Assessment and Formative Feedback In Research Methods (AFFIRM)

Literature Review

By Bob Rotheram, National Teaching Fellow, Reader in Assessment, Learning and Teaching, Leeds Metropolitan University.

N.B. This is an incomplete draft, mainly written in April 2007 as part of 'AFFIRM', a small in-house Leeds Met project. It may be helpful to others, but it should not be quoted without the author's permission.

Introduction

'AFFIRM' is part of Leeds Met's efforts to promote research-informed teaching. A pilot project, it is concerned with computer-assisted assessment (CAA) related to the teaching of research methods. Over a 12-month period in 2007-8 it will create, test and evaluate a small computerised bank of quality-checked items to support undergraduate and taught postgraduate-level research methods teaching. Although the items are likely to have several potential uses, i.e. to be 'reusable learning objects' (RLOs), their primary function will probably be as components of various types of formative assessment. To enhance their effectiveness in this, particular efforts will be made to ensure items provide rich formative feedback to learners. In addition, they will contain metadata' (descriptive information about the item) to assist location and selection.

This literature review, written at an early stage, is intended to inform the design and execution of the project.

Objective testing

AFFIRM will be producing a bank of items for eventual use in 'objective tests'. Characteristically, objective tests:

> *"require a user to choose or provide a response to a question whose correct answer is predetermined"* (Bull and McKenna, 2001: 14).

It must be noted immediately that the reputation of objective testing is decidedly mixed. For a start, Seale (2002: 2) quotes the caution from the CAA Centre (now defunct) that:

> *"it is worth remembering that an objective test is only as objective as the test designer makes it."*

Putting that aside, if possible, one common and persistent reservation about CAA is to do with the suitability of multiple-choice questions (MCQs) for assessing higher-level skills. Some sources (e.g. Bates and Poole, 2003; Biggs, 2003) regard MCQs as appropriate only for testing the lowest levels of knowledge and comprehension. But other writers, such as Mogey and Watt (1996), Pritchett (1999) and Jenkins (2004), are more positive. Pritchett argues that objective questions can be set which test higher-level abilities.

> *"Producing a multiple-choice test that is both reliable and valid is a task that requires some care and skill. To produce one that can also differentiate between various levels of intellectual ability is even more demanding but, with practice, it is possible."* (Pritchett 1999: 36)

Bull and Danson (2004) observe that CAA allows a range of question types and may permit questions involving multimedia, which would not be possible if limited to the use of paper.

The reputation of computerised objective testing is partly dependent on the discipline for which it is being considered. It is more readily accepted in computing, sciences and mathematics than humanities and the arts (Bull and McKenna, 2001, McKenna, n.d.). Gipps (2005) notes that CAA involving multiple-choice and short-answer questions, both of which can be marked automatically, is more likely to be used in disciplines relying heavily on factual information, such as geography, mathematics and engineering. However, she reveals something of her attitude towards objective testing when she says, *"there is the potential to go much wider than this"* (Gipps 2005: 173).

At the positive end of the spectrum of opinion on objective testing:

> *"The most optimistic view is that item-based testing may be appropriate for examining the full range of learning outcomes in undergraduates and postgraduates, provided sufficient care is taken in their construction"* (Conole and Warburton, 2005: 21).

Yet even if one only accepts objective testing of knowledge and understanding, this will often have value:

> *"In many courses there is a body of underpinning knowledge which must be learned to enable progression during the later stages of the course"* (Bull and McKenna, 2001: 10).

Interestingly, Biggs is prepared to make a virtue of the limitations (in his eyes) of objective testing. He notes a common concern about using objective CAA summatively – that given time and access to the item bank, students can rote learn the correct responses. However, he shrugs this off by advocating the use of CAA systems *"precisely for those items that require rote learning, such as terminology, rules, etc."*, whilst perhaps requiring a high pass mark (Biggs, 2003: 223).

Multiple-choice tests are often criticised on the grounds that chance can play a large part in the successful completion of the examination (Pritchett, 1999). For example, true-false questions give a 50:50 chance of success, and in MCQs with four options there is a 25% chance of getting the right answer by guessing. However, the literature shows that there are ways of mitigating this problem, some of which will be outlined later.

Hence it may be seen that the reputation of objective testing is mixed. Reservations are primarily about its ability to test skills other than knowledge and comprehension. This type of testing, though, is likely to have some value on many programmes. Another noteworthy fact about objective testing is that it is more readily accepted in some disciplines than others. How acceptable it is to teachers of research methods is at present unclear.

Item banks

The terms *item* and *question* are often used interchangeably in discussions about CAA (Sclater 2005). However, Sclater claims the CAA community generally prefers the term *item*, because, along with the question, it may incorporate other elements including answers, feedback, scoring information and perhaps metadata describing it. Consistent with this, items may be stored together in an *item bank*, which may be defined as:

> *"a collection of items for a particular assessment, subject or educational sector, classified by metadata which facilitates searching and automated test creation"* (Sclater, 2005: 1).

This review, and the AFFIRM project, will adopt this usage.

Still on the matter of terminology, Sclater (2005) notes that the JISC-funded Item Bank Infrastructure Study (IBIS) called a group of items gathered together for a particular test a *pool* in preference to the term *assessment* because it may be seen as more flexible. Thus a pool could contain a substantial number of items of similar difficulty and a subset might be drawn at random from the pool for an assessment.

