History and Challenges of e-assessment

The 'Cambridge Approach' perspective - e-assessment research and development 1989 to 2009

Patrick Craven
[1]Cambridge Assessment

Key words: *e-learning, e-assessment, challenges, CBT, e-portfolio, knowledge, skills, competence*

Abstract:

A review of the history of e-assessment research and development within the Cambridge Assessment Group charting a 'fit for purpose' evidence-based approach to the introduction of valid and reliable e-assessment solutions. The paper will draw on examples from OCR Examinations vocational and general assessment applications in the UK, Cambridge ESOL tests for English language and Cambridge International Examinations solutions for diverse customers and locations around the world.

Despite continuous improvements to technology applications the mainstream education industry has remained largely immune to its use within high stakes national assessment systems. The paper concludes with some observations about why this might be and why they may continue to present significant challenges to the penetration of e-assessment solutions within compulsory education.

1 The Cambridge Approach

The foundation of effective and trustworthy assessment is defined by the 'Cambridge Approach' to assessment development. These principles have been applied to the organisation's e-assessment solutions since the first attempts to create Word Processing Functions Tests through to the more recent large scale deployments of on screen marking and computer-based tests for 14 to 19 year olds in support of school exams. For an assessment to be 'fit for purpose' it is vital to consider the relationship between assessment type and the complexity of skill/understanding you are trying to measure, and recognise the extent to which purpose must always drive assessment form if the outcomes are to remain valid and reliable [1]. Without robust validity and reliability an assessment will have little value and if stakeholder confidence is lost in the outcomes its associated currency will also be eroded.

Such principles create justifiable constraints on the wide scale adoption of simple forms of assessment such as multiple-choice questions. This is not to say that multiple-choice questions are not appropriate for some aspects of assessment, when used correctly they are an effective and efficient form of assessment. However, if we are to retain the integrity of assessment practice it is important for assessment design to be based on the best fit with the construct, concept or skills that we wish to assess [2, 3]. There is a danger that e-assessment will lead some to adopt quick and convenient modes of assessment that are not fit for purpose and this in turn will impact the credibility of the measurement. The Cambridge Approach seeks to ensure that technology is used to enhance the process; reliability, validity, accessibility, efficiency, feedback, speed etc but does not dilute the quality.

There are also sound ethical reasons why the mode of delivery, screen versus paper, should be studied prior to any widespread launch of an e-assessment solution [4]. Without such diligence there is always the risk that unintended consequences may arise from what appear to be simple deviations from practice [5, 6]. More specific studies into mathematics [7, 8], writing [9] and comprehension [10] have shown that there are differences and although not significant we should remain vigilant about the implications of impact on performance. Assessment can have a profound impact on the future life chances of candidates and so assessment agencies and educators have a duty to ensure that research and development is conducted in a professional and rigorous manner [11].

It is also important to consider the implications of marking digital evidence or evidence presented on screen. Not all e-assessment involves delivery of an assessment on-screen, or automated marking, and so it is not always just the candidate that is affected by change [12, 13, 14]. A professional approach to evidence-based research and development must cover all aspects of assessment and consider all the people, students and examiners, who come into contact with the system.

2 History of e-assessment research & development

2.1 Experience and diversity

Over twenty years experience of research and development in this domain has allowed the Cambridge Assessment Group to pilot e-assessment solutions ranging from the computer-mediated to more complex applications of true computer-based assessment. These applications cover subject areas as diverse as Maths, Science, Languages, Geography and Vocational Sector competence. Assessment techniques range from question papers, tasks and projects to portfolios of performance-based evidence – e-assessment studies have been explored in every area and some with more success than others. This section considers the range of applications we have studied.

