The validity of an innovative on-screen assessment

Andrew Boyle
Principal research officer
e-assessment
Qualifications and Curriculum Authority (QCA)
83 Piccadilly
London
W1J 8QA

www.qca.org.uk

boylea@qca.org.uk

Background

The Qualifications and Curriculum Authority (QCA) is the regulatory authority for 3 – 19 education in England. The QCA 'maintains and develops the national curriculum and associated assessments, tests and examinations; and accredits and monitors qualifications in colleges and at work' (QCA, 2005a).

As part of that role QCA is developing a new national curriculum test for information and communication technology (ICT) under contract from the Department for Education and Skills (DfES). This test will assess all pupils at the end of key stage 3 (KS3); the lower secondary phase of education in England. It is intended that this test will be used on a statutory basis for all pupils from 2008.

The KS3 ICT test is an innovative on-screen assessment. The test is designed to address the construct of ICT capability, which is defined as follows:

> 'ICT capability is about having the technical and cognitive proficiency to access, use and communicate information using technological tools.
>
> Learners demonstrate this capability by purposefully applying technology to solve problems, analyse information, develop ideas, create models and exchange information.
>
> They are discriminating in their use of information and ICT tools.' (DfES, 2004; see also: Peppiatt, 2004)

Given this definition of the assessed construct, it was felt inappropriate to use traditional models to design this test. Instead, a sophisticated virtual toolkit (a set of simulated office-type applications[1]) was developed. Also, there was a walled garden of virtual assets (simulated web pages and data files of varying types held in a

[1] The toolkit (or 'office suite of applications') is 'virtual' or 'simulated' in the sense that it is a generic set of applications designed for use in a test. It is not a fully functioning suite of office-type applications.

structure of file directories). Then, pupils were presented (by email) with authentic-looking tasks to perform using the toolkit applications and assets.

There are no marks in this test in the conventional sense; rather, pupils' actions using the virtual applications to respond to the tasks are tracked and aggregated. This aggregation is carried out with reference to a new document that QCA has developed. The 'Rules Base' is a sophisticated branching database which starts – at the macro level – from national curriculum level descriptions, which are then disaggregated through various granularities of detail. The most fined-grained entities that can represent meaningful ICT capability are known as 'opportunities'. Opportunities are constituted from an aggregation of pupils' mouse clicks, keystrokes and so forth.

The table below illustrates a small part of the Rules Base:

Level Description sub-division	**Granularities of the Rules Base**	
	Process indicators	**Elaborations**[2]
(A) Pupils select the information they need for different purposes, check its accuracy and organise it in a form suitable for processing.	i. (A) Select information/assets for specific purposes	(b) Check accuracy by finding information/assets from more than one source (i, ii)
		(c) Check validity by finding information/assets from more than one source (i, ii)
		(d) Select relevant parts of the information/assets gathered, ignoring irrelevant parts (i, ii, iii)
	ii. (A) Organise information/assets for processing	(r) With guidance select technology tools for problem solving and decision making (i, iii, iv)
		(y) Select and apply technology tools for information analysis (ii)

Table 1: A small section of the Rules Base, as used in the 2005 pilot

The test's awarding process operates by setting thresholds for the numbers of levelled opportunities that a pupil needs to fire to demonstrate sufficient evidence that s/he has reached a given national curriculum level. This awarding procedure is known as the sufficient evidence model (see Research Machines, 2005a and 2005b).

A demonstration of the test is available on the web site: www.ks3ictpilot.com.

[2] Opportunities are not technically a part of the Rules Base but are developed as a more finely grained sub-set of elaborations.

Evaluation

The key stage 3 ICT test development project is run at QCA by a small project team. The head contractor on the project is Research Machines PLC (RM), who have several sub-contractors providing a range of specialist expertise; for example, in measurement issues.

In addition to the test development work of the project team and the contractor, this project has been subject to an ongoing programme of evaluation work. This evaluation is provided by the author of this paper.

The project evaluator is a QCA employee. However, he is not a member of the team that is responsible for developing the test. Further, he has no explicit responsibilities for delivering any specific facet of the test development. As such, this is an independent evaluation.

