ALT-J, Research in Learning Technology
Vol. 13, No. 1, March 2005, pp. 17–31

A review of computer-assisted assessment

Gráinne Conole* and Bill Warburton
University of Southampton, UK

Pressure for better measurement of stated learning outcomes has resulted in a demand for more frequent assessment. The resources available are seen to be static or dwindling, but Information and Communications Technology is seen to increase productivity by automating assessment tasks. This paper reviews computer-assisted assessment (CAA) and suggests future developments. A search was conducted of CAA-related literature from the past decade to trace the development of CAA from the beginnings of its large-scale use in higher education. Lack of resources, individual inertia and risk propensity are key barriers for individual academics, while proper resourcing and cultural factors outweigh technical barriers at the institutional level.

Introduction

Assessment is a critical catalyst for student learning (for example, Brown *et al.*, 1997) and there is considerable pressure on higher education institutions to measure learning outcomes more formally (Farrer, 2002; Laurillard, 2002). This has been interpreted as a demand for more frequent assessment. The potential for Information and Communications Technology (ICT) to automate aspects of learning and teaching is widely acknowledged, although promised productivity benefits have been slow to appear (Conole, 2004a; Conole & Dyke, 2004). Computer-assisted assessment (CAA) has considerable potential both to ease assessment load and provide innovative and powerful modes of assessment (Brown *et al.*, 1997; Bull & McKenna, 2004), and as the use of ICT increases there may be 'inherent difficulties in teaching and learning online and assessing on paper' (Bull, 2001; Bennett, 2002a).

This review of CAA sits within a context of increased ICT use, student diversity, financial constraints and the shift from quality assurance to quality enhancement. The review presents a fresh look at the topic by describing key features of CAA,

* Corresponding author. School of Education, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. Email: g.c.conole@soton.ac.uk

ISSN 0968-7769 (print)/ISSN 1741-1629 (online)/05/010017–15

DOI: 10.1080/0968776042000339772

identifying issues and highlighting potential areas for research. It describes progress made in identifying and addressing critical factors associated with implementing CAA.

CAA in context

The shift towards online testing is well documented (for example, Bennett, 2002a) and different forms of CAA are illustrated in Figure 1. Bull and McKenna recently defined CAA as 'the use of computers for assessing student learning' (2004). Computer-based assessment involves a computer program marking answers that were entered directly into a computer, whereas optical mark reading uses a computer to mark scripts originally composed on paper. Portfolio collection is the use of a computer to collect scripts or written work. Computer-based assessment can be subdivided into stand-alone applications that only require a single computer, applications that work on private networks and those that are designed to be delivered across public networks such as the web (online assessment).

Six ways in which the strategic application of a learning technology such as CAA may add value to the efficiency and effectiveness of the learning process have been identified, along with six factors that may adversely influence it (ALT, 2003). The issues around CAA are similar to those identified for other learning technologies in terms of design and delivery and associated support needs (Seale, 2003). CAA has obvious similarities with the development of Managed Learning Environments in terms of the encountered difficulty of institutional implementation and wide-scale

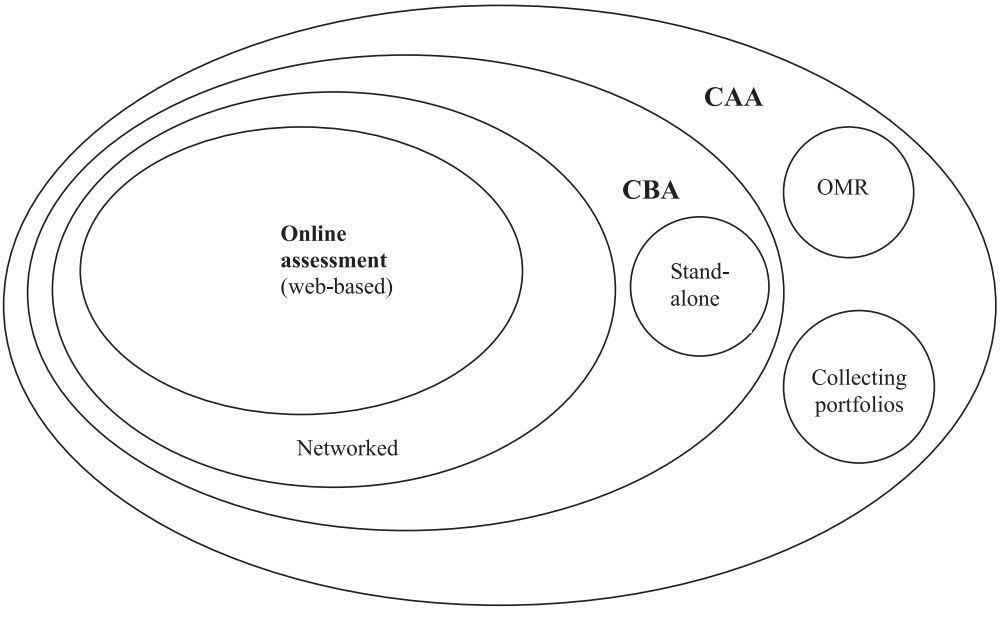

Figure 1. Different types of CAA

use (Sommerlad *et al.*, 1999). However, CAA differs from other learning technologies in that the stakes are much higher, particularly where it is used for examinations (QAA, 1998).