Across the spectrum of subject and vocational sectors supported by the Cambridge

Assessment Group there are examples of many types of assessment. To assess knowledge and understanding a range of testing and examination techniques are typically used. These range from highly objective questions types, such as multiple choice/response or true/false, to the more subjective disciplines that require open-ended constructed response items or creative/technical essays. To assess skills or competence it is more usual to set tasks and assess through performance or observation. In either domain there will be a mixture of assessment that is entirely independent and external or internal assessments (local judgements), which are then externally moderated and validated.

True computer-based forms of assessment, where the technology conducts the marking, are best associated with highly logical assessment forms. Objective assessment, sometimes referred to as convergent assessment as the markschemes focus on one right answer, can be automated in a very effective way. Subjective assessment, sometimes referred to as divergent assessment as the markschemes will contain a variety of acceptable answers, does not lend itself as easily to total automation but still presents many possibilities for e-assessment. In the latter context e-assessment solutions use technology to facilitate the process of assessment delivery and processing (Diagram 1), but do not attempt to automate that process. To ensure the integrity of the assessment aims are maintained it is important to recognise when it is appropriate to use each form of assessment. There is a danger that changing the form of assessment to suit the technology will impact the precision and purpose of the assessment and so e-assessment must always be adopted in an informed manner.

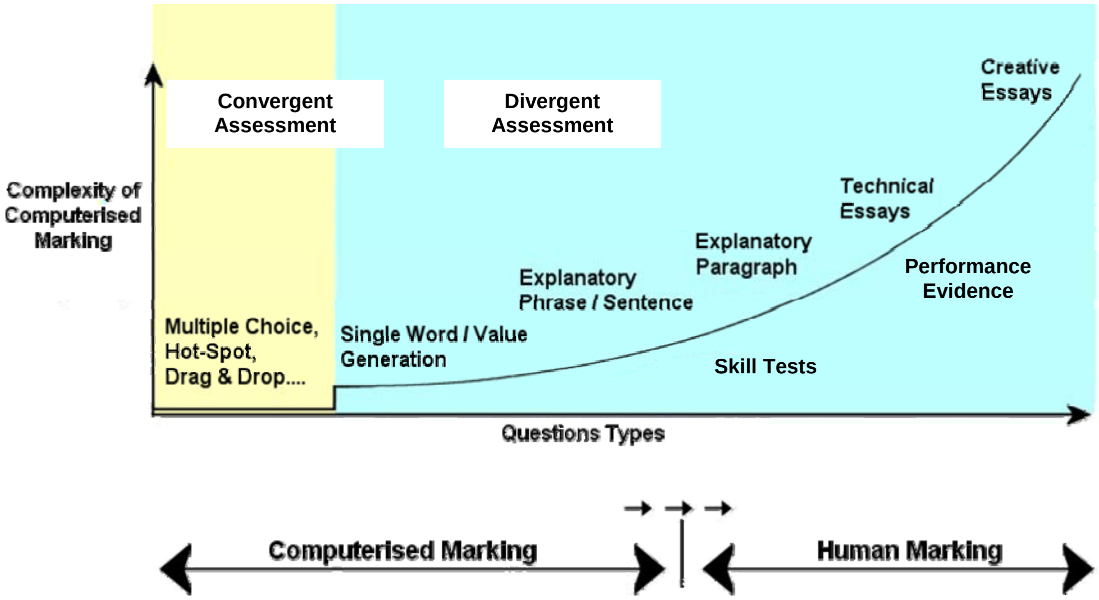

Diagram 1 – Assessment spectrum and relationship with marking complexity

It is ironic that some of the more objective subject and sector areas bring with them particular technical challenges when e-assessment solutions are considered. Mathematics and the Sciences lend themselves well to objective forms of assessment, and so potentially to e-assessment solutions, but they also present unique challenges. The use of and interaction with formulae and symbols is not easy using conventional keyboards or peripheral devices and yet the use of such characters is commonplace within examination of the subjects and forms an integral part of the assessment. It is relatively easy to present such information as part of questions or tasks but more demanding to allow candidates to include use of such characters or symbols in their responses. A simple solution would be to convert all response types to a range of multiple choice options but the question then becomes one testing recognition and selection rather than construction and understanding. Changes for the sake of convenience or limitation of technology should be avoided.