The evaluation has had informal and formal aspects. Informal aspects include the provision of measurement advice on a range of issues, working with specialists in the contracting consortium to specify important measurement issues for this innovative test model, and similar work.

The evaluation produced formal output in 2005 which addressed several audiences. Principal amongst these were interim and final evaluation reports (QCA, 2005b and QCA, 2005c, respectively). These reports were delivered, in the first instance, to the Senior Responsible Officer of the QCA project team, and – thereafter – to the DfES in July and October 2005, respectively.

The evaluation judged the project against five objectives. In 2005 these objectives addressed the entirety of the project, and hence a disparate set of issues:

- Validity
- The scalability of infrastructure software and support processes
- Accurate formative and summative reports
- Test security
- School experience

This conference paper will describe the evaluation of the first objective; validity.

A particular issue for the evaluation was the standard of proof that could be used to judge the validity of the 2005 pilot test. Being a pilot year – and with two further pilots due in 2006 and 2007 – it would have been unreasonable to demand that the 2005 pilot had a flawless record on validity.

However, it was also important that a stringent standard of proof was applied to the evaluation of validity. To square this circle, the following construct was devised: it was posited that the 2005 test did not have to be demonstrably perfect in order to be declared to be a successful pilot. Rather, the professional judgement of the evaluator was applied to state whether it seemed likely that the tests would be able to be delivered in 2008 to the high quality that is required for national curriculum tests.

Validity

Definitional issues

Validity has been widely agreed to be the central concept in understanding the quality and appropriateness of a test and its uses. It has had many definitions; however, in the current context it has not been appropriate to adopt a single definition of the concept wholesale.

Rather, it is easier to understand the practical import of validity for this test development by examining several of its features. Whilst this does not provide a incontrovertible explanation of validity, it may allow the reader to appreciate what validity and its investigation meant in the 2005 pilot of the KS3 ICT test.

Firstly, the American publication *Standards for Educational and Psychological Testing* (AERA et al, 1999) can be used to provide a central characteristic of validity. That is, the test developer and sponsor are under an active duty to provide strong evidence that the test is valid. If such evidence is absent, or questionable, then the best interpretation is that the test has not been demonstrated to be valid.

Further, validity was taken to be made up of several facets. These included:

- Face validity
- Reliability
- Fairness for all pupils
- Content evidence of validity
- Concurrent evidence of validity
- Level setting procedures and process

The main body of this paper will briefly define each facet of validity, and then describe findings for the facets.

A further issue that has concerned the evaluation and the project more widely is whether validity can be taken as a single unitary concept with a number of facets, or whether it is better to conceptualise a group of distinct validities, or types of validity.

The evaluation has adopted the position that validity should be viewed as a single, indivisible construct. This is for the following reasons:

- To allow a single evaluative judgement to be made as to whether the 2005 pilot test was valid or not.
- To emphasise that all facets of validity are necessary conditions – for example to negate any tendency to promote a particular facet of validity as *prima inter pares*.

Validity findings

The next substantial section of this paper reports a selection of findings with respect to different facets of validity. Where relevant, mitigations to validity issues that are known to have been instituted since the 2005 evaluation are also described.

Face validity

Face validity can be defined as:

> the extent to which a test (and its outcomes) is perceived to be accurate, appropriate and useful by non-technical users

The main finding on face validity was that there was evidence that test users found the instrument face valid for most levels of pupils to whom it was addressed.

However, there was concern from users that the test was not face valid for the highest level of pupils who sat the test in 2005. (This is related to a wider issue that is referred to below at page 8.)

Whilst users' perceptions of the test in 2005 were generally positive, there were several issues that were observed on visits to schools. These are listed below.

Firstly, some teacher respondents felt that there had been too much text in emails that conveyed task instructions, and also that some of the vocabulary in instructions had been too difficult. These phenomena were felt to impact particularly on pupils who were weak readers generally.