A number of item banks already exist in the higher education sector. Disciplines covered include: electrical and electronic engineering; mathematics; economics; mathematics in economics; medicine; pharmacology; chemistry; bioscience. Also some publishers (e.g. McGraw-Hill) are beginning to produce banks of questions to accompany particular textbooks. Of particular relevance to AFFIRM are the 260 questions on the Oxford University Press website, supporting a research methods textbook (Bryman, 2004). In addition, the Multimedia Educational Resource for Learning and Online Teaching (MERLOT) (www.merlot.org) when searched on "research methods quiz" provides gateways to:

- a tutorial and online quiz on sampling in social research (Shafie, 2005);
- activities and quizzes related to conducting psychological research (Bradley, 2006);

Many regard item banks as having an important role to play in the wider adoption of CAA. Conole and Warburton (2005) report their own 2003 survey of 50 respondents, mostly 'academic enthusiasts', a number of whom cited subject-specific shared question banks and the value of exemplars as *"important drivers for the large-scale uptake of CAA."* Bull (2000) was an early advocate of nationally coordinated and supported item banks to overcome issues of security, copyright and organisation, all of which are potential obstacles to take-up and effective use.

Bull and Danson (2004) are of the opinion that institutional item banks have the potential for sharing material across traditionally disparate departments where teaching content is shared. [N.B. This is likely to be the case with research methods teaching at Leeds Met and many other universities.] They say that, if successful, item banks can be hugely beneficial to both new and established academics. Sharing, reuse and adaptation can, of course, also help to overcome some of the upfront time investment in creating items and pools. However, Bull and Danson strike a note of caution, saying that peer review of questions is essential for sharing ownership of item banks across the community. This is obviously related to the issue of quality, which is discussed further below.

Other advantages of item banks include the possibility of grading items according to difficulty, perhaps allowing the creation of learning hierarchies and better ways of structuring the curriculum (Bull and Danson, 2004).

Grading items and describing them in various ways can also make them more discoverable and more likely to be reused. Garrison and Anderson's (2003) favoured way of doing this is via 'metatags' and including the items in a repository of educational objects. Tags commonly include: difficulty of the question, topic, academic level, and the skill or knowledge component addressed (Bull and Dalziel, 2003). Tagging can lead to relatively sophisticated 'computer-adaptive testing' (CAT) where the questions put to individual students depend on their previous responses, thus tailoring the test to their level of ability (Bull and Dalziel, 2003). Metadata can also facilitate the exchange of items between repositories (Sclater, 2005).

Ownership, copyright and intellectual property *"raise a cluster of thorny issues"* (Bull and Dalziel (2003: 177). One solution they offer is using 'open source software'-style (OSS) licences, such as the General Public Licence (GPL), sometimes known as 'copyleft'. This approach was used by Rotheram with his Social Policy Question Resource (Rotheram, 2005). However, as Bull and Dalziel (2003) note, it may conflict with an institution's intellectual property policy. They propose that copyright, intellectual property and some other, related issues be addressed by institutions on a national and international level. Sclater (2005) is another commentator recommending clarity both about ownership and usage. Bull and Danson (2004) mention one bank (economics) where growth was hindered by problems with intellectual property rights.

Sclater (2005) says that if items are to be shared it is important to take steps to avoid them being plagiarised. In

his view, potential authors should, at the very least, be reminded of their responsibility not to plagiarise the work of others.

Finally in this section, it must be noted that an item bank requires maintenance (Rotheram 2005). The size of this task varies according to discipline. In some subject areas, e.g. social policy, many items may have a short 'shelf life', for various reasons: they soon become outdated; 'correct' answers become incorrect in the face of new developments; web links in questions or feedback no longer work. In contrast, most mathematics items will have an extremely long 'shelf life' and require little maintenance. We may speculate at this stage that this may be true of mathematical, quantitative research methods items. However, the AFFIRM bank is likely also to contain items related to qualitative research methods. How much maintenance will they require?

Question-setting: general issues

The literature consistently makes the point that question-setting is a skilled job (see, for example, Pritchett, 1999; Bull and McKenna, 2001; Rotheram, 2005) and is likely to remain so (Conole and Warburton, 2005). It takes time and practice to avoid elementary mistakes, to go beyond simple probing of factual knowledge and produce clearly expressed, worthwhile questions with indisputably correct answers (Zakrzewski, 1997; Bull and Dalziel, 2003; Gipps, 2005; Rotheram, 2005).

Therefore staff beginning to set objective questions are likely to benefit from training and support. Indeed Bull and Danson (2004) relay the view of Boyle and O'Hare (2003) that training in item construction and analysis should be obligatory for staff involved in developing CAA tests and that items should be peer-reviewed and trialled before use. In addition, staff will need time to become skilled, create a stock of quality items and become familiar with computer systems. However, as Gipps (2005) observes, these requirements are often underestimated, which rather reduces the claims some make for the efficiency of CAA.

Obtaining items can be difficult. Rotheram (2005) reported little success in drawing in other question-setters, despite offering £15 per accepted item. That said, Harvey and Mogey 1999) appear to have fared better. They recommend several ways of gathering contributions to item banks:

- getting students to set up a FAQ;
- ask other students to research the answers;
- ask yet other students to provide feedback for others on where their peers are going wrong;
- inviting colleagues (perhaps across institutions) to provide a set number of questions each;
- go to staff development events on writing good questions;
- searching publishers' sites.

An apparently successful, idea, adopted by the University Medical Assessment Project (UMAP), is to hold question-setting workshops. It reports (UMAP, 2006) that more than 50 workshops have helped amass over 5,000 assessment items, including multiple-choice and extended matching set questions (see 'Question types', below).