As far back as the late 1980's Cambridge Assessment was exploring the possibility of developing computer-based assessment solutions for vocational qualifications. The Word Processing Functions Test had been introduced alongside the well respected RSA Secretarial Skills Typing Exams and this was identified as a suitable target for e-assessment development. It was already assessed via computer output and aimed at Level 1/Level 2 skills within the National Qualifications Framework (NQF). The first iteration of the test (QWIZ) was aimed at 3 leading MS-DOS word processing applications and required candidates to work within the applications where their use of functions was tracked and scored. This implementation remained true to the principles of the assessment – practical rather than theory – and attempted to assess the application of knowledge in an authentic and valid manner [15, 16]. The system was able to do this with a high degree of success but soon fell foul of the planned obsolescence of desktop applications and required almost annual updates to keep pace with the latest versions of word processing software.

This development burden was further impacted by the introduction of MS Windows and the significant difference that was then introduced between old and new software applications. The complexity of a program-based e-assessment solution then became too complex to maintain [17, 18]. By this stage alternative forms of assessment had begun to emerge in the market where the true use of the application was no longer assessed, but simply knowledge of how the software worked. This was typically demonstrated by multiple choice questions containing screenshots of the software application. For Cambridge Assessment this was no longer an assessment of 'can do' but of 'knows how' and had deviated too far from the original assessment objectives.

Attention then turned to the leading global desktop IT application qualification OCR Computer Literacy and Information Technology (CLAIT). With over 250 thousand candidates a year and available to centres regardless of the software they were using it seemed like an ambitious but achievable target. Working in conjunction with the University of East Anglia the aim was to develop a computer-based assessment solution that would assess the output from task-based assessment regardless of software package used [19, 20]. The first three assessment engines were developed for word processing, spreadsheets and database applications and proved to be highly reliable, efficient and valid forms of assessment [21, 22]. The simple principle was that digital output was saved to or converted to generic file formats that were then uploaded to a remote automated assessment system that processed and marked the submissions. Results reports were emailed back to the centre and submitted with paper-based output for certification claims.

The system was so effective that it won the British Computer Society award for Applied Artificial Intelligence but despite this success there were still practical issues surrounding its sustainability. It soon became apparent that even a system that attempted to remain distant from software upgrade paths could not keep pace with the rate of change associated with ICT and operating system development. The programme was phased down in 2004 in favour of computer-mediated solutions [23, 24]. In this area of development the goal is not to automatically assess the outcomes but to use technology to facilitate the supply of digital evidence to the awarding body. This is a concept that is proving very fruitful; from hundreds of thousands of digitised scripts submitted for on-screen marking supporting mainstream UK school GCSE and GCEs, CLAIT and other vocational IT qualifications uploading digital files, the e-scape project supporting portfolio-based evidence for GCSE product design through a range of successful National Vocational Qualification (NVQ) programmes using eportfolios. The early evidence suggests that computer-mediated solutions may prove a more viable long-term solution for complex forms of assessment.

Where multiple choice is already the preferred mode of assessment we have seen gradual, but optional, forms of e-assessment introduced for a range of qualifications. At OCR the national key and basic skills qualifications (numeracy and literacy) now assess several thousand candidates each week using computer-based tests. Cambridge ESOL has successfully developed and introduced computer-based tests for the listening and reading (comprehension) components of a number of their qualifications. These are now offered to thousands of candidates around the world. However, it should be noted that such services are often still offered as an option only as many education centres still lack the infrastructure or capacity to offer a-assessment for high stakes assessment at scale. The requirement to offer a dual-delivery service remains a considerable constraint on a global assessment agency's ability to expand e-assessment services. Within multi-national assessment programmes it is not ethical to disenfranchise centres and candidates simply because they do not have sufficient access to technology. Despite these challenges there is widespread opinion that the use of technology to support assessment will be a continuing trend over the next few decades [25, 26].

