

Regulatory research into on-demand testing



Andrew Boyle
February 2010
Product code: Ofqual/10/4725

Contents

Executive summary	3
Introduction.....	4
Method	5
Review	5
Focus groups	5
Findings.....	7
Review	7
Meaning of the term ‘on-demand testing’.....	7
Use of on-demand testing.....	8
Arguments in favour and against on-demand testing.....	10
Previous regulatory actions.....	13
Proposals for future regulatory action	15
Focus groups	18
On-demand testing as a driver of e-assessment uptake.....	18
In favour or against increased use of on-demand.....	19
What GCSEs and A levels could learn from other areas	19
Regulatory approaches to facilitate on-demand testing	20
Key challenges in doing on-demand testing	20
Discussion and next steps.....	22
Appendix 1 References	25
Appendix 2 Focus groups discussion guide	27

Executive summary

This document reports regulatory research into on-demand testing. It bases its findings on two sources of evidence: prior reports commissioned or written by qualifications regulators and focus groups that addressed the topic in November 2009.

The review of prior research divides findings into the following categories:

- meaning of the term 'on-demand testing'
- use of on-demand testing
- arguments in favour and against on-demand testing
- previous regulatory actions
- proposals for future regulatory action.

Focus group findings were based on five questions which were developed in the light of prior research.

The concluding discussion section notes qualifications regulators' longstanding interest in e-assessment and, in particular, the prominence that on-demand testing has played as an element of e-assessment work. It suggests that 'higher level' regulatory documents are likely to be part of hybrid regulatory systems and that the strengths and weaknesses of such systems need to be clearly understood.

The discussion section concludes by suggesting further actions. It suggests that principles expressed in a 2009 AQA report form a good basis for regulatory principles to facilitate on-demand testing, but also states that any revision of the regulatory principles for e-assessment should be comprehensive, rather than merely to address on-demand testing issues. The concluding section also proposes that regulators should provide a technical document showing how standards can be maintained in an on-demand world, and that they need to be in touch with a range of agencies to address issues of assessment burden and pedagogy.

Introduction

In 2009 researchers from AQA delivered a report on on-demand testing (Wheadon *et al*, 2009) written to the specification of the qualifications regulators – Ofqual (England), DCELLS (Wales) and CCEA (Northern Ireland). A follow-up event was held in November 2009 to discuss on-demand testing, attended by a diverse group of stakeholders who listened to presentations about on-demand testing, took part in plenary discussions and smaller focus group sessions.

This document summarises comments made by participants in the focus groups. However, four focus groups would not by themselves provide a set of data from which to draw robust findings. Therefore, the report of remarks made in the focus groups is supported by a review of research conducted by or on behalf of qualifications regulators which pertain to on-demand testing. The aim of this is to give context to the topic of on-demand testing (for instance to show why regulators are particularly interested in on-demand testing as an element of their work on e-assessment). Also, a secondary purpose is to reference a range of reports that have been authored or commissioned by regulators – including published and unpublished sources.

Method

Review

The review takes as its universe of possible papers research or policy reports written for or on behalf of UK qualifications regulators that refer to on-demand testing as facilitated by e-assessment. This provides a set of papers sufficient in size, diversity and quality to derive findings about on-demand testing. Of course, there are many other research papers that reference on-demand testing other than those written by regulators. However, this body of papers is suitable to support the aims of the current research.

The papers are reviewed instrumentally and synthetically. Key issues are chosen and illustrations of these issues are drawn from specific papers. This review is not an exhaustive or formally organised listing of all papers produced by regulators on the topic of on-demand testing. Also, the topics highlighted in the review link to topics noted in analysis of focus group discussions; the purpose being to contextualise the things that were said in those groups.

Focus groups

The on-demand testing event was attended by over 30 delegates. These invitees came from across the spectrum of stakeholders in e-assessment. They included representatives of: qualifications regulators from across the UK, general and vocational awarding organisations, the Department of Children, Schools and Families, Becta, the English Secondary Schools Students Association, as well as representatives of school staff such as exams officers and network managers. Many of the invited participants were experts in e-assessment, for example, e-assessment managers and researchers from awarding organisations. As such, it was reasonable to characterise the research participants as an informed group of people.