On what kinds of topic should items be created? Beginning with very general considerations, the central thesis of Biggs (2003), a widely-respected author, is that assessment should be 'constructively aligned' with all other elements of the programme. Another guiding principle, surely, is the argument of Bates and Poole (2003) that the form of assessment should match the skills being taught.

Moving to more concrete matters, Race (2007) recommends thinking about which parts of the curriculum best lend themselves to resource-based learning (about which, computerised questions may be devised). He suggests they may include:

1. Important background material.
2. 'Need to know before …' material.
3. 'Remedial material'.
4. 'Nice to know' material.
5. Much-repeated material.
6. Material which is best 'learned-by-doing'.
7. Material where students need individual feedback on their progress.
8. Material you don't like to teach!
9. Material that students find hard to grasp first time.
10. Material which may be needed later, at short notice.

One implication of these recommendations is that question-setters should know well the programme(s) and module(s) on which their questions might be used.

Another important consideration is educational *level*. What intellectual challenges are appropriate for the learners for whom the questions are being posed? So far, this has been mentioned only in passing, with references to "higher-level skills" and "the lowest levels of knowledge and comprehension." In discussing levels, the educational literature (including that on CAA) frequently refers to Bloom's (1956) taxonomy of educational objectives. In summarising this, Pritchett (1999: 33) points out that it is a hierarchy and each level incorporates those below it:

1. *Knowledge.* Ability to recall previously learned material, e.g. theories, terms, conventions, classifications, categories, principles, methodologies.
2. *Comprehension.* Understanding the meaning of learned material, to be able to translate or convert it from one form to another, to interpret material and to be able to extrapolate data.
3. *Application.* Ability to apply what has been learned in one context to another situation.
4. *Analysis.* Ability to identify component parts of material, their relationships and the principles underlying their organisation.

5. *Synthesis.* Drawing together material from different sources to produce a unique item such as a plan or theory.
6. *Evaluation.* Ability to estimate the appropriateness of a certain item according to particular criteria.

Conole and Warburton (2005: 20) say that outcomes at the lower end of Bloom's taxonomy are traditionally assessed on a convergent basis (i.e. only one 'correct' answer), while higher-order outcomes are most readily assessed divergently (a range of responses and analyses is permissible). However, Pritchett (1999: 33-4) offers some examples of questions which probe different levels of Bloom's taxonomy, bolstering her argument that it is possible for objective tests to go beyond probing knowledge and comprehension. In similar vein, Bull and McKenna (2001: 17) suggest some 'question words' appropriate for each of Bloom's levels. Passages from both these excellent sources are reproduced extensively in Appendix 1.

Question types

Bull and McKenna (2001) identify several question types as being suitable for CAA, including:

- multiple choice
- true/false
- assertion/reason
- multiple response
- graphical hotspot
- text/numerical (involve input of text or numbers at keyboard)
- matching
- 'sore finger' (something is out of keeping with the rest)
- ranking
- sequencing

A more advanced question type discussed by Bull and McKenna (2001) is the 'multiple true/false' (MTFQ) in which students are presented with a set of data followed by three or more statements. The student is required to determine whether each statement is true or false. Bull and McKenna (2001: 32) say a series of these questions on a specific topic can test a more complex understanding of an issue.

Bender (2003) and Dennis et al. (n.d.) also consider Extended Matching Set Questions (EMSQ). Here, a list of possible answers (options) are supplied with an associated list of questions. Both lists may be quite long and the same option may be used once, more than once, or not at all. The claimed advantages of EMSQs include less 'cueing' of the correct answer than with other question types such as MCQs and fewer problems with writing plausible distracters.

Harvey and Mogey (1999) encourage variety, incorporating a range of multimedia and objective types when question-setting.

More information on question types and advice on how to set them is to be found in Appendix 2.

Feedback

Ramsden (2003: 187) in his general text on learning and teaching in higher education says:

> *"It is impossible to overstate the role of effective comments on students' progress in any discussion of effective teaching and assessment."*

In their text on CAA, Bull and McKenna (2001: 38) observe:

> *"Feedback helps to motivate students to learn and needs to be timely and constructive."*

Computers, of course, can provide extremely timely (instant!) feedback. This is one of the principal advantages of CAA and can benefit all parties. Pritchett (1999) encourages using the ability of computers to give rapid feedback to students on individual performance and to lecturers on the performance of the group, effectiveness of the examination of individual questions.

But speed is not the only benefit of computerised objective testing. Jenkins (2004) thinks CAA (in various forms, not just MCQs) can be used in a wide range of contexts, especially for formative feedback. Bull and Danson (2004) emphasise CAA's ability to provide diagnostic and formative feedback and enhance student learning in ways which are not possible with paper-based assessments. Exemplifying this, Race (2007) suggests computer feedback could be done by a sound file, exploiting the benefits of tone-of-voice. Independently, Rotheram (2007) has experimented successfully with directly recording feedback to students (albeit on essays rather than objective tests) as MP3 files, which are highly portable. Student reaction to the experiment was very positive.

For more detailed advice on providing feedback in objective testing, please see Appendix 3.

Quality

The point has already been made that items and item banks are more likely to be adopted if they are of high quality. If it is a matter of priorities, Harvey and Mogey (1999) recommend choosing quality rather than quantity. Establishing an editorial board early in the life of a project provides an important quality control mechanism, helping to raise the standard of the items (Rotheram 2005).

Bull and Dalziel (2003) note that questions being added to a bank are typically piloted prior to inclusion, to allow statistical measures to be gathered. They suggest a routine process of piloting a few new questions each time a test is delivered, allowing the bank to grow, whilst ensuring the quality and appropriateness of items. This is particularly important if high-stakes summative assessment is to be attempted (Bull and Danson, 2004).