Students are perceived as increasingly litigious (Baty, 2004), and the clear scoring schemes of objective tests open the results of CAA tests to scrutiny that can render deficits in practice, apparently highlighting the need for risk analysis and management strategies (Zakrzewski & Steven, 2000).

Methodology

A search was conducted of CAA-related literature from the past decade to trace its development from the beginnings of large-scale use. Criteria for inclusion were direct or indirect reference to the implementation or evaluation of large-scale CAA. Some earlier material was also included because it forms the foundation of the literature. Search keywords were fed into electronic literature indexes, catalogues and search engines. The review is presented as a narrative tracing the progress made in identifying critical factors associated with implementing CAA and in overcoming operational and cultural barriers. The paper considers the findings from the literature in terms of CAA developments, exploring research activities around the design, delivery and analysis of online assessments. It identifies barriers and enablers to the uptake of CAA and considers emergent patterns of uptake as well as the associated implementation issues.

Categorising assessment

A multiple choice item consists of four elements: the stem of the question, options, correct responses and distractors (Figure 2). Tests are collections of subject-specific items, possibly drawn from item banks (Sclater, 2004). There are a variety of different question types (e.g. multiple choice, multiple response, hotspot, matching, ranking, drag and drop, multiple steps and open ended) and feedback mechanisms (including automatic feedback in objective testing, model answers, annotated test, or mixed mode with intervention from the teacher).

Assessment can be categorised as either summative (administered for grading purposes) or formative (to give feedback to assist the learning process). Diagnostic assessment is used by tutors to determine students' prior knowledge, and self-assessment is where students reflect on their understanding (O'Reilly & Morgan, 1999; Bull & McKenna, 2004). Other categorisations include formal/informal (invigilated or not) and final/continuous (at the end of a course or throughout). Sclater and Howie distinguish six different applications of CAA: 'credit bearing' or high-stakes summative tests, continuous assessment, authenticated or anonymous self-assessment, and diagnostic tests that evaluate the student's knowledge before the course to assess the effectiveness of the teaching (Sclater & Howie, 2003).

Six kinds of cognitive learning outcomes are distinguishable according to Bloom *et al.*'s (1956) taxonomy: knowledge recall is at the most fundamental level, rising through comprehension, application, analysis, synthesis and evaluation. Others have

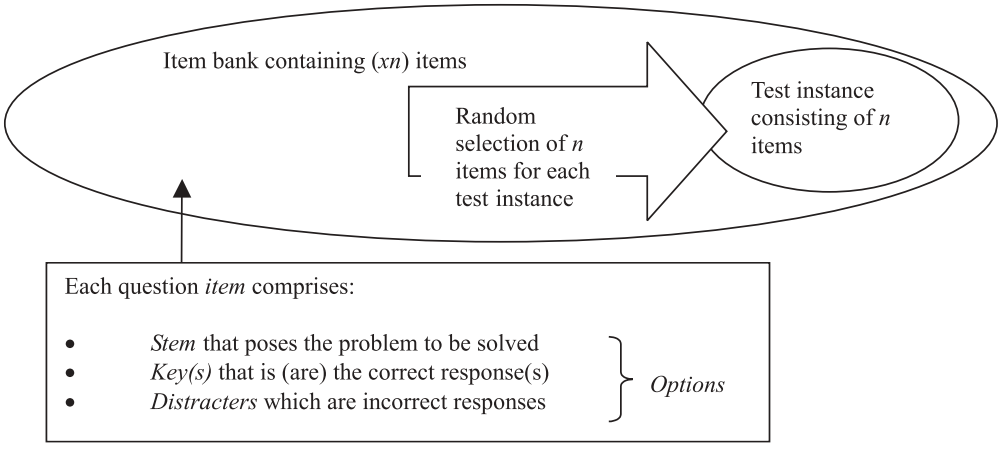

Figure 2. Test selection from an item bank

extended or adapted this (Krathwohl *et al.*, 1964; Imrie, 1995; Anderson *et al.*, 2001). Although tutors in higher education assess the full range of outcomes (Miller *et al.*, 1998, p. 56), it is the view of many that higher education should specialise in developing the highest level skills; for example, evaluative ('critical') skills (Barnett, 1997).

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 informed responses and analyses is permissible) (McAlpine, 2002a). Convergent assessments can be readily constructed using objective items, but divergent assessment has traditionally relied on longer written answers or essays, the automated marking of which presents real challenges (for example, Mason & Grove-Stephensen, 2002).

Current research and development activities

Bull and McKenna (2004) provide a valuable overview of CAA, McAlpine (2002a) outlines the basic principles of assessment and Mills *et al.* (2002) describe some of the current research directions.

CAA research and development activities can be grouped into three areas. The first centres on design, ranging from the development of individual items through to the specification of full CAA systems. The second is concerned with implementation and delivery. The third focuses on analysis, test scoring and the development of appropriate reporting systems.