Relatedly, it was also felt that the message pane in the email applet could become very narrow; and that this could decrease the readability of the instructions texts. Finally, it was also observed that, in some classrooms, pupils using more that one application could get a very cluttered screen. The figure below illustrates one such screen that was observed:

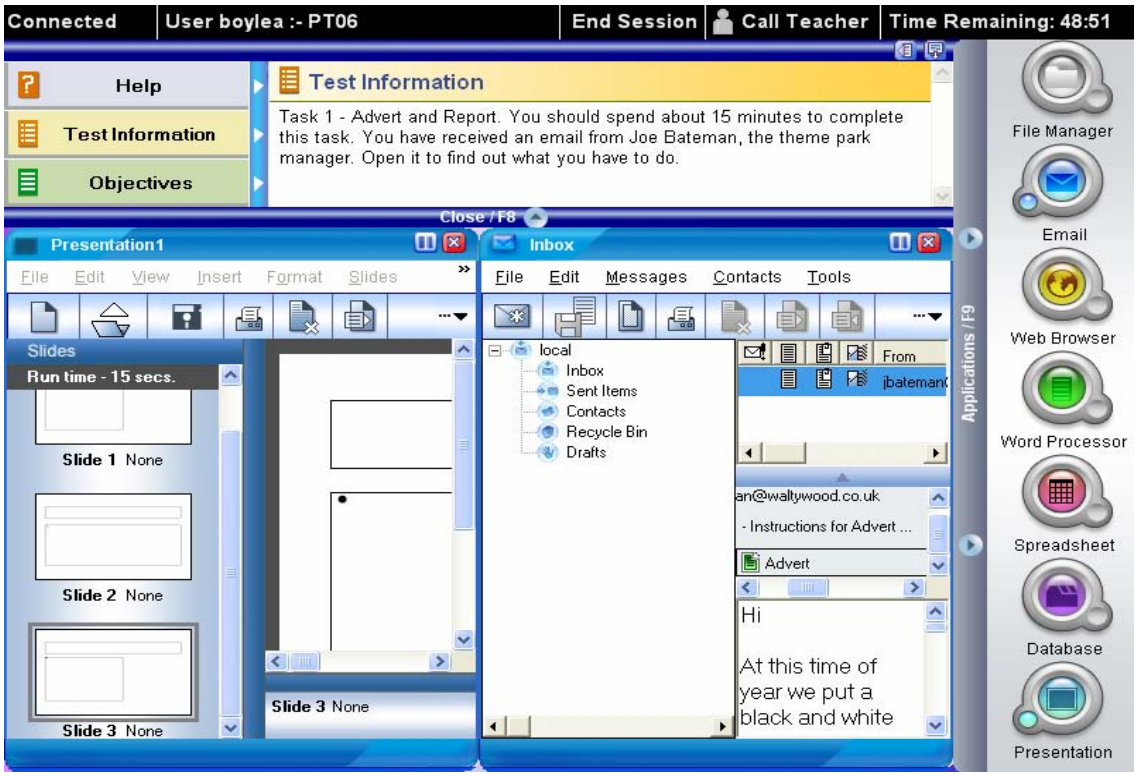

Figure 1: An example of a cluttered screen

The issue of cluttered screen has been the harbinger of a wider issue related to sources of difficulty in the test. This has taken several forms. One example would be a pupil who read an email, then clicked on a file attachment, did some work (say working on a formula in a spreadsheet), but then struggled to get back to the original email to find out the next step in the instruction.

It was from observations such as that described above that it was hypothesised that this novel test might have sources of difficulty that were quite different from those which pertain in conventional tests. Further, it was not clear of the extent (strength) of such difficulties, their relevance to ICT capability (the construct being assessed) and their differential impact across groups of pupils.

A range of work has been done in 2006 test development to mitigate some of these problems observed in 2005. Firstly, the project continues to control the readability of task instructions through a variety of methods. This includes both review by relevant experts (e.g. ICT curriculum, English as an Additional Language (EAL) and Special Educational Needs (SEN) specialists) and the calculation of readability indices to establish whether the reading difficulty of instructions is appropriate, given the age of pupils taking the test.

It is acknowledged that conventional readability analyses have only limited usefulness. Their efficacy has been doubted both in the case of 'paper-and-pencil' exam questions (Allan et al, 2005) and in the case of reading texts from screen (Dyson, 2004). However, the development of an appropriate index of readability of on-screen test instructions would be a substantial task. The use of conventional readability analyses is defended as a 'rule of thumb' – a rough check to ensure that instructions are sufficiently readable; used alongside the several expert reviews described above.