Alongside these more simple forms of assessment the Cambridge Assessment Group continues to pioneer innovative and challenging forms of e-assessment. Interactive and scenario-based assessment has been explored through the ENIGMA Science simulations, primarily used to support formative assessment, and the Cambridge International Examinations Geography IGCSE fieldwork simulations. Both ventures have been successful and proved extremely useful in the assessment of applied underpinning skills and knowledge. This is an area that has proved difficult to standardise in traditional assessment settings and so some clear benefits are being realised. Organisations should note that the cost of production for simulation-based assessment is considerably higher than traditional materials and so it is advisable to seek to achieve reusability of the simulation scenarios in future assessments.

Alongside the work of ETS (US) and Intelligent Assessment (UK) the Cambridge Assessment Group have also explored the potential for true automated short and extended/essay responses. There are examples of where this work has proved to be a suitable alternative to a human marker where double marking was previously in place [27, 28] but there are few, if any, examples of where the technology has proved reliable enough to implement as the single marker for high stakes assessment programmes. This remains a rich area of e-assessment research as organisations seek ways to make the marking more reliable [29, 30] or look to apply the techniques to shorter text-based responses [31]. The systems have also proved to be reasonably reliable where an assessment programme works off a limited number of pre-set essay topics but that brings with it additional security concerns regarding exposure of assessment material. Such concerns become very real in an era when numerous web-based services exist to provide model answers for known essay topics. In practice they are currently used within formative assessment higher education programmes or to replace one of the human examiners in a dual marking system. It is worth noting that they usually become most unreliable at the top end of achievement, where responses may be acceptable but so unique that they deviate too strongly from the published markscheme. Such a situation is deemed unacceptable for Cambridge Assessment when considering deployment in a high stakes environment.

Capture and assessment of audio files has also been explored by Cambridge ESOL and subject areas such as modern foreign languages within OCR and CIE. In this domain a computer-mediated approach lends itself well to acceptance of digital audio files for transfer to human markers and this now forms an integral part of quality computer-based assessment systems. Despite developments in the field of automated call centre technology there are still too many constraints placed on what can be assessed orally for it to be used for automatic assessment in mainstream speaking examinations. For the assessment solution to remain valid the technology should not unduly constrain what can be assessed (spoken) and must be highly reliable. Voice recognition technology, if the intention is to use it for authentic speaking assessment, may still be too immature for this purpose in its current state.

As mentioned earlier there is much wider scope for the application of technology to support assessment; often described as computer-mediated assessment (CMA). In this context there is less chance that the nature of the assessment tasks or expectation of evidence will be constrained to fit the available technology solution. The Cambridge Assessment on-screen marking programme is an example of wide scale adoption of CMA where technology is used to facilitate the capture and transfer of evidence to human markers. It also provides an opportunity to explore new methods of marker allocation and quality assurance [32, 33, 34] but it is likely that such changes will be introduced in a gradual manner while computer-based and paper-based assessment services exist in parallel.

2.2 Key drivers and challenges

It is important to consider the key drivers and challenges for e-assessment when deployed at scale on an international basis. A summary of the common drivers in relation to adoption of technology-based assessment solutions is given below:

- Constant pressure to improve reliability of operational and assessment processes
- Increased demand for quicker end to end services
- Increased demand for but decreasing availability of ‘expert assessment personnel’
- Ongoing desire to review and control overheads/running costs
- Maturing customer capability and confidence in the use of technology
- Ever increasing availability of technology options and solutions
- Assessment innovation options enabled by the use of technology

Despite numerous reasons for adoption of technologies it is sobering to consider what little impact technology has had on education and training practice and assessment forms in particular. Perhaps this is no surprise when we consider how important it is to maintain public confidence in any national assessment system [35]. When married with the fact that valid assessment must always remain true to the core values of content and subject domains we can understand why many disciplines have been reluctant to adopt inappropriate modes of assessment. The Driving Standards Agency (DSA) driving theory test in the UK is often cited as an example of e-assessment adopted on a massive scale for high stakes testing. This is true, but what is conveniently overlooked is that the theory component of the test plays a secondary and supplementary role to the practical part of the driving test. In this context the theory component is lower stakes than the practical and was always assessed via a series of objective questions, in a relatively unregulated manner, well before e-assessment was introduced.