Design

Item development

CAA allows for more complex item types compared with paper-based assessments, including the use of audiovisual materials and more complex interactions between

learner and computer. One finding from the literature is that direct translation of paper-based assessments into online assessments is inappropriate; there is a need to revisit question formulation, reflecting on what it is intended to test. The process of creating CAA questions therefore raises fundamental issues about the nature of paper-based questions as well. The use of steps for some kinds of questions (for example in mathematics) has also proved valuable in terms of enabling teachers and researchers to get a better understanding of the student learning experience and how they tackle questions (Ashton *et al.*, 2003). This raises important issues about how CAA software systems record and report on student interactions. The reporting mechanisms available within CAA systems provide richer data about the students than were available from paper-based assessments. Thus the development of online assessments has important pedagogical implications.

The format of an assessment affects validity, reliability and student performance. Paper and online assessments may differ in several respects. Studies have compared paper-based assessments with computer-based assessments to explore this (for example, Ward *et al.*, 1980; Outtz, 1998; Fiddes *et al.*, 2002). In particular, the Pass-IT project has conducted a large-scale study of schools and colleges in Scotland, across a range of subject areas and levels (Ashton *et al.*, 2003, 2004). Findings vary according to the item type, subject area and level. Potential causes of mode effect include the attributes of the examinees, the nature of the items, item ordering, local item dependency and the test-taking experience of the student. Additionally there may be cognitive differences and different test-taking strategies adopted for each mode. Understanding these issues is important for developing strategies for item development as well as to produce guidelines for developing appropriate administrative procedures or statistically adjusting item parameters.

In contrast to marking essays, marking objective test scripts is a simple repetitive task and researchers are exploring methods of automating assessment. Objective testing is now well established in the United States and elsewhere for standardised testing in schools, colleges, professional entrance examinations and for psychological testing (Bennett, 2002b; Hambrick, 2002).

The limitations of item types are an ongoing issue. A major concern related to the nature of objective tests is whether multiple choice questions (MCQs) are really suitable for assessing higher-order learning outcomes in higher education students (Pritchett, 1999; Davies, 2002), and this is reflected in the opinions of both academics and quality assurance staff (Bull, 1999; Warburton & Conole, 2003). 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 (Farthing & McPhee, 1999; Duke-Williams & King, 2001). MCQs and multiple response questions are still the most frequently used question types (Boyle *et al.*, 2002; Warburton & Conole, 2003) but there is steady pressure for the use of 'more sophisticated' question types (Davies, 2001).

Work is also being conducted on the development of computer-generated items (Mills *et al.*, 2002). This includes the development of item templates precise enough to enable the computer to generate parallel items that do not need to be individually

calibrated. Research suggests that some subject areas are easier to replicate than others—lower-level mathematics, for example, in comparison with higher-level content domain areas.

Item banks

Item banks are collections of questions, often produced collaboratively across a subject domain that can be grouped according to difficulty, the type of skill or topic. Sclater and colleagues have recently conducted a detailed review of item bank developments within the United Kingdom (Sclater, 2004), covering metadata, security, access and authentication and legal issues. Sclater positions item banks as the crucial driver of CAA, and McAlpine (2002b) argues for the routine adoption of item banks as a means of countering challenges from students about fairness, validity, security or quality assurance. Other researchers have looked at setting up, maintaining and adapting item banks (Mills *et al.*, 2002). Security issues are particularly important with high-stakes assessments; use of larger pools is one strategy that can help deter cheating. Controlling item exposure and maintaining the security of item banks is likely to continue to be an active area of research and development.

Computer-adaptive testing

Objective items' ability to explore the limits of a participant's ability is developed by computer-adaptive testing (CAT). CAT involves issuing questions of a difficulty level that depends on the test-taker's previous responses. If a question is answered correctly, the estimate of his/her ability is raised and a more difficult question is presented, and *vice versa*, giving the potential to test a wide range of student ability concisely. For instance, Lilley and Barker (2003) constructed a database of 119 peer-reviewed items and gave both 'traditional' (non-adaptive) and CAT tests to 133 students drawing on Item Response Theory as a model. Students' results from the CAT test correlated well with their results from the traditional version and they did not find the CAT test unfair. Because CAT items are written to test particular levels of ability, they have the potential to deliver more accurate and reliable results than traditional tests.

CAA software systems

CAA software tools vary in how they support the design, delivery and analysis of online assessments and differ in cost, flexibility and scalability. The TOIA system provides a range of tools for managing and reporting on the assessment process and has produced a free Question and Test Interoperability (QTI)-compliant system (TOIA, 2004). TRIADS delivers different question styles in a variety of modes to facilitate the testing of higher order learning skills (McKenzie *et al.*, 2002). Commercial tools, like Perception (Question Mark Computing Ltd, 2004), represent considerable investments in terms of initial outlay and maintenance agreements. CAA systems vary

widely in the number of question types supported and in the control administrators have in scheduling assessments. For example, Perception supports 18 objective item types, TOIA nine (although the commercial version has more) and Blackboard six. Assessment tools in Virtual Learning Environments tend to be less sophisticated but have proved important in getting practitioners using CAA. More costly systems tend to be sold on the basis of commensurate scalability and flexibility with dedicated support; however, scalability is still problematic (Danson *et al.*, 2001; Cosemans *et al.*, 2002; Stevenson *et al.*, 2002; Harwood, 2004). Software also varies in the number of assessments that can be taken simultaneously (Question Mark Computing Ltd, 2004).