The functionality of the virtual toolkit has been improved following teachers' and pupils' feedback. Improvements have been made to the operating system; for example, by adding minimise and maximise buttons. The addition of these functions will bring the toolkit closer to standard operating systems such as Windows. It is felt that such improvements to the test interface, allied to increasing familiarity amongst teachers and pupils as the test moves towards high-stakes roll-out, will make problems such as the illustrated cluttered screen issue less frequent.

The issue of the potentially novel sources of difficulty in the test poses a more fundamental question. Many have written about the great potential of e-assessment to provide richer, more authentic assessment (cf. Boyle, 2005). However, there is virtually no research investigating potential sources of difficulty in such novel assessment models.

A sub-project has been set up to remedy this omission. A team from the University of Leeds has been contracted to hypothesise, categorise and investigate sources of difficulty. A particular issue is to establish the construct relevance of such difficulties. The project will output a structured taxonomy – basically mapping an unknown terrain for the information of future researchers. It will also produce practical advice for task writers; helping them to improve the tasks in future tests – to make sure that sources of difficulty are relevant to the assessed construct.

Reliability

Reliability is a crucial aspect of a test's validity. Whilst validity is the overarching concept by which to judge a test's quality and appropriateness, if a test is not reliable then, in effect, it is not measuring anything at all.

There are many features of a definition of reliability which could be rehearsed profitably in a longer, or more specifically-focused, paper. In the current context it is sufficient to say that most reliability analysis techniques seek to estimate the extent to which a test data set implies that the instrument produces replicable measurement. In many conventional testing programmes, the extent of replicability is mimicked by calculating the internal consistency of data produced by a single test administration. Although internal consistency analyses are practical to conduct, they are not necessarily the most direct or principled method for collecting data for a reliability study. A more theoretically-principled data collection approach is the 'test-

retest method'; that is, getting a sample of the test-taking population to sit the same version of a test twice and monitoring the extent to which the result (e.g. the national curriculum level awarded) remains the same over the two administrations.

The most commonly-used internal consistency reliability indices have been criticised for only providing an implicit description of the extent to which a test reliably classifies pupils into levels (Wiliam, 2000). In contrast, some reliability analysis techniques produce an explicit quantification of the reliability of classification into levels as their main output. Such classification consistency analyses are most often carried out following test-retest data collections.

Reliability findings from the 2005 pilot are best described as exploratory: various reliability analyses were conducted (using internal consistency and test-retest data collections) on various entities (e.g. test tiers, forms, scores for specific levels, pre-test, full summative pilot test data sets, and so on).

Results were extremely varied. (RM (2005c) and QCA (2005c) contain a full description of reliability results and methods used.) Since the 2005 evaluation, substantial research has been conducted to specify the types of reliability study that will be most appropriate for this novel type of test. It has been decided that the key concept in reliability analysis of the KS3 ICT tests will be classification consistency. Further, such analyses will be based on sophisticated test-retest data collections.

If such reliability investigations are successful (and there are good reasons to suppose that they will be), then the KS3 ICT test could even be a vanguard for a principled and transparent approach to the investigation and reporting of the reliability of high-stakes tests.

Fairness for all pupils

It is important that any high-stakes test is fair for all those who take it. In this context, fairness refers to groups of pupils with identifiable demographic characteristics (e.g. gender, EAL status, Free School Meals (FSM) entitlement and to pupils with SEN).

The following is a useful definition of 'fairness for all pupils':

> Fairness … addresses the question of whether students given the same quality of preparation and who have the same degree of motivation would be likely to perform similarly in the examinations in question. Fairness involves the extent to which the test administration and scoring practices are comparable across identifiable groups of students. … Our use of the term 'fairness' in this fashion is not intended to convey that the performances of particular subgroups should be more or less equal, although that use of the term is sometimes made. Differences in group performance may be due to differences in preparation, e.g. quality of teaching, access to support, motivation, as well as to any differences among the subgroups, such as English language proficiency. (International panel, 2002)

Thus, for a test to be fair, it does not mean that all groups of pupils must score at the same level. Rather, it means that differences must be proportionate, represent the underlying abilities of pupils and be consistent with other information on groups of pupils' typical abilities.