This is not to say that highly objective forms of assessment, such as multiple-choice, are bad but they can be damaging when used in an inappropriate way. It is also not a given that they can only assess low-level areas of knowledge and understanding. When used well they can assess some complex and high order cognitive skills as the Cambridge Assessment Thinking Skills Assessments (TSA) prove [36, 37, 38]. The crucial issue is to ensure that convenience does not outweigh authenticity when selecting modes of assessment.

There are other challenges not addressed in detail by this paper, and these can broadly be described as practical and commercial constraints. The cost of developing innovative and complex new forms of e-assessment is high and if there is little potential to transfer those technical solutions to other domains the incentive to invest will be low. Despite the relentless advances made around power, portability and accessibility of technology it is still not a sufficiently robust and reliable medium for delivery in all areas of education and training [39]. If we add the global dimension to this equation the reach of robust and reliable technology solutions is further reduced. Even where an institution has access to a robust and reliable technology infrastructure it is often not possible to access this system at scale – for national assessment several hundred students may need to access an assessment at the same time and that is just at an institution level. The provider of the assessment service may then need to offer those tests to several thousand candidates at the same time and this presents some real challenges to resilience.

Despite the limitations of paper-based assessment it is a highly reliable form of delivery once the materials are at the institution. There are few things that can go wrong on the day of the assessment itself and this is very attractive in a high stakes assessment environment [40]. When faced with such challenges it is not surprising that many national assessment systems stick with the known and traditional modes of delivery unless there is sufficient external intervention to make them shift. Consider for a moment the UK initiative to move all consumers to digital TV by 2011 – would so many people have upgraded their equipment if the old service was not being removed?

The final area that should be considered when adopting e-assessment solutions is one of accessibility. This is accessibility in the context of the Disability Discrimination Act (DDA). There are many enabling features of technology that can be highly empowering for an individual who wishes to access an assessment. Most advances have been made in the fields of peripheral devices and text to speech software and these are easy enough to incorporate in e-assessment solutions if they are considered at the design stage. There are some concerns that e-assessment will seek to use innovative modes of delivery simply because they exist (eg simulations, interactive diagrams etc) without thinking through the rationale for including them in the assessment instrument. This brings us back to the principles of the Cambridge Approach and 'fit for purpose' assessment – if the instrument is no longer measuring what you set out to measure one must begin to question the validity of the outcomes. Accessibility concerns should not be viewed as a reason for not adopting innovative e-assessment solutions but due care and attention should be paid to them during the initial design of those solutions.

3 Conclusion - Fit for purpose applications

This paper presents an insight into the qualities of assessment objectives that may lend themselves to an e-assessment solution and conversely those qualities that will provide the greatest challenge. If an assessment model already makes use of a series of atomised objective questions it should lend itself well to automated e-assessment, if the target audience can accommodate the technical and capacity requirements.

There is also increasing scope for the application of mediated e-assessment solutions as the creation of digital evidence, or conversion of paper to digital, becomes more prevalent. This area may hold the greatest opportunity for developers of e-assessment services as it is less likely to compromise the original intention of the assessment. Although the full benefits of automated assessment cannot be realised there are still numerous possibilities for the acceptance of a wider and more authentic range of evidence from candidates and dramatic improvements in turnaround of marking and issue of results.