One of the aims of the event was to gauge participants' opinions on on-demand testing. A detailed, yet flexible, discussion guide was developed in the light of prior regulatory research (it is Appendix 2 below), and peer reviewed by a qualitative researcher before use. The guide consisted of five core questions, with optional follow-up questions and prompts for each. The five core questions were:

1. Previous research participants have told us that on-demand testing will be an important driver of e-assessment uptake.

Do you agree?

2. Broadly speaking, how would you feel about increased use of on-demand testing in qualifications?

Would you be in favour or against?

3. What could on-demand testing in GCSEs and A levels learn from on-demand testing in other areas (e.g. vocational, professional or licensure qualifications)?
4. The regulators are proposing to develop a regulatory approach to facilitate on-demand testing.

What is your view of this?

5. What are the key challenges (difficulties) in doing on-demand testing?

Discussion facilitators were asked to take a flexible approach to the guide's implementation.

Four focus groups were constituted. Each one was facilitated by a member of staff from a qualifications regulator and bullet point notes were taken of each group's discussions. These notes were checked by each facilitator before being submitted for analysis.

Analysis methods were chosen so as to provide the most robust set of findings possible. As with all qualitative research, the findings from these focus groups give an indication of the range of opinions that exist on particular topics, rather than a quantitative summary of their extent across a representative sample of respondents. This is a general feature of qualitative research, which is nonetheless widely considered to be a valid research method.

Focus group methodologists have stated that 10–12 focus groups should be run for 'saturation point' to be reached; that is, in order for there to be no more substantial issues that would be expressed in response to a set of questions by further groups. Clearly, with four focus groups, this research data set is less than optimal. In mitigation of this concern, it may be noted that group members could be considered as experts with well-developed opinions and their views are therefore worthwhile on that basis. Further, the findings from the focus groups have been closely cross-referenced with findings from the review. The findings section below points out wherever focus group comments support or contradict those summarised in the review. Finally, there were several findings in the focus groups that appeared to contain internal inconsistencies. The nature of such inconsistencies is pointed out, to alert the reader to an issue that requires further clarification.

So, subject to the caveats set out here, it is believed that the data set gathered is sufficient in quantity and quality to sustain the findings reported below.

Findings

Review

Meaning of the term ‘on-demand testing’

On-demand testing could also be referred to as ‘when-ready testing’. It is not a new concept – for many years candidates have been able to sit a graded music examination or the practical driving test at a time of their choosing. These two examples are ‘performance assessments’, not assessments in which large quantities of new content need to be written for each version of a test. Performance assessments are perhaps the most common form of on-demand testing that was carried out before the advent of viable e-assessment. In contrast, on-demand testing facilitated by e-assessment would typically involve many versions of test papers being generated from electronic databases of questions (or ‘item banks’).

On-demand testing is a concept with a range of meaning. The purest form of on-demand testing could be caricatured as the ‘Martini model’ – ‘any time, any place, anywhere’. But more moderate versions of on-demand testing could be seen as a relatively slight leap from current practices of modular assessment for GCSE and A level. An important distinction within the concept of on-demand testing is between a single multi-day test window (e.g. where there are ten test sessions over a five-day week) or multiple single-day sessions (e.g. 12 sessions per year – students being able to sit a test on the 15th day of every month). A case may be made in favour of either of these approaches.

Wheadon *et al* (2009, p. 9) helpfully discern four ‘flavours’ of on-demand testing:

A Unique to candidate – any time

Unique tests are provided for each candidate. No test is used more than once. Tests can be taken at any time on any day suitable for the candidate and/or centre. Other than for very low-volume subjects, this is likely to require the generation of tests automatically from very large item banks.

B Unique to session – many sessions

Unique tests are provided for a large number of sessions. There may be one unique test per session taken by all candidates in all centres, or a number of tests which are used only for a specified session and are taken by specified sub-samples of candidates/centres in order to pre-test items and establish grade boundaries. Sessions may be grouped to form windows of assessment of one or more days, at intervals during the year. There may be enough tests to allow multiple sessions in a single day for most days of the year. The capability to provide a large number of sessions is likely to require the generation of tests automatically from large item banks.

C Unique to session – few sessions

Unique tests are provided for a small number of sessions in any single academic year. There may be one unique test per session taken by all candidates in all centres, or a number of tests which are used only for a specified session and are taken by specified sub-samples of candidates/centres to pre-test items and establish grade boundaries. The dates and times of the test sessions are fixed by the awarding body. Because of the small number of tests, they could be generated manually or semi-automatically from smaller item banks. This is the current model used for the AQA GCSE science tests, which at present has three test series a year. It was also the model for key skills tests offered six times a year.