Findings from the 2005 pilot on fairness for all pupils were as follows:

- The test appeared to be fair in all substantial respects for boys and girls.

- Pupils entitled to FSMs scored less well in the pilot than those who were not so entitled. However, this lower scoring was consistent with patterns of lower scoring for pupils entitled to FSMs across a wide range of national tests and certificated examinations (regrettably though this may be). Thus, the judgement was that this test was fair to pupils entitled to FSMs.
- Analysis was conducted to compare the scoring of pupils who spoke English as their first language with 'others'. The findings in this category were inconclusive; pupils in the 'others' category scored less well than those who spoke English as a first language. However, it was not clear whether this represented those pupils' genuinely lower abilities.

  It would be more helpful if a more appropriate approach could be found to analyse the performance of pupils with EAL; such an approach would need to reflect the diversity of these pupils – in terms of their general language competence and their literacy in particular.
- The scoring of pupils with SEN was compared with that of pupils without special needs. All pupils were found to have scored more lowly than their teachers' initial estimates. However, pupils with SEN scored particularly lowly, when compared with pupils without SEN.

  As such, it still remains for the test to demonstrate that it is fair for pupils with special needs.

Content evidence of validity

Content evidence of validity can be defined as 'whether a test adequately targets and represents the whole domain of performance upon which it purports to report'. In the current case, the whole performance domain is defined in the national curriculum for ICT.

Findings on content evidence of validity included:

- More than 70 per cent of Rules Base elaborations were included in the test at all national curriculum levels (three to six), thus complying with the pre-agreed standard for coverage of the Rules Base.
- A QCA Teacher Review Group gave their view as to coverage of the ICT programme of study (which – in some ways – is similar to a syllabus). This group of teachers thought that approximately 80 per cent of the programme of study was covered.
- Several of the non-covered aspects of the programme of study relate to the 'Communication' part of ICT. One teacher commented as follows:
  [It is] hard to see how 'Reviewing, modifying and evaluating' can be covered … many of the statements include the terms 'share', 'discuss', and 'reflect'. Similarly with 'Breadth of Study', 'Working with others'. Generally it's about the 'C' in ICT and how this is assessed.
- There were several issues concerning the counting of the amount of material in the tests. These can be summarised as follows:
  - There was a weakness in the method used for counting opportunities prior to the 2005 test. This consisted of difficulties in assessing which opportunities were distinct and unique, and in how to most meaningfully count the numbers of opportunities that were available to any one pupil (for example, if a pupil could select either 'paste link' or simple 'paste' at a given moment in the test, then it would be more sensible to count those two options as one available opportunity, rather than two).
  - Whichever way it was counted, it was clear that there was less test material in the top level (level six) than was acceptable in the 2005 test.

This caused several problems, in the areas of: face validity, reliability and level awarding – as can be seen elsewhere in this paper.

Concurrent evidence of validity

In 2005, concurrent evidence of validity was sought using three data collection methods:

- Skilled observers watching pupils' tests and then assigning a national curriculum level to each performance.
- Skilled moderators viewing detailed reports produced by the test software to describe pupils' performance in the test.
- Collection of teachers' assessment of pupils' national curriculum levels.

Unfortunately, neither of the first two methods produced data that were good enough to use for convincing analysis to establish concurrent evidence of validity. Therefore, this facet of validity had not been demonstrated in 2005.

In 2006, renewed efforts will be made to provide concurrent evidence of validity. This will be done using moderated test reports, and teacher assessment.

Level setting procedures and process

Following the conduct of the 2005 test pilot, pupils were awarded national curriculum levels. The 2005 test could award levels three to six against national curriculum descriptions. Also, a pupil could be awarded no level (or 'level n') if the test had not provided enough evidence for a judgement to be made about which level s/he should be assigned to.

Level awarding was carried out using a *post-hoc* procedure, which was informed by input from a small panel of teachers, the views of RM educational specialists (and their sub-contracting advisors), and of QCA ICT curriculum experts.