The underlying theme of this paper is that technology should not be deployed simply for the opportunity to realise significant cost-savings or efficiency gains, although these might be useful components of an e-assessment solution. The key drivers should align with the initial aims of the assessment to ensure that the e-assessment solution continues to meet those objectives or even better, leads to quality or service improvements.

Only by taking this approach, the Cambridge Approach, can we ensure that technology finds its rightful place within the overall mix of assessment research and development. Over time the technological challenges will lessen, user confidence will improve and accessibility issues will be addressed at design stage. The challenge of ensuring assessment solutions are 'fit or purpose' will never disappear as long as stakeholder confidence in the outcomes remains a priority for all concerned. This one key factor will continue to remain the single biggest challenge for developers of e-assessment solutions.

References:

[1] Harding, R. and Craven, P., ICT in Assessment: A three-legged race, presented at Qualifications Curriculum Authority (QCA) Research Seminar, London, 2001

[2] Bloom, B.S., et al, Taxonomy of Educational Objectives Handbook I: Cognitive Domain, McKay, New York, 1956.

[3] Knight, P., A Briefing on Key Concepts – Formative and summative, criterion and norm-referenced assessment, LTSN Generic Centre Assessment Series No. 7, ISBN 1-904190-05-7, 2001

[4] Raikes, N., Greatorex, J and Shaw, S, From Paper to Screen: some issues on the way, International Association for Educational Assessment Conference, 2004

[5] Johnson, M. and Green, S., On-line assessment: the impact of mode on student performance, British Educational Research Association Conference, 2004

[6] Johnson, M. and Green, S, On-line assessment: the impact of mode on students' strategies, perceptions and behaviours, British Educational Research Association Conference, 2004

[7] Bennett, R., et al, Does it Matter if I take my Mathematics Test on Computer? A second empirical study of mode effects in NAEP, The Journal for Technology, Learning, and Assessment, Volume 6, Number 9, 2008

[8] Johnson, M. and Green, S., On-line Mathematics Assessment: The impact of mode on performance and question answering strategies, The Journal for Technology, Learning, and Assessment, Volume 4, Number 5, 2006

[9] Bennett, R., et al, Does it Matter if I take my Writing Test on Computer? An empirical study of mode effects in NAEP, The Journal for Technology, Learning, and Assessment, Volume 5, Number 2, 2006

[10] Pommerich, M., Developing computerised versions of paper-and-pencil tests: Mode effects for passage-based tests, The Journal for Technology, Learning, and Assessment, Volume 2, Number 6, 2004

[11] Raikes, N., and Harding, R., The Horseless Carriage Stage: replacing conventional measures, Assessment in Education, Volume 10, Issue 3, 2003

[12] Johnson, M. and Greatorex, J., Judging Text Presented on Screen: implications for validity, E-Learning, 5(1), 40-50, 2008

[13] Shaw, S. and Imam, H., On-screen essay marking reliability: Towards an understanding of marker assessment behaviour, IAEA Conference, 2008

[14] Raikes, N., On-screen Marking of Scanned Paper Scripts, UCLES Research Conference, 2002

[15] Dowsing, R.D., Long, S. and Sleep, M.R., Assessing Word Processing skills by computer, Information Services and Use, Volume 18, Number 1-2, p.15-24, 1998

[16] Dowsing, R.D. and Long, S., An Evaluation of the Impact of AI Techniques on Computerised Assessment of Word Processing Skills, 9th International Conference on Artificial Intelligence and Education (AIED 99), Le Mans, France, IOS Press, 1999

[17] Long, S., Computer-based Assessment of Authentic Word Processing Skills, Doctoral Thesis, School of Information Systems, University of East Anglia, Norwich

[18] Long, S., Outcomes of the first live pilot of WP Marker, Project Report, School of Information Systems, University of East Anglia, Norwich

[19] Dowsing, R. D., and Long, S., The Algorithmic Basis for IT Skills Automated Assessment, Computers in Advanced Technology Conference (CATE'99), New Jersey, USA, IASTED/Acta Press, 1999

