moodle	QMP	TOIA	SToMP
Test for a particular marker in a particular place	-	-	-	Y	-	
Test for any marker in a particular place	-	-	-		-	Y
Individual cumulative marks for each alternative	-	-	-	Y	-	Y
Individual feedback message for each alternative	-	-	-	Y	-	Y

Table F: Hotspot question types

6. Text entry
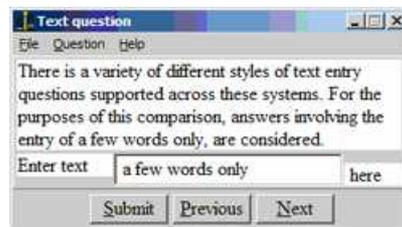
Figure 5: Text entry question

Question types	WebCT	Blackb	Moodle	QMP	TOIA	SToMP
Test for exact match, case optional, multiple templates	Y	Y	Y	Y	Y	Y
Test for inclusion in response (i.e. substring)	Y	-	-	Y	-	Y
Test can ignore misspellings	-	-	-	Y	-	-
Individual cumulative marks for each alternative	Y	-	Y	Y	-	Y
Only one "correct" mark		Y			Y	
Individual feedback message for each option	-	-	Y	Y	-	Y
One "right" and one "wrong" feedback message		Y			Y	
Only one "always displayed" feedback message	Y					

Table G: Text entry question types

7. Numeric
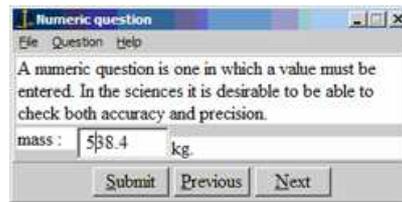
Figure 6: Numeric question

Question types	WebCT	Blackb	Moodle	QMP	TOIA	SToMP
Can have randomised values in a question	Y	Y	Y	-	-	Y
Answers can be checked for accuracy	Y	Y	Y	Y	-	Y
Answers can be checked for precision	Y	Y	Y	-	-	Y
Answers can be checked for alternative accuracies	-	-	-	Y	-	Y
Answers can be checked for alternative precisions	-	-	-	-	-	Y
Individual cumulative marks for each alternative	-	-	-	Y	-	Y
Only one "correct" mark	Y	Y	Y	-	-	-
Individual feedback message for each option	-	-	-	Y	-	Y
Only one feedback message	Y	-	Y	-	-	-

Table H: Numeric question types

Note that in the above question styles in which negative marking is supported, TOIA and QMP do not support any method for stopping the final mark for a question becoming negative.

Question rendering

Figure 7: Pair matching rendered in drop-down box style in WebCT

Figure 8: Pair matching in matrix style, rendered in TOIA
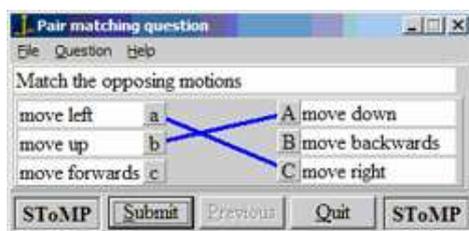

Figure 9: Pair matching rendered in SToMP
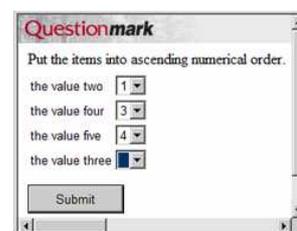
Figure 10: Rank ordering rendered in Questionmark Perception

All systems render conventional multiple choice questions using radio buttons, and render multiple selection questions using check boxes.

Pairing questions are rendered using drop down boxes in Bb, Moodle and WebCT. An example is shown in figure 7. TOIA and QMP offer drop down boxes or matrix styles as shown in figure 8. The SToMP system alone offers a visual pairing operation, as shown in figure 9.

Text and numeric entry questions vary little between systems, with a text edit box of defined style being provided. Hotspot questions in QMP require the author to provide the markers, whereas in SToMP the markers are predefined as circles. For rank ordering style questions, Bb and QMP both display the items in a fixed order, and then the user selects an ordinal position for each item from numeric drop-down lists, as shown in figure 10. SToMP rank ordering questions require the user to actually move the items into the desired order.