Level awarding was conducted according to the 'sufficient evidence model'. Features of the sufficient evidence model included:

- In order to be awarded a national curriculum level, pupils had to gain a specified number of opportunities that were targeted at the level to be awarded.
- As well as demonstrating a certain number of opportunities at the level to be awarded, pupils also had to demonstrate that they would have been awarded the level(s) below as well.
- Level six awarding was performed on a different basis to other levels in 2005. To be awarded level six, pupils had to demonstrate that they were a 'sound level five' – that is, that they had gained a number of opportunities well above the level five cut score – and they also had to have gained a small number of level six opportunities (in fact, the number was one).

The numbers, and percentages, of pupils awarded each level are summarised in this table:

National Curriculum level awarded	Number of pupils	Percentage	Cumulative percentage
'n'	7,715	16.9%	16.9%
3	6,066	13.3%	30.3%
4	15,332	33.7%	63.9%
5	13,731	30.2%	94.1%
6	2,696	5.9%	100%
Total	**45,540**	**100%**	

Table 2: Numbers and percentages of pupils being awarded national curriculum levels

However, prior to the introduction of the KS3 ICT test, summative assessment at the end of this key stage in this subject has been implemented via teacher assessment (TA). Each teacher must make an assessment of each pupil and allocate him/her to a national curriculum level. A comparison was made between the levels assigned by the pilot test and by TA. This comparison is shown in the figure below:

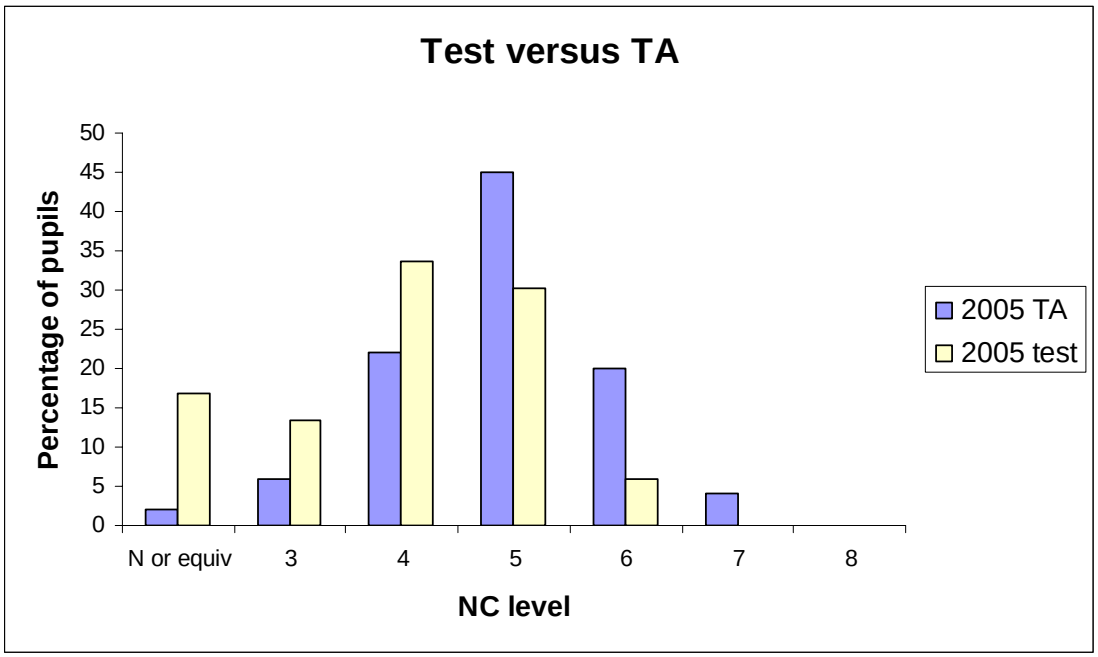

Figure 2: Comparison of levels awarded by test and teacher assessment

The graph above makes plain two aspects of the comparison between test and TA levels:

- Levels awarded by TA are approximately one level higher than the levels awarded by the test – i.e. the distribution of test results has shifted one level 'to the left' when compared with the TA distribution.
- There is a much larger proportion of pupils who have been awarded no level ('level n') from the test as compared to TA.