D Re-usable – centre selected dates

A bank of re-usable tests is created when the specification is first taught. Centres request a test to administer on a date chosen and specified by the centre. The awarding body provides a test not taken recently at the centre or neighbouring centres and/or not taken by the candidates. This model is currently used for some Entry level qualifications.

Use of on-demand testing

Basic and key skills qualifications will be accredited until at least August 2010, and will certify attainment until August 2012 (QCA, 2007a; see also QCA, 2004a). They deliver on-demand assessment for the level 1 and 2 tests of:

- application of number/adult numeracy
- communication/adult literacy
- ICT (QCDA, undated).

QCA piloted a national test in information and communication technology at key stage 3. The test was available across a four-week test window. In 2006, over 2,000 schools took part in the pilot (Boyle, 2006, p. 62).

Wheadon *et al* (2009, p. 41) summarise use of on-demand testing in the following table.

Regulatory research into on-demand testing

	Subject to Ofqual regulation	Supplies items	Supplies on-demand tests	Stakes of on-demand tests	Purpose of assessment	Number of on-demand tests available	Number of test windows per year	Period of notice required before taking test
Universities Medical Assessment Partnership (UMAP)	×	✓	×	High	Progression through medical school	Supplies items rather than tests	Supplies items rather than tests	Supplies items rather than tests
City & Guilds	✓	×	✓	High	Professional & vocational qualifications	900 approx.	Continuous	None to six months
Scottish Qualifications Authority (SQA)	×	✓	✓	Medium	Selection to higher education & employment	800 approx.	Continuous	Not applicable
IFS School of Finance	✓	×	✓	High/Low	Licence to practise within industry sector/life skills	40–50	Continuous	One day minimum
CEM Centre	×	×	✓	Low	Formative & diagnostic assessment in primary & secondary schools	8	Variable: 1 per year of 7 months to continuous	None to one month
EDI	✓	×	✓	Medium	Diagnostic & summative	350 approx	Continuous	Four weeks
Driving Standards Agency	×	×	✓	Low	Licence to practise a life skill	11	Continuous	Ten-day minimum
Organisation I	×	×	✓	High	Licence to practise within industry sector	2	Test 1: six Test 2: 1 of 30 days	Test 1: One week Test 2: Four weeks
Organisation II	✓	×	✓	Medium/High	Licence to practise within industry sector	20	Five days per week, except two shutdown periods	Seven days to three months

Note: The last two organisations listed in this table did not wish to be acknowledged.

Arguments in favour and against on-demand testing

A number of arguments have been advanced in favour of and against on-demand testing. Arguments in favour have been derived from educational and assessment principle, but also from practical considerations. While on-demand testing is sometimes seen as an important driver of e-assessment, it is one among several. As well as arguments in favour of on-demand testing, there are a number of arguments against its use, or at least comments that emphasise its limitations. It may be that arguments for and against tend to come from different sources: those for tend to come from providers of assessment and those against (or reservations) from consumers.

On-demand testing has been characterised as an important contributor to desirable outcomes in the future of education in several influential reports. Gilbert's report presented a vision of the education system in 2020 and a key driver of her vision of personalised learning was frequent, targeted assessment using on-demand approaches (Wheadon *et al*, 2009, p. 8). In his report advocating a broadly based diploma as the qualification of choice for England, Tomlinson proposed the following:

- use on-demand assessments tailored to the individual:
 - as a tool to allow less motivated pupils to 'bank' credits as they go along
 - to moderate teachers' assessments. (Boyle, 2005, p. 8)

In addition to on-demand testing as an important matter of principle, it has also been alleged to be a practical facilitator to the uptake of e-assessment. It has already been noted how QCA ran the trials of the key stage 3 ICT test using a four-week test window to allow schools to schedule whole-cohort test sessions on limited numbers of computers. Also, awarding body staff taking part in focus groups in 2006 suggested that on-demand testing was an important way in which the 'logjam' of e-assessment could be broken (Boyle, 2007, p. 20).