Conole and Warburton (2005) and Bull and Danson (2004) recommend using the capabilities of CAA software to analyse how students, items and tests perform. Software commonly allows analysis of questions to find those which are best at discriminating between the top and bottom of classes, and those which need some reworking.

Conole and Warburton (2005), Sclater (2005) and Bull and Dalziel (2003) mention Classical Test Construction and Latent Trait Analysis as methods of studying the statistics on tests and items respectively. These may allow judgements to be made on whether particular questions are too easy or difficult for particular groups or whether some distracters within question are implausible (e.g. because they are rarely selected) or to give an indication of common misconceptions (Bull and Danson 2004).

McAlpine (2002) recommends that assessments overall, and individual items carrying a high proportion of the overall marks, should have a difficulty level of about 0.5, so that the mean mark is about half of the marks available. When a test is comprised of individual items worth a low proportion of the total marks, she suggests they should vary in difficulty, so that candidates of all abilities may be fully tested. However, items should not have 'facility values' (indicators of difficulty) above 0.85 or below 0.15 because beyond these limits *"they are contributing little to the measurement of the candidates"* (McAlpine 2002:15).

Scoring

One problem with objective testing, highlighted earlier in this review, is that of candidates guessing. Pritchett (1999) offers some strategies for coping with guessing:

- Increase the number of choices in MCQs, but she advises against having more than six choices – it is difficult to come up with sufficient plausible 'distracters' (wrong answers). Also, a large number of choices could become confusing to candidates – it increases the time needed for reading and decision-making.
- Use multiple-response questions with more than one correct answer and require candidates to select several of the options.
- Penalising wrong answers.

Conole and Warburton (2005) note that concerns about students guessing answers are dealt with in two main ways:

- discounting a test's guess factor;
- adjusting the marking scheme away from 'one correct answer equals one mark' to include negative marking.

One method of discounting guessing is to score questions as normal and apply a formula for guess correction at the end. A standard formula, for MCQs all with the same number of options, is:

SCORE = R-W/(N-1)

where R=number of correct answers; W=number of incorrect answers; N=total number of options per item (including the correct answer) (Bull and McKenna, 2001: 41). Alternatively, a relatively high mark may be required in order to pass. However, Bull and McKenna (ibid.) point out that corrective scoring is controversial and some practitioners feel it is unnecessary. They observe that even if all questions are true/false the chance of obtaining at least 70% by random selection in a 45-question test is less than 1 in 300.

Negative marking can be done in various ways. A simple system would be to award +1 for a correct answer, 0 for no answer, -1 for a wrong answer (Bull and McKenna, ibid.) A more complex approach is *confidence-based marking* (Bender, 2003; Conole and Warburton, 2005), where marks are awarded for a response predicated on a student's confidence that the correct response has been given. Bender (2003) discusses 'confidence assessment' as part of the marking scheme for MCQs and MTFQs. The student not only has to select the correct answer but must also express his/her confidence in the answer: unsure scores +1 if correct and 0 if incorrect; fairly sure scores +2 or -2; a high level of confidence scores +3 if correct but a penalty of -6 if incorrect.

The rationale for negative marking is that in some settings (e.g. clinical practice) guessing should be dissuaded and that being confident but incorrect can literally be lethal. On the other hand, it could be perceived as harsh and may deter answers based on near-certain knowledge as much as wild guesses (Pritchett, 1999).

Whatever marking system is to be used, it needs to be communicated clearly to students (Pritchett, 1999; Bull and McKenna, 2001; Laurillard, 2002).

Computerisation

Computerising objective tests brings a crop of issues. Arguably the most important step is to use software that allows easy export and import of items to and from other packages (Rotheram 2005). For this to happen smoothly, *interoperability* is crucial. Interoperability will help:

- if an institution changes its virtual learning environment (VLE);
- to enable sharing of items between institutions;
- use of communal resource repositories.

Other reasons in favour of interoperability are (Conole and Warburton 2005: 23) to:

- enable student assessment data to be transferred to institutional student record systems – a point also made by Bull and Dalziel (2003);
- preserve users' investments in existing questions and tests when moving to different institutions or to different CAA systems.

Sclater (2005) reported that some projects surveyed by IBIS had items trapped in proprietary formats and lacking adequate metadata, fundamentally limiting their wider use.

How might interoperability be achieved? Conole and Warburton (2003: 23) take the view that a good starting point is for software to be 'IMS QTI-compliant' (IMS Global, 2000). The QTI specification separates questions from presentation, allowing questions which meet the standard to be imported and exported between IMS-compliant systems (Bull and Danson 2004).

Rotheram (2005) commented that Questionmark Perception (www.qmark.com) is fairly good but not perfect in this respect. The JISC-funded 'TOIA' (www.toia.ac.uk) and 'RELOAD' (www.reload.ac.uk) projects and the JORUM repository (www.jorum.ac.uk) are taking IMS QTI-compliance into account and may therefore become the tools of choice for many UK HE institutions (Rotheram 2005).

Respondus (available at Leeds Met) is a tool for creating and managing tests. It claims to be IMS QTI-compliant and therefore suitable for importing and exporting items (e.g. from and to Leeds Met's VLE, WebCT). Rotheram's (limited) experience suggests that its performance in import/export is less than perfect, even with basic item types such as multiple-choice and multiple-response. Therefore users need to check carefully how faithfully items have been transferred, and perhaps make corrections/adaptations.