The earliest CAA systems were stand-alone applications, whereas current systems are either connected by a private network or are delivered by browser. Although most CAA assessments are still constructed from MCQs (Warburton & Conole, 2003) there have long been demands for more flexible question types, including those that might be difficult or impossible to rendered on paper (Bull & Hesketh, 2001; Davies, 2001; Bennett, 2002b). This is reflected in the increasing number of question types supported by CAA software, including 'new' question types. It is clearly in the interests of vendors to maximise the number of item types supported by their product. This is seen by users as an important metric of flexibility and a means by which vendors differentiate themselves from their competitors, although many are simply elaborations of basic question types, which makes it difficult to compare CAA products (Paterson, 2002).

Interoperability

Interoperability is important for transferring questions between systems. Practitioners are using different ICT tools to support their teaching, and hence may want to design questions within one tool and deliver tests in another. One contentious issue is whether current CAA systems are truly interoperable (Sclater *et al.*, 2002).

Lay and Sclater (2001) identify two more reasons for interoperability. First, whether the item banks will be accessible when current CAA systems are no longer in use, and second whether student assessment data can be transferred to institutional student records system. Another important driver for interoperability is to preserve users' investments in existing questions and tests when moving between institutions or to different CAA systems. The IMS (2003) Consortium's QTI specification is a valuable starting point, but clearly there is a need for further work (Sclater *et al.*, 2002; Sclater & Howie, 2003).

Delivery

Compliance with published standards for CAA practice

The recent code of practice for the use of information technology in the delivery of assessments (BS 7988: 2002) acknowledges that increased use of CAA 'has raised

issues about the security and fairness of IT-delivered assessments, as well as resulting in a wide range of different practices' (BSI, 2002, p. 11). It aims to enhance the status of CAA and encourage use by demonstrating its fairness, security, authenticity and validity. However, the code's focus on the delivery of CAA tests could lead to the relative neglect of earlier stages in the preparation and quality assurance of assessments. As Boyle and O'Hare (2003, p. 72) contend, 'A poor assessment, delivered appropriately, would [still] conform to BS'. They identified the American Educational Research Association (2003) *Standards for Educational and Psychological Testing*, the Association of Test Publishers *Guidelines for Computer-based Testing* and the Scottish Qualifications Authority (2003) *Guidelines for Online Assessment in Further Education* as better guides to practice.

Scaling up CAA

An obstacle to the uptake of CAA is that it is often implemented by individuals on an *ad hoc* basis with no overarching strategy or institutional IT infrastructure. This may delay or prevent embedding (Bull, 2001; Boyle & O'Hare, 2003). Bull asserts that 'Retooling is a challenge which impacts on research and development, requiring a high level of resourcing for academic and support staff in order to maintain pace with technological and software developments' (2001, p. 11). The risks of small-scale development include practitioner isolation and under-funding, although possible benefits include practitioners being in control of the process (Kennedy, 1998). Higher education institutions that are implementing CAA centrally encounter risks and benefits on a different scale (Danson *et al.*, 2001; Cosemans *et al.*, 2002; Warburton & Harwood, 2004). Scaling up for full-scale institutional deployment covers every possible use and seems likely to depend more upon the resolution of cultural than technical issues. Bull (2001, p. 11) points out that 'the organisational and pedagogical issues and challenges surrounding the take-up of CAA often outweigh the technical limitations of software and hardware.' This finding is mirrored in the issues associated with the uptake of other learning technologies (for example, Seale, 2004).

For individual practitioners, concerns about risks are likely to continue (Harwood & Warburton, 2004). While operational barriers might be overcome with incremental advances in technology, cultural obstacles are proving more durable.

Critical factors governing the uptake of CAA

Traditional assessment practices are now reasonably well understood. Even so, not all traditional assessments run smoothly (for example, Goddard, 2002). Many barriers and enablers for traditional assessment are also relevant to CAA. The emergence of CAA has forced the re-examination of these dormant issues in traditional practice. An argument for establishing good CAA practice at an institutional level is that it triggers the re-examination of assessment practice generally (Bull & McKenna, 2004).

Several approaches have been taken to identifying factors governing uptake. Stephens and Mascia conducted the first UK survey of CAA use in 1995 using a

10-item questionnaire that attracted 445 responses. They identified the need for institutional support (training and resourcing), allowing time to develop CAA tests, making CAA a fully integrated part of existing assessment procedures (rather than an afterthought) and subject-related dependencies. Important operational factors were familiarisation with the tools, well-planned procedures that addressed security and reliability issues, and involvement of support staff (Stephens & Mascia, 1997).

Four years later the CAA Centre conducted a national survey of higher education that focussed on use and attitudes. The survey built on that of Stephens and Mascia (Bull, 1999) but had over twice as many items, many of which were multi-part. It attracted more than 750 responses from academics, quality assurance staff and staff developers (McKenna, 2001). Warburton and Conole (2003) piloted an adapted, online version of the 1999 survey, which received 50 responses, mostly from academic CAA enthusiasts.