The fact that the test and TA reported different distributions was not necessarily a problem, as the test was intended to implement the national curriculum level descriptions, not to equate to the pre-existing distributions of TA levels.

However, in order to better understand the difference in the level distributions in 2005, the QCA appointed an independent panel of experts in ICT and level awarding. This panel reported in February 2006 (Independent panel, 2006). The panel concluded that level awarding procedures were sound in principle. However, it also found that there were 'flaws' in the 2005 test and its delivery. But, it further stated that none of these flaws was so serious that it could not be remedied if actively addressed before 2008. The panel also made 18 recommendations to improve the test.

Conclusion – an overall evaluation of validity

The findings reported above include many positive outcomes for the key stage 3 ICT test in 2005. However, there are also substantial areas of work still to address. Once again, as described in this paper, there is evidence that the majority of these outstanding issues are being addressed by the project team and their contractors.

The independent panel investigating the levels awarded in 2005 concluded that although there were flaws in the 2005 test and its delivery, none were so major as to prevent a successful statutory introduction of the test in 2008. This formulation is semantically different from, but substantially the same as, that put forward by the final evaluation report (QCA, 2005c); that is, the 2005 test was valid, *given that this was a pilot year*.

Once again, this positive evaluation of the test's validity stems from the judgement that the problems that did occur in 2005 were consistent with a test in pilot phase, and could be resolved if they were actively addressed before 2008. The large amounts of development work that are already under way (a small portion of which is reported in this paper) engender confidence that the test can be delivered successfully as planned.

Bibliography

Allan, S., McGhee, M. & van Krieken, R. (2005) *Using readability formulae for examination questions.* Research report for QCA. Online, available: http://www.qca.org.uk/14991.html.

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME) (1999) *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Boyle, A. (2005) *Sophisticated Tasks in E-Assessment: What are they? And what are their benefits?* **in** Danson, M. (ed) Ninth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, 5th and 6th July 2005. Online, available: http://www.caaconference.com/.

Department for Education and Skills (DfES) (2004) *Creating effective learners. Pedagogy and practice: teaching and learning in secondary schools. Unit 15: Using ICT to enhance learning*. Online, available: http://www.standards.dfes.gov.uk/keystage3/downloads/sec_pptl043804u15using_ict.pdf.

Dyson, M. (2004) *How physical text layout affects reading from screen*. Behaviour & Information Technology, November–December 2004, Vol. 23, No. 6, 377–393.

Independent panel (2006) *QCA key stage 3 information and communication technology 2005 Pilot Report of the Independent Review Panel on Awarding Procedures.* Unpublished report.

International panel (2002) *Maintaining GCE A level standards: the findings of an independent panel of experts*. Online, available: http://www.internationalpanel.org.uk/.

Peppiatt, M. (2004) *Data-rich reporting and standards setting following rules-based marking* **in** Ashby, M. (ed) Eighth International Computer Assisted Assessment (CAA) Conference Proceedings, Loughborough University, 6th and 7th July 2004. Online, available: http://www.caaconference.com/.

Qualifications and Curriculum Authority (QCA) (2005a) *What we do*. Online, available: http://www.qca.org.uk/3657.html.

Qualifications and Curriculum Authority (QCA) (2005b) *Interim evaluation of the 2005 pilot of the key stage 3 ICT tests*. Online, available: http://www.qca.org.uk/7254.html.

Qualifications and Curriculum Authority (QCA) (2005c) *Final evaluation of the 2005 pilot of the key stage 3 ICT tests*. Online, available: http://www.qca.org.uk/7254.html.

Research Machines (RM) (2005a) *Test Specification: Measurement Methodology.* Unpublished project specification document.

Research Machines (RM) (2005b) *Test Specification: Test Development & Awarding Specification.* Unpublished project specification document.

Research Machines (RM) (2005c) *Validity and Reliability of the 2005 Pilot Test*. Unpublished project report.

Wiliam, D. (2000) *Validity, Reliability, and all that Jazz.* Education 3-13, Vol. 29, No. 3. (2000), pp. 9-13.

All web references were live at 13th March 2006.