It should be noted that achieving interoperability is an active field of development and the goal has by no means been achieved. For an indication of some of the issues confronting researchers in this area, see Olivier and Liber (2003).

To enhance student learning it is important to use software which allows rich feedback on each question. Questionmark Perception performs quite well in this respect (Rotheram, 2005), but some packages, e.g. Hot Potatoes (http://hotpot.uvic.ca/), only provide an indication of whether an answer is correct. As for Respondus, Rotheram has found that it appears not to be able to render fully to test users the feedback in imported items that it was clearly storing. Given the importance of feedback, this is not a trivial matter.

Bull and Danson (2004) note the limited CAA functionality of widely-used virtual learning environments (VLEs) such as Blackboard and WebCT. However, they acknowledge the ability of VLEs to:

- quickly provide self-assessment questions which can be incorporated in other learning materials;
- introduce CAA into the curriculum;
- (usually) provide a simple interface for question construction.

Whatever system is used, the advantages of 'single sign on', allowing students to move easily from a VLE to the CAA system, are recognised by Bull and Dalziel (2003).

Thought needs to be given to assessment delivery (Bates and Poole, 2003). For example, what feedback to give, and when; whether to allow repeated attempts; whether to block progress until some content has been mastered; who sees the results (student, teacher or both).

Then there are issues of security. Web-based systems rely on web browser applications so security issues, such as invoking a secure connection, should be addressed. In addition, how to be sure that the person answering is the registered student? This is mentioned by Garrison and Anderson (2003: 101), who suggest that developments in computer security (e.g. in the airline industry) will probably find their way into educational systems.

Also, efforts must be made to ensure that assessments look and behave as expected across browsers and platforms (Bull and Danson 2004: 16).

Part IV of the Disability Discrimination Act 1995 requires educational establishments to make 'reasonable adjustments' (which should be 'anticipatory') for disabled people. Therefore online assessments should be as accessible as reasonably possible and the best advice, e.g. from TechDis (n.d.), is that accessible materials benefit all learners and that accessibility should be designed in from the beginning.

Research methods teaching

[To be added later]

Implications for AFFIRM

- It will be important to explore, at various stages of the project, the attitudes of staff and students towards the use of CAA to assist the teaching of research methods.
- An editorial board should be appointed.
- Consideration should be given to the use, if any, of existing resources, especially those associated with Bryman (2004).
- Item-creators will need to be recruited and trained.
- Ways of encouraging and facilitating contributions will need to be considered. Item-writing events might be fruitful. Could students be rewarded for providing items which are accepted?
- Creativity should be encouraged, as should use of as wide a variety of items as possible.
- The importance of incorporating rich feedback should be emphasised.
- It may be worth trying to enrich feedback by giving some of it via sound files.
- Questions should be confined to issues on which there are unambiguously right or wrong answers.
- The project should focus on the testing of knowledge and understanding, where creation of

questions is easiest. Assess higher-level skills in other ways.

- The scoring of items requires careful thought, especially whether to build in penalty marking and confidence-based marking. [Rotheram suggests that AFFIRM doesn't do either of these things because research methods questions do not seem to raise any special issues. Guessing may be coped with in other ways, e.g. by requiring a fairly high pass mark in tests.]
- There should be further consideration of the software to be used to create, store, deliver items, pools and tests. Some key questions are:
  - How adequate are Respondus and WebCT?
  - Might use be made of the Leeds Met repository being created as a JISC project led by Wendy Luker?
  - Can Prof Janet Finlay provide useful advice?
  - Are the resources created by the TOIA project likely to be useful, if still available? (TOIA's funding has now ceased.)
- Other software issues to explore include:
  - interoperability;
  - IMS QTI compliance;
  - recording of results;
  - integration with institutional systems for student assessment records.
- AFFIRM should take care from the outset to make items as accessible as reasonably possible. There will be a need for comparison against TechDis guidelines and for testing, e.g. with JAWS software.
- Metadata should be incorporated from the outset. The project will need to study how best to do this. The 'RELOAD Editor' may be of use. See: http://www.reload.ac.uk/ex/ReloadQSv1.pdf [Accessed 2.4.07]
- Evaluation of items should be planned for. What is to be evaluated? How? When? Timing will require care when students are involved.
- Intellectual property issues should be considered carefully. [Rotheram's preference is for operating in a spirit of openness and collaboration with staff at institutions, with some form of 'copyleft' arrangement.]

References

Bates, AW and Poole, G (2003) *Effective Teaching with Technology in Higher Education: Foundations for Success*, San Francisco, Jossey Bass.

Bender, DA (2003) "MCQ, EMSQ or multiple true/false questions?", *Bioscience Education e-journal*, Volume 2, November. Available online at http://bio.ltsn.ac.uk/journal/vol2/beej-2-L1.htm [Accessed 25.3.07]

Biggs, J (2003) *Teaching for Quality Learning at University* (2nd ed.), Maidenhead, Open University Press.

Bloom, BS et al. (1956) *Taxonomy of Educational Objectives: Cognitive domain*, New York, David McKay Co. Inc.

Boyle, A and O'Hare, D (2003) "Assuring quality computer-based assessment development in UK higher education" – in Christie, J (ed.) *7th International CAA Conference*, Loughborough University, 8-9 July 2004.