The greatest institutional barrier was seen to be cost both in terms of personal time and the expense of commercial software. Unrealistic expectations coupled with inherent conservatism, and lack of technical and pedagogic support were also cited. Respondents were less concerned with Managed Learning Environment integration, security or copyright issues. Another obstacle was the perceived steep learning curve associated with the technology and constructing specialised CAA question types. Of particular concern was the difficulty of constructing objective items that reliably assess higher-learning outcomes (cf. Boyle & O'Hare, 2003). There was a perceived credibility gap between what CAA proponents promise and what respondents thought could be delivered; lack of support, cultural resistance and technophobia were cited less often. Related issues about usability, academics working in isolation and individual inertia were also raised. Subject-specific shared question banks and the value of exemplars were cited as important drivers for the large-scale uptake of CAA, but the provision of 'evangelists' and adherence to institutional guidelines was thought less crucial.

Academic commitment was cited as an important enabler; faculty support for CAA seems limited; external funding is the principle way support for CAA at this level is rendered. Other important factors included the need to embed CAA within normal teaching. Effective interoperability (particularly between CAA systems and Virtual Learning Environments) and integration of multimedia were also cited. Most systems were web-based, although many respondents delivered CAA using closed networks, and a small percentage used optical mark reading. Only one-third were invigilated, and most of the summative CAA tests restricted the percentage weighting to one-third or less of the overall mark, although it was noteworthy that some tests were worth up to 100%. CAA was used to test to a range of group sizes including very large groups (more than 200 students). Subject-specific differences in the uptake of CAA were obvious (Bull, 1999; Warburton & Conole, 2003). Interestingly, Quality Assurance staff identified few enabling factors, perhaps indicating their largely negative perception of CAA (Bull, 1999; Bull & McKenna, 2000; Bull & Hesketh, 2001).

The centrality of cultural issues was evident, with 90% of barriers and 65% of enablers being identified as cultural in the 1999 survey. Hambrick's (2002) study of

large-scale applications of formal online assessment in the US K-12 school system are split equally between cultural and operational factors. Zakrzewski and Steven (2000) conducted a formal risk assessment; one-third of the factors they identified were cultural.

Analysis

Item analysis and scoring

One of the benefits of CAA is the opportunity to record student interactions and analyse these to provide a richer understanding of learning. A variety of analyses can be run to assess how well individual questions or students perform. Weak items can then be eliminated or teacher strategies adapted. Automatically recorded data can be used in a variety of ways; for example, looking at the relationship between answering speed and accuracy. Care is needed, however, in interpreting results, for incorrect responses may indicate a more sophisticated understanding by the student than might at first appear; for example, incorrect use of grammar in a foreign language test might be a result of higher cognitive understanding by the student. Gitomer *et al.* (1987) found both high-ability and low-ability examinees increased their processing time for more difficult items, but that low-ability examinees spent more time on encoding the stem while high-ability examinees spent more time on subsequent processing.

Assessments need to be both valid and reliable. An advantage of CAA is that it offers consistency in marking. A range of methods are possible for scoring from simple allocation of a mark to the correct response through to varied, compound and negative scoring. Two main methods are used for item statistics, Classical Test Theory and Latent Trait Analysis (LTA) (Rasch analysis and Item Response Theory). The former is simpler and evaluates at the level of a test, whereas the latter looks at individual questions. More details on these can be found elsewhere (McAlpine, 2002c; Mills *et al.*, 2002). Boyle *et al* (2002) explore the use of Classical Test Theory, Item Response Theory and Rasch analysis with a set of 25 questions used by 350 test-takers. They concluded that the present approach by many practitioners to CAA of neglecting the rigorous quality assurance of items is untenable, and that this is particularly problematic for high-stakes assessment. Boyle and O'Hare (2003) recommend that training in item construction and analysis should be obligatory for staff who are involved in developing CAA tests and that items should be peer-reviewed and trialled before use. Statistics can be used to aid curriculum design and quality control (Bull & McKenna, 2004). Feedback of this type can be motivating in terms of enabling a student to identify areas of weakness and evaluate their performance against stated learning outcomes.

Concerns about the risk of students guessing answers are addressed in two main ways: first by discounting a test's guess factor, and second by adjusting the marking scheme away from simple tariffs where 'one correct answer equals one mark' to include negative marking. Confidence-based assessment is where marks awarded for a response are predicted on a student's confidence that the correct response has been given (Davies, 2002; Gardner-Medwin & Gahan, 2003; McAlpine & Hesketh, 2003).

Concerns about the risk of cheating in summative tests might be reduced by strategies such as providing 'blinker screens' and proper invigilation, by interleaving participants taking different tests and by randomising item and response order (BSI, 2002; Bull & McKenna, 2004; Pain & LeHeron, 2003). Other tactics include the use of item banks.

Future directions

Research is still needed into understanding the differences between paper-based and online assessments across different subject areas and levels. Additionally we need to explore how this knowledge can be used to produce more effective questions and appropriate administration of tests. Greater understanding is also needed into effective question design and how this links to learning outcomes. It seems likely that CAA will continue to require specialised skills in particular for the construction of good items. Concerns about cost-benefit may be partially addressed by the integration of other technologies such as multimedia as indicated by the graphical, animated (FLASH) and open-ended JAVA item types provided by more recent CAA systems (Question Mark Computing Ltd, 2004; TOIA, 2004).