Sharing questions between these systems

Each of these systems is able to export questions of the above types in QTI xml format. Moodle exports in QTI v2.1, but all the others export in QTI v1.2. None of the systems except TOIA, however, export QTI in a form that can be read by any other system without some changes being required. These vary from system to system and from question style to question style. Some are due to the ambiguities inherent in the QTI v1.2 specification, others are just simple errors and others are due to proprietary styles of coding in the xml.

Most of the changes that are required are completely predictable, and can be dealt with by programming relatively simple translator programs for each combination, where there are sufficient numbers of questions to be converted to make it worthwhile. Greater difficulties lie with the changes that are required because of the different feedback and marking features of the different systems. For self assessments, individual feedback for each type of student error is clearly preferable to generic 'right' and 'wrong' answers, particularly if a student is going to try the question again. Where questions are submitted to the question bank having been prepared using systems that do not support this feature, however, the additional feedback would have to be created by a subject expert, and that can be a laborious process if many questions are involved.

Another problem with questions coming from different systems and backgrounds can involve specific environmental details that are irrelevant or wrong within a new context. These can vary from question text referring to, for example "the diagram above", which may be alongside or below in a different assessment system, to feedback referring to course notes specific to one institution. The generalisation of such detail cannot realistically be automated, and so responsibility for such changes lies with the question-bank editor in the first instance.

Questions of types other than those listed above may from time to time be offered to the question bank. Such cases will have to be dealt with in an ad hoc manner, according to the worth of the questions. In principle, most questions can be re-cast into different forms (e.g. into different question types) but there can be pedagogic differences implicit in such changes that would need to be addressed.

Conclusions

It is clear from the above that the movement of questions across a range of systems is possible, given that the questions match certain criteria. Exceptions will occur for certain question types and for certain systems, but most of these can be addressed in systematic ways.

For the two most common question types, multiple choice and multiple selection, the criteria for ease of interoperability are mostly to do with the ways in which the feedback and marking are organised. For example, a multiple choice question that has a different feedback message for each option could be moved to any testing system except early Bb systems. If a question is supplied with one 'right' and one 'wrong' feedback message, then clearly this can be converted to the more general style by replicating the 'wrong' message for each wrong option. No one form of a multiple selection question will work properly in all the relevant systems, but two versions would suffice.

Clearly, some feedback information or mark detail would be lacking for some systems, but this would have to be accepted as the consequence of choosing to use those systems.

The most effective ways of maximising the usability of the questions is thus currently seen as:

1. holding multiple versions of questions in the database where human intervention is necessary to effect the expansion or contraction of feedback information
2. holding fewer versions where the conversion of the question to a form required for particular target systems can be automated
3. holding the original version, in order to take advantage of all the information possible from the original form as systems change
4. holding questions in free-to-use on-line systems as far as possible, to minimise the work involved in getting the assessments in front of students.

The value of the question bank will be significantly increased if the submission of new questions for sharing is made as simple as possible, and so an on-line interface for such submissions should also be developed as part of any item bank project. Automated validity checks can be built into such a submission system, but since the quality of exported xml is a function of the source system being used, it must be accepted that the item bank editor is best placed to do this whilst checking each item for generality, feedback, marks and for inter-system conversions. An automated system that simply rejected badly formed questions would, it is conjectured, severely limit the quantity of question submitted and would thus damage the prospects of building a system for academics to share such academic resources.

In the longer term, numeric features written into version 2 of the IMS QTI specification may well have beneficial effects upon the use of values in all types of questions, and in particular in questions requiring numeric answers, in commercial systems. It is also likely that the use of academic or other non-VLE based systems such as SToMP, with features particularly fitted to use in the sciences, will become more widespread. The reliance upon the more basic question types and the assumed exclusivity of 'formative' or summative' assessments will also, it is hoped, be reduced as academics become more familiar with working with the tools and resources.

The creation of national assessment item banks is clearly timely for those disciplines in which the use of electronic assessment is becoming widespread. The technical means are now in place for the effective sharing of many questions between a different systems, and some specialist questions that are sparsely supported at present are likely to be more widely supported in the near future. There are significant academic efficiency gains to be obtained from the use of electronic assessment systems, and further gains can be made by the sharing of good quality questions. Students also stand to gain from the concomitant development of better quality questions for formative assessments to assist their learning as well as the better assessment of their attainments.