Bradley, ME (2006) *Cyberlab for Psychological Research: activities and quizzes*, Available online at: http://faculty.frostburg.edu/mbradley/activities.html [Accessed 2.4.07]

Bryman, A, (2004) *Social Research Methods* (2nd ed.), Oxford, Oxford University Press. Associated collection of multiple-choice questions available online at: http://www.oup.com/uk/orc/bin/9780199264469/01student/mcqs/ [Accessed 29.3.07]

Bull, J (2000) *The Implementation and Evaluation of Computer-Assisted Assessment*, CAA Centre, University of Luton. Available online at: http://caacentre.lboro.ac.uk/dldocs/Annual_Report_2K.pdf [Accessed 24.3.07]

Bull, J and Dalziel, J (2003) "Assessing question banks" – in Littlejohn A (ed.) *Reusing Online Resources: a sustainable approach to e-learning"*, London, Kogan Page.

Bull, J and Danson, M (2004) *Assessment Series No. 14: Computer-assisted Assessment*, York, LTSN Generic Centre. Available online at: http://www.heacademy.ac.uk/embedded_object.asp?id=20388&prompt=yes&filename=ASS093 [Accessed 22.3.07]

Bull, J and McKenna, C (2001) *Blueprint for Computer-Assisted Assessment*, Luton, Computer-Assisted Assessment Centre. (Republished in 2004 by RoutledgeFalmer)

Conole, G and Warburton, B (2005) "A review of computer-assisted assessment" – in *ALT-J, Research in Learning Technology*, Vol. 13, No.1, March, pp.17-31.

Dennis, L et al. (n.d.) *Extended Matching Questions (EMQ)*, Available online at: http://www.cs.nott.ac.uk/~smx/PGCHE/EMQ.html [Accessed 25.3.07]

Garrison, T and Anderson, T (2003) *E-Learning in the 21st Century*, London, RoutledgeFalmer.

Gipps, C (2005) "What is the role for ICT-based assessment in universities?" – in *Studies in Higher Education*, Vol.30, No.2, April, pp. 171-180.

Harvey, J and Mogey, N (1999) "Pragmatic issues when integrating technology into the assessment of students" – in Brown, S, Bull, J and Race, P *Computer-Assisted Assessment in Higher Education*, London, Kogan Page.

IMS Global Learning Consortium, (2000) *IMS Question & Test Interoperability Specification: A Review*. Available online at: http://www.imsglobal.org/question/whitepaper.pdf [Accessed 24.3.07]

Jenkins, M (2004) "Unfulfilled Promise: formative assessment using computer-aided assessment" – in *Learning and Teaching in Higher Education*, Issue 1, pp. 67-80.

Laurillard, D (2002) *Rethinking University Teaching: a conversational framework for the effective use of learning technologies*, London, RoutledgeFalmer.
McAlpine, M (2002) *Principles of Assessment*, CAA Centre, University of Luton, Bluepaper No.1, February. Available online at: http://www.caacentre.ac.uk/dldocs/Bluepaper1.pdf [Accessed 28.3.07]
McKenna, C (n.d. but probably 2001) *Academic Approaches and Attitudes Towards CAA: A Qualitative Study*, University College, London. Available online at: http://magpie.lboro.ac.uk/dspace/bitstream/2134/1821/1/mckenna_academic.pdf [Accessed 17.1.07]
MERLOT (Multimedia Educational Resource for Learning and Online Teaching):. Available online at: www.merlot.org [Accessed 29.3.07]
Mogey, N and Watt, H (1996) *The use of computers in the assessment of student learning*, Learning Technology Dissemination Initiative, Heriot-Watt University. Available online at: http://www.icbl.hw.ac.uk/ltdi/implementing-it/using.pdf [Accessed 24.3.07]
Olivier, B and Liber, O (2003) "Learning content interoperability standards" – in Littlejohn, A (ed.) *Reusing Online Resources: a sustainable approach to e-learning*, London, Kogan Page.
Pritchett, N (1999) "Effective question design" – in Brown, S, Race, P and Bull, J (eds.) *Computer-Assisted Assessment in Higher Education*, London, Kogan Page.
Race, P (2007) *The Lecturer's Toolkit: a practical guide to assessment, learning and teaching* (3rd ed.), London, Routledge.
Ramsden, P (2003) *Learning to Teach in Higher Education* (2nd ed.), London, RoutledgeFalmer.
Rotheram, B (2005) *Social Policy Question Resource (SPQR): a review.* Available online at: http://www.swap.ac.uk/docs/learning/SPQRreport.doc [Accessed 25.3.07]
Rotheram, B (2007) "Using an MP3 recorder to give feedback on student assignments", *Educational Developments*, 8.2, pp.7-10, London, SEDA.
Sclater, N (ed.) (2005) *Item Banks Infrastructure Study (IBIS): Executive Summary*, Bristol, JISC. Available online at: http://www.toia.ac.uk/ibis/IBIS-Executive-Summary.pdf [Accessed 28.3.07]
Seale, J (2002) *Using CAA to support student learning*, York, LTSN Generic Centre. Available online at: http://www.heacademy.ac.uk/embedded_object.asp?id=17325&prompt=yes&filename=ELN004 [Accessed 23.3.07]
Shafie, D (2005), *Sampling in Social Research.* Available online at: http://teach.citl.ohiou.edu/sampling/ [Accessed 2.4.07]
TechDis (n.d.) *Staff Pack on e-Assessment*, http://www.techdis.ac.uk/resources/sites/staffpacks/Staff%20Packs/E-Assessment/index.xml [Accessed 28.3.07]
UMAP (2006) *Workshop output summary*. Available online at: http://www.umap.org.uk/workshops/ [Accessed 2.4.07]
Zakrzewski, S (1997) *The Luton Experience*, Learning Technology Dissemination Initiative, Heriot-Watt University. Available online at: http://www.icbl.hw.ac.uk/ltdi/assessit/luton.htm [Accessed 24.3.07]

Appendix 1: Using Bloom's Taxonomy

The material here is from Pritchett (1999: 33-4), except for the 'question words', which are taken from Bull and McKenna (2001: 17).