Automated essay marking may address concerns about the difficulty of assessing higher-level outcomes. However, this is not currently part of many CAA systems and requires significant resources in terms of skills and time (Christie, 2003).

In terms of system design, the pressure for greater interoperability may force the development of standards for test interoperability beyond items, and current *de facto* standards such as Questionmark Markup Language are increasingly being supplemented by the development of 'open' standards such as the IMS Consortium's QTI specification (IMS, 2003). Sclater and Howie's (2003) 'ideal' CAA system is meant to fulfil institutional needs that current 'off the shelf' commercial CAA system arguably do not. They identify 21 possible roles that a full-scale CAA system might require: six functional roles (authors, viewers and validators of questions, test authors, viewers and validators); three associated with the sale of items or assessments; three administrative roles; six associated with processing results; and three to do with delivery (test-taker, timetabler and invigilator).

Conclusion

This paper has provided a review of current activities in the design, delivery and analysis of online assessment. Many of the barriers and enablers to effective implementation mirror those found in the uptake and use of other learning technologies; however, CAA differs because it is perceived as more high risk and central to the learning and teaching process. Much of the research raises fundamental questions about the whole learning and teaching process; in particular, raising complex pedagogical questions about the role of assessment and its relationship to learning.

The role of technology and how it might impact on assessment is still in its infancy and we need to develop new models for exploring this. Similarly we need to understand

more about the barriers and enablers to using these tools effectively and mechanisms for addressing these. In light of this there are real issues in terms of what we might expect or want students to learn. In addition there is a general shift across education from assessment of products or outputs to assessing the processes of learning. Therefore we need to consider what we want to be assessing and how best to do this.

Similarly, issues arise given the new forms of communication and collaborations that are now possible. How can we measure the interactions that occur within an online discussion forum and how can we attribute this in terms of student learning? What about those who do not contribute—'the lurkers'—are they opting out or learning differently (e.g. vicariously by reading and reflection on the postings of others)? Conole (2004b) lists a number of questions CAA researcher need to address, including exploration of which new forms of assessment might arise as a result of the impact of technologies. CAA raises fundamental questions about the whole learning and teaching process that we need to research and address if we can achieve the desired goal of maximising the potential technologies offer to improve learning, teaching and assessment.

References

AERA (1999) *Standards for educational and psychological testing* (Washington, DC, AERA).

ALT (2003) The future of higher education: the Association for Learning Technologies response to the white paper. Available online at: http://www.alt.ac.uk/docs/he_wp_20030429_final.doc (accessed 19 December 2003).

ATP (2003) *Guidelines for computer-based testing* (Washington, DC, ATP).

Anderson, L., Krathwohl, D. & Bloom, B. (2001) *A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives* (New York, Longman).

Ashton, H. S., Schofield, D. K. & Woodgar, S. C. (2003) Piloting summative web assessment in secondary education, in: J. Christie (Ed.) *Proceedings of the 7th International Computer–Assisted Assessment Conference,* Loughborough University, 8–9 July 2003.

Ashton, H. S., Beavers, C. E., Schofield, D. K. & Youngson, M. A. (2004) Informative reports—experiences from the Pass-IT project, in: M. Ashby and R. Wilson (Eds) *Proceedings of the 8th International Computer–Assisted Assessment Conference,* Loughborough University, 6–7 July 2004.

Barnett, R. (1997) *Higher education: a critical business* (Buckingham, Society for Research into Higher Education and the Open University Press).

Baty, P. (2004, March 12) Litigation fees top £15m as academic disputes grow, *Times Higher Education Supplement,* 2004, p. 7.

Bennett, R. E. (2002a) Inexorable and inevitable: the continuing story of technology and assessment. *Journal of Technology, Learning and Assessment,* 1. Available online at: http://www.bc.edu/research/intasc/jtla/journal/pdf/v1n1_jtla.pdf (accessed 17 July 2004).

Bennett, R. E. (2002b) *Using electronic assessment to measure student performance.* The State Education Standard, National Association of State Boards of Education.

Bloom, B., Engelhart, M., Furst, E., Hill, W. & Krathwohl, D. (1956) *Taxonomy of educational objectives: the classification of educational goals. Handbook 1: cognitive domain* (London, Longman).

Boyle, A., Hutchison, D., O'Hare, D. & Patterson, A. (2002) Item selection and application in higher education, in: M. Danson (Ed.), *6th International CAA Conference* (Loughborough, Loughborough University).

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

Brown, G., Bull, J. & Pendlebury, M. (1997) *Assessing student learning in higher education* (Routledge, London).

BSI (2002) *Code of practice for the use of information technology (IT) in the delivery of assessments* (London, BSI).

Bull, J. (1999) Update on the National TLTP3 Project The Implementation and Evaluation of Computer-assisted Assessment, in: M. Danson (Ed.), *3rd International CAA Conference,* Loughborough University, 16–17 June 1999.

Bull, J. (2001) *TLTP85 implementation and evaluation of computer-assisted assessment: final report.* Available online at: www.caacentre.ac.uk/dldocs/final_report.pdf (accessed16 March 2003).