[20] Dowsing, R. D., Long, S. and Craven, P., An Analysis of the difference between traditional and computer-based assessment of IT Skills, ASCILITE 2000, 477—486, 2000

[21] Dowsing, R.D. and Long, S., Building a Computer-Based Assessor for IT Skills with enough intelligence, International Conference on Intelligent Systems and Applications (ISA 2000), Wollongong, Australia, ICSC Academic Press, 2000

[22] Long, S., Dowsing, R.D. and Craven, P., Knowledge-based Systems for Marking Professional IT Skills Examinations, Proceedings of ES2002, Second SGAI International Conference on Knowledge-based systems and Applied Artificial Intelligence, Cambridge, UK, 2002

[23] Long, S., Lessons in the Development and Deployment of Automated IT Skills Accreditation, Fourth International Computer Assisted Assessment Conference, Loughborough, UK

[24] Dowsing, R.D. and Long, S., Trust and the automated assessment of IT Skills, Special Interest Group on Computer Personnel Research Annual Conference, Kristiansand, Norway, Session: 3.1 IT Careers, p90-95, 2002

[25] Bennett, R., Inexorable and Inevitable: The Continuing Story of Technology and Assessment, The Journal of Technology, Learning and Assessment, Volume 1, Number 1, 2002

[26] Harding, R., What have Examinations got to do with Computers in Education?, UCLES Conference Report, 2003

[27] Dikli, S., An Overview of Automated Scoring of Essays, The Journal of Technology, Learning and Assessment, Volume 5, Number 1, 2006

[28] Wang, J. and Stallone-Brown, M., Automated Essay Scoring Versus Human Scoring: A Comparative Study, The Journal of Technology, Learning and Assessment, Volume 6, Number 2, 2002

[29] Ben-Simon, A. and Bennett, R., Toward More Substantively Meaningful Automated Essay Scoring, The Journal of Technology, Learning and Assessment, Volume 6, Number 1, 2007

[30] McCallum et al, Improving text classification by shrinkage in a hierarchy of classes, ICML-98, 1998

[31] Siddiqi, R. and Harrison, C., On the Automated Assessment of Short Free-Text Responses, IAEA Conference Paper, 2008

[32] Bramley, T., A Rank-Ordering method for equating tests by expert judgement, Journal of Applied Measurement, Volume 6, Issue 2, 202-223, 2005

[33] Newton, P., et al, 'Paired Comparison Methods', in Eds, 2007

[34] Suto, I. and Nadas, R., Towards a new model of marking accuracy: An empirical investigation of an international biology examination, IAEA Conference Paper, Cambridge, UK, 2008

[35] Desmond, T. and Desmond, M., Large Scale Assessment – maintaining public confidence in high stakes state examinations (State Examinations Commission, Ireland), IAEA Conference Paper, Cambridge, UK, 2008

[36] Robinson, P., Thinking Skills and University Admissions, presentation to Cambridge International Examinations Seminar: Thinking Skills for the 21st Century, Cambridge, UK, 2002

[37] Chapman, J., The Development of the Assessment of Thinking Skills, Assessment of Thinking Skills Conference, 2005

[38] Black, B., Critical Thinking – a definition and taxonomy for Cambridge Assessment: Supporting valid arguments about Critical Thinking assessments administered by Cambridge Assessment, IAEA Conference Paper, Cambridge, UK, 2008

[39] Thomson Prometric Study, Drivers and Barriers to the adoption of e-Assessment for UK Awarding Bodies, Learning Solutions, 2005

[40] Hermans, P., The quality management of large-scale computer based assessments (CITO, Netherlands), IAEA Conference Paper, 2008

Author(s):

Patrick Craven - Principal Analyst
Cambridge Assessment - Assessment Research and Development Division
1 Regent Street, Cambridge, CB2 1GG
craven.p@cambridgeassessment.org.uk