Knowledge-testing

- What word means the same as…?
- What is the most important difference between…?
- Which one of the following sequences shows the correct order of…?
- What are the major classifications of…?
- Which method is the most useful for…?
- What evidence best supports the theory of…?

Question words: list, define, label, describe, name.

Comprehension-testing

The possibilities which appear here presume that the appropriate answers have not been taught or discussed in class, otherwise they would be merely knowledge recall items.

- Which one of the following is closest in meaning to the term…?
- The statement '…' means that…
- (Various facts are presented.) Which of the reasons listed below best explains this?

Question words: interpret, discuss, predict, summarise, classify.

Application

This involves being able to apply what has been learned about one set of circumstances to another set of conditions.

- Indicate under which of the following circumstances a problem-solver will be able to use prior experience.

Question words: apply, demonstrate, show, relate.

Analysis

Analysis of parts/elements includes recognising unstated assumptions, distinguishing fact from opinion, distinguishing conclusions from the supporting facts.

Analysis of relationships includes identifying cause-and-effect relationships, relationships between ideas, distinguishing between relevant and irrelevant arguments.

Analysis of organising principles includes such abilities as recognising an author's point of view, theoretical perspective or purpose.

- (An attributed statement is made) Which one of the following assumptions is being made by the author?
- Read the following two statements and select the answer which best expresses their relationship.
- Which one of the following best expresses the perspective of the author?

Analysis may be better expressed with a larger piece of stimulus material about which various questions may be posed.

Question words: analyse, arrange, order, explain, connect, infer, compare, categorise.

Synthesis
This involves the production of some novel response based on material drawn from several sources. It is most easily done through free-response methods, but it can be done through MCQs. It is probably more easily assessed objectively in some disciplines than in others. As with other complex skill levels, it may be best to provide a piece of stimulus material around which a set of questions may be posed.

Question words: integrate, modify, invent, design, compose, plan, formulate, arrange.

Evaluation
Can be assessed to some extent by objective testing, especially over a series of MCQs. However, some aspects may be better tested in a less constrained format, e.g. a longer piece of stimulus material on which a set of questions is based. Evaluation could be based on the internal content of the stimulus or on external evaluative criteria. In either case, the student should be required to identify explicit criteria for use in the evaluation.

Question words: appraise, judge, evaluate, defend, rank, conclude, discriminate, recommend.

Appendix 2: Some question types and how to set them

Effective design of *multiple-choice* questions (Pritchett 1999: 30-32):

1. Construct each item to test an important learning outcome. Avoid testing for trivial details.
2. The stem of the item should contain only one question.
3. Use simple and clear expression.
4. Put as much of the wording as possible in the stem of the question.
5. There should be only one correct answer.
6. Give the stem of the question in the positive form and avoid negatives. If a negative expression must be used it should be emphasised by the use of capitals or underlining.
7. Make sure that all the answer choices are grammatically consistent with the stem of the question and that they are all parallel in form. Clues may be given in various ways, e.g.:
   - the use of 'a' or 'an' at the end of a stem if some of the options begin with a vowel;
   - similarity of wording in the stem and in the correct answer;
   - stating the correct answer in textbook language or stereotyped phraseology may cause it to be selected because it looks better;
   - stating the correct answer in more detail than other choices;
   - including modifiers (may, usually, sometimes) may cause an item to be chosen;
   - including absolutes (always, never, all) in the distracters may allow them to be rejected easily;
   - including two all-inclusive statements allows the others to be rejected because at least one of the two must be correct;
   - including two responses with the same meaning allows them to be rejected since clearly they cannot both be correct.
8. Avoid verbal clues which may enable selection of the correct answer or rejection of incorrect alternatives.
9. Make all the responses approximately the same length. An option which is longer than the rest may give a clue that it is the correct answer.
10. Avoid the use of 'all of the above' or 'none of the above'. They may allow the correct choice to be made on the basis of incomplete information or not knowing the correct answer.
11. Vary the position of the right answer in a random way.
12. Locate questions on the same topic together.
13. Make sure that each item is independent of other items in a test. One item should not give clues to the correct answer in another. Nor should correct answers in one item depend on correctly answering a previous one.
14. The layout of the questions should be clear. List alternatives on separate lines and use letters rather than numbers for them. If the stem is an incomplete statement, the choices should begin with lower case letters and end with a question mark.

In relation to MCQs, Bull and McKenna (2001: 25-6) also advise:

- avoid unnecessary and irrelevant material;
- put as of the question as possible in the stem;
- distracters based on common student errors or misconceptions are very effective;
- correct statements that do not answer the question are often strong distracters;
- do not create distracters which are so close to the correct answer that they may confuse students who really know the answer.

With *multiple-response* questions, Bull and McKenna (2001: 27) point out that withholding the number of correct responses makes guessing the correct answers more difficult.

Matching questions are particularly good at assessing a student's understanding of relationships. Bull and

McKenna (2001: 29) offer several tips for writing good matching questions:

- provide clear directions;
- keep the information in each column as homogeneous as possible;
- allow the responses to be used more than once;
- arrange the list of responses systematically if possible (chronological, alphabetical, numerical);
- include more responses than stems to help prevent obtaining the answer by a process of elimination.