Bull, J. & Hesketh, I. (2001) Computer-assisted assessment centre update, in M. Danson & C. Eabry (Eds) *5th International CAA Conference,* Loughborough University, 2–3 July 2001.

Bull, J. & Mckenna, C. (2000) Computer-assisted assessment centre (TLTP3) update, in: M. Danson (Ed.) *4th International CAA Conference,* Loughborough University, 21–22 June 2000.

Bull, J. & McKenna, C. (2004) *Blueprint for computer-assisted assessment* (London, RoutledgeFalmer)

Christie, J. (2003) Automated essay marking for content—does it work?, in J. Christie. (Ed.) *7th International CAA Conference,* Loughborough, 8–9 July 2001.

Conole, G. (2004a) *Report on the effectiveness of tools for e-learning,* report for the JISC commissioned Research Study on the Effectiveness of Resources, Tools and Support Services used by Practitioners in Designing and Delivering E-Learning Activities.

Conole, G. (2004b) *Assessment as a catalyst for innovation,* invited keynote and paper for the Quality Assurance Agency for Higher Education, Assessment Workshop 5. Available online at: http://www.qaa.ac.uk/scottishenhancement/events/default.htm (accessed 10 September 2004).

Conole, G. & Dyke, M. (2004) What are the affordances of Information and communication technologies?, *ALT-J,* 12(2), 111–123.

Cosemans, D., Van Rentergem, L., Verburgh, A. & Wils, A. (2002) Campus wide setup of question mark perception (V2.5) at the Katholieke Universiteit Leuven (Belgium)—facing a large scale implementation, in: M. Danson (Ed.) *5th International CAA Conference,* University of Loughborough, 2–3 July 2001.

Danson, M. (2003) *Implementing online assessment in an emerging MLE: a generic guidance document with practical examples* (Bristol, JISC).

Danson, M., Dawson, B. & Baseley, T. (2001) Large scale implementation of question mark perception (V2.5)—experiences at Loughborough University, in: M. Danson (Ed.) *6th International CAA Conference,* Loughborough University, 2–3 July 2001.

Davies, P. (2001) CAA must be more than multiple-choice tests for it to be academically credible?, in: M. Danson & C. Eabry (Eds) *5th International CAA Conference,* Loughborough University.

Davies, P. (2002) There's no confidence in multiple-choice testing, in: M. Danson (Ed.) *6th International CAA Conference,* Loughborough University, 4–5 July 2002.

Duke-Williams, E. & King, T. (2001) Using computer-aided assessment to test higher level learning outcomes, in: M. Danson & C. Eabry (Eds) *5th International CAA Conference,* Loughborough University, 2–3 July 2001.

Farrer, S. (2002) End short contract outrage, MPs insist, *Times Higher Education Supplement.*

Farthing, D. & McPhee, D. (1999) Multiple choice for honours-level students?, in: M. Danson (Ed.) *3rd International CAA Conference,* Loughborough University, 16–17 June 1999.

Fiddes, D. J., Korabinski, A. A., McGuire, G. R., Youngson, M. A. & McMillan, D. (2002) Are mathematics exam results affected by the mode of delivery?, *ALT-J,* 10(6), 1–9.

Gardner-Medwin, A. & Gahan, M. (2003) Formative and summative confidence-based assessment, in: J. Christie (Ed.) *7th International CAA Conference,* Loughborough University, 8–9 July 2003.

Gitomer, D. H., Curtis, M. E., Glaser, R. & Lensky, D. B. (1987) Processing differences as a function of item difficulty in verbal analogy performance, *Journal of Educational Psychology,* 79, 212–219.

Goddard, A. (2002, November 1) Universities blamed for exam fiasco, *Times Higher Education Supplement*, p. 7.

Hambrick, K. (2002) *Critical issues in online, large-scale assessment: An exploratory study to identify and refine issues.* Unpublished Ph.D. thesis, Capella University.

Harwood, I. (2004) What happens when computer-assisted assessment goes wrong?, *British Journal of Educational Technology*, submitted.

Harwood, I. & Warburton, W. (2004) Thinking the unthinkable: using project risk management when introducing computer-assisted assessments, in: M. Ashby & R. Wilson (Eds) *8th International CAA Conference*, Loughborough University, 6–7 July 2004.

Imrie, B. (1995) Assessment for learning: quality and taxonomies, *Assessment and Evaluation in Higher Education*, 20, 171–189.

IMS (2003) *QTI Enterprise v1.1 specs, examples and schemas.* Available online at: http://www.imsglobal.org/specificationdownload.cfm (accessed 2 March 2003).

Kennedy, N. (1998) *Experiences of assessing LMU students over the web.* Leeds Metropolitan University.

Krathwohl, D., Bloom, B. & Masia, B. (1964) *Taxonomy of educational objectives the classification of educational goals* (London, Longman).

Laurillard, D. (2002,) *Rethinking university teaching a conversational framework for the effective use of learning technologies* (2nd edn) (London, RoutledgeFalmer).

Lay, S. & Sclater, N. (2001) Question and test interoperability: an update on national and international developments, in: M. Danson & C. Eabry (Eds) *5th International CAA Conference*, Loughborough University, 2–3 July 2001.