True/false questions, say Bull and McKenna (2001: 30) have limitations including: 50% chance of guessing the right answer; difficulty in writing answers which are unambiguously true or false; they don't discriminate well between students of different abilities. Even so, they offer some suggestions for writing true/false questions:

- include only one main idea in each item;
- (as with MCQs) use negatives sparingly;
- try using them in combination with other material, such as graphs, maps, written material;
- use statements which are unambiguously true or false;
- avoid lifting statements from assigned reading, lecture notes, etc. which might allow simple recall to give the correct answer;
- avoid using words which signal the correct answer. E.g. 'none', 'never', 'always', 'all', 'impossible' tend to be false, while qualifiers such as 'usually', 'generally', 'sometimes', 'often' are likely to be true.

Text match response questions (also known as fill-in-the-blank or gap-fill) have the advantage of requiring the student to supply the correct answer, so are less likely to be able to guess the correct response. However, care needs to be taken when programming the responses allowed as 'correct' (Bull and McKenna 2001: 30).

Mathematical expressions may allow the use of random parameters to generate large numbers of 'different' questions (Bull and McKenna 2001: 30-1).

Graphical hotspot questions require identification of a particular location on the screen. This is useful in subjects in which the interpretation of visual materials is required (Bull and McKenna 2001: 31)

A more advanced question type is the *'multiple true/false' (MTFQ)* in which students are presented with a set of data followed by three or more statements. The student is required to determine whether each statement is true or false. Bull and McKenna (2001: 32) say a series of these questions on a specific topic can test a more complex understanding of an issue. N.B. Whereas a single true/false question carries a 50% chance that a student will guess the correct answer, a five-part multiple true/false question carries only a 3% chance of arriving at the correct answer solely by guessing. These questions are better than single true/false questions in discriminating between students of different abilities. Also discussed briefly by Bender (2003).

Bender (2003) and Dennis et al. (n.d.) also consider *Extended Matching Set Questions (EMSQ)*. Here, a list of possible answers (options) are supplied with an associated list of questions. Both lists may be quite long and the same option may be used once, more than once, or not at all. The claimed advantages of EMSQs include less 'cueing' of the correct answer than with other question types such as MCQs and fewer problems with writing plausible distracters. In comparison with MCQs and MTFQs, Bender (2003: 2) says:

> *"EMSQs are better for assessing the application of … knowledge, but on their own are unlikely to distinguish between the student who has not learnt the basic facts and the student who cannot apply the knowledge."*

However, he concedes that:

> *"As with much in teaching, learning and assessment, there is a great deal of fashion and opinion, but little hard evidence for either side [EMSQs vs. MCQs or MTFQs]."*

Assertion/reason items combine elements of multiple choice and true/false questions and, according to Bull and McKenna (2001: 33), allow testing of more complicated issues and require a higher level of learning. Questions consist of two statements, an assertion and a reason. First, the respondent has to determine whether each statement is true. If both are true, s/he must then determine whether the reason correctly explains the assertion. Assertion/reason questions are said to be suitable for exploring cause and effect and identifying relationships.

Appendix 3: Providing feedback

Harvey and Mogey (1999) recommend providing:

- different levels of feedback and help, dependent on the number of times students have given the wrong response;
- feedback which encourages students to continue, e.g. by giving advice on where they might have gone wrong.

Race (2007: 162):

- Create a separate response for students choosing each of the options.
- Try out draft self-assessment questions and feedback with 'live' students, e.g. by using them as class-based exercises in lectures or tutorials. This can lead to refinements or discarding those where it may not be straightforward to devise self-sufficient feedback responses.

Race (2007: 167-8): Feedback should:

- really *respond* to what students have done (give more than the correct answer, say what was wrong with their answer);
- remind students of what exactly they did (their answer should remain in sight)
- give appropriate praise without being patronising;
- not make those who got things wrong feel like complete idiots (e.g. if a question was difficult, acknowledge this).

After reviewing a selection of the CAA literature available in 2002, Rotheram wrote the following in the briefing for question-setters in his Social Policy Question Resource (SPQR) project (Rotheram, 2005):

> *Your job as a question-setter is to go beyond supplying appropriate questions and the correct answer. The project asks you also to provide helpful feedback for all option choices within a question. As a **minimum**, you should include:*
>
> - *a statement as to whether the chosen option is correct or not;*
> - *(if appropriate) a statement which gives the correct answer and a sentence or so on why it is correct;*
> - *a few words to clear up possible misunderstanding about the wrong choice;*
> - *the source of the information on the correct answer (to help the student find out more and, importantly, to give staff a starting point if they wish to amend a question, e.g. by supplying more up-to-date data).*
>
> *Please remember to keep the language clear and not discouraging.*
>
> *You might want to go beyond the minimum feedback and also include, say:*
>
> - *a related question for students to think about;*
> - *a web link to further information. If you do this, choose a link that is likely to have a reasonable 'shelf life' (i.e. it isn't likely to disappear quickly). Sometimes it may be better to link to the home page of a website (e.g. HM Prison Service), rather than the precise page (which might be short-lived) containing a particular statistic.*

Both Race (2007) and Rotheram (2007) recommend providing some feedback via sound files, because of the richness contained in the speaker's tone of voice and the speed with which a substantial amount of feedback can be provided. Rotheram (2007) goes further, suggesting that, if possible, sound files should be recorded directly in the highly-portable, widely-playable MP3 format.

Bob Rotheram
National Teaching Fellow
Reader in Assessment, Learning and Teaching
Office of Pro-Vice-Chancellor (ALT)
Leeds Metropolitan University

E: b.rotheram@leedsmet.ac.uk
M: 07824 482506
T: 0113 812 9045