Lilley, M. & Barker, T. (2003) An evaluation of a computer adaptive test in a UK university context, in: J. Christie (Ed.) *7th International CAA Conference*, Loughborough University, 8–9 July 2003.

Mason, O. & Grove-Stephensen, I. (2002) Automated free text marking with paperless school, in: M. Danson (Ed.) *6th International CAA Conference*, Loughborough University, 2–3 July 2001.

McAlpine, M. (2002a) *Principles of assessment* (Luton, CAA Centre).

McAlpine, M. (2002b) *Design requirements of a databank* (Luton, CAA Centre).

McAlpine, M. (2002c) *A summary of methods of item analysis* (Luton, CAA Centre).

McAlpine, M. & Hesketh, I. (2003) Multiple response questions—allowing for chance in authentic assessments, in: J. Christie (Ed.) *7th International CAA Conference*, Loughborough University, 8–9 July 2000.

McKenna, C. (2001) Academic approaches and attitudes towards CAA: a qualitative study, in: M. Danson & C. Eaborg (Eds) *5th International CAA Conference*, Loughborough University, 2–3 July 2001.

McKenzie, D., Hallam, B., Baggott, G. & Potts, J. (2002) TRIADS experiences and Developments, in: M. Danson (Ed.) *6th International CAA Conference*, Loughborough University, 2–3 July 2002.

Miller, A., Imrie, B. & Cox, K. (1998) *Student assessment in higher education: a handbook for assessing performance* (London, Kogan Page).

Mills, C., Potenza, M., Fremer, J. & Ward, C. (Eds) (2002) *Computer-based testing—building the foundation for future assessment* (New York, Lawrence Erlbaum Associates).

Outtz, J. L. (1998) Testing medium, validity and test performance, in: M. D. Hakel (Ed.) *Beyond multiple choice evaluating alternative to traditional testing for selection* (New York, Lawrence Erlbaum Associates).

O'Reilly, M. & Morgan, C. (1999) Online assessment: creating communities and opportunities, in: S. Brown, P. Race & J. Bull (Eds) *Computer-assisted assessment in higher education* (London, Kogan Page), 149–161.

Pain, D. & Leheron, K. (2003) WebCT and online assessment: the best thing since SOAP?, *Educational Technology and Society*, 6, 62–71.

Paterson, J. S. (2002) What's in a name? A new hierarchy for question types, in: M. Danson (Ed.) *6th International CAA Conference*, Loughborough University, 4–5 July 2002.

Pritchett, N. (1999) Effective question design, in: S. Brown, P. Race & J. Bull (Eds) *Computer-assisted assessment in higher education* (London, Kogan Page).

QAA. (1998) *University of Bath quality audit report.* Available online at: http://www.qaa.ac.uk/revreps/instrev/bath/comms.htm (accessed 4 September 2003).

Question Mark Computing Ltd (2004) *Perception: Windows based authoring,* Available online at: http://www.questionmark.com/uk/perception/authoring_windows.htm (accessed 29 August 2004).

Sclater, N., Low, B. & Barr, N. (2002) Interoperability with CAA: does it work in practice?, in: M. Danson (Ed.) *Proceedings of the 6th International Conference on Computer-Assisted Assessment,* Loughborough University, 4–5 July 2002.

Sclater, N. & Howie, K. (2003) User requirements of the ultimate online assessment engine, *Computers and Education,* 40, 285–306.

Sclater, N. (Ed.) (2004) *Final report for the Item Banks Infrastructure Study (IBIS)* (Bristol, JISC).

Seale, J. (Ed.) (2003) *Learning technology in transition—from individual enthusiasm to institutional implementation* (Swets and Zeitlinger, The Netherlands).

Sommerlad, E., Pettigrew, M., Ramsden, C. & Stern, E. (1999) *Synthesis of TLTP annual reports* (London, Tavistock Institute).

Stephens, D. & Mascia, J. (1997) *Results of a (1995) Survey into the use of computer-assisted assessment in institutions of higher education in the UK,* University of Loughborough.

Stevenson, A., Sweeney, P., Greenan, K. & Alexander, S. (2002) Integrating CAA within the University of Ulster, in: M. Danson (Ed.) *Proceedings of the 6th International conference on computer-assisted assessment,* University of Loughborough, 4–5 July 2002.

Scottish Qualifications Authority (2003) *Guidelines for online assessment in further education* (Glasgow, Scottish Qualifications Agency).

TOIA (2004) *The TOIA Project web site.* Available online at: http://www.toia.ac.uk (accessed 31 August 2004).

Warburton, W. & Conole, G. (2003) CAA in UK HEIs: the state of the art, in: J. Christie (Ed.) *7th International CAA Conference,* University of Loughborough, 8–9 July 2003.

Warburton, W. & Harwood, I. (2004) Implementing perception across a large university: getting it right, *Perception European Users Conference* (Edinburgh, Question Mark Computing).

Ward, W. C., Frederiksen, N. & Carlson, S. B. (1980) Construct validity of free response and machine-scorable forms of a test, *Journal of Educational Measurement,* 7(1), 11–29.

Zakrzewski, S. & Steven, C. (2000) A model for computer-based assessment: the Catherine wheel principle, *Assessment and Evaluation in Higher Education,* 25, 201–215.