Although on-demand testing has been thought of as a driver or facilitator of e-assessment uptake, it is not the only such driver. In 2004, the consultancy 'Exam on Demand (ExoD) and the Assessment Advisory Group' suggested that the following factors were important to the development of e-assessment:

- agreement of the on-screen question types that could be used to assess individual GCSE and AS/A-level syllabuses, plus vocational qualifications
- regular reviews of the regulatory structures to maintain relevance and flexibility
- supply of good quality e-learning materials to support personalised learning

- development of technical, physical and human resources by schools and colleges to manage e-assessment
- schools and colleges developing the organisational structures to support personalised learning and on-demand assessment (Boyle, 2007, p. 12).

Focus group respondents reported in Boyle (2007, p. 19) suggested that the following were drivers and inhibitors¹ of e-assessment:

¹ On drivers and inhibitors of e-assessment see also: Thomson (2005 and 2006).

Drivers
<p>(Ts) Young people are engaged by on-screen assessment.</p> <p>(Ts) Schools should follow industry in making full use of ICTs:</p> <p style="padding-left: 40px;"><i>... if you look in industry at the way that process happens, this is very much now data driven. And why shouldn't we be keeping up with using technology to simplify the process, reduce workload and automate things as much as we can?</i></p> <p>(Ts) Back-up and support of senior leadership teams.</p> <p>(Ts) e-assessment must have a genuine benefit – for example in permitting simplification of processes or the assessment of novel constructs.</p> <p>(Ts) When-ready testing is particularly suitable for 16-to-18-year olds.</p> <p>(Ts) Having a portal that teachers can access from home computers – e.g. for marking, etc.</p> <p>(ABs) Test centres' technology being upgraded following government initiatives.</p> <p>(ABs) Awarding bodies' commercial interests to identify products to plug gaps in the market.</p> <p>(ABs) Changes in QCA's codes of practice to facilitate on-demand assessment.</p>
Inhibitors
<p>(Ts) Lack of access to technology by teachers and students due to low numbers of computers and high student and/or teacher-to computer ratios in schools.</p> <p>(Ts) and (ABs) Exams being required to run only in fixed sessions.</p> <p>(EOs) Difficulties in booking rooms to run test sessions.</p> <p>(ABs) The requirement to demonstrate that on-demand tests are comparable.</p> <p>(ABs) Separate <i>principles</i> for e-assessment representing an extra hurdle not required for on-paper assessment.</p> <p>(ABs) Public perception that e-assessment is 'dumbed-down' assessment.</p>

Legend: Ts - teachers; ABs - Awarding bodies; EOs - exams officers

Drivers and inhibitors that relate to on-demand testing are in bold.

Similarly, Hedra (2008) recommended on-demand testing as an action for Ofqual, but as one among the following four:

- clarify position on e-assessment and expand on collaborative projects
- state position on the implications of e-assessment components e.g. use of item banks, on-demand assessments

- consider the possibility of making a statement requesting awarding bodies that at least one GCSE subject is delivered onscreen
- state position on technology platform.

Senior educationalists setting out visions of the future of education have seen the potential of on-demand testing to support profound change. Professionals in the assessment provision industry perceive that on-demand testing could help to overcome some of the practical difficulties that inhibit e-assessment take up in present-day schools and colleges. However, there is also a strand of opinion that expresses scepticism towards on-demand testing. Some of this opinion comes from industry professionals, but rather more of it comes from users of tests – university students, teachers and examinations officers.

It is acknowledged from several sources that guaranteeing the measurement properties of on-demand tests is difficult; awarding body informants spoke about the difficulties of demonstrating the validity, reliability and comparability of on-demand testing (Boyle, 2007, p. 21). University students and teachers in focus groups were concerned about the comparability and integrity of results from on-demand tests. In particular, there were concerns that cheating would be easier if questions were used in multiple versions of tests (Wheadon *et al*, 2009, p. 29). An exams officer felt that the advent of on-demand testing in further education had led to tutors and students taking examinations less seriously. She said:

I think it can impact on them. The tutor knows that the 'on demand' can be scheduled 24/48 hours, they might actually be inclined to say to their group, 'okay, it's Wednesday, got a test on Friday'. Now I don't think that's fair to the candidate, not to give them very much notice to prepare. And also, the other thing I think is a bit of a shame, the online exams seem to have been (I don't know) cheapened a little bit, by the students and the tutors. ... they don't have the same regard that they would if it was a paper-based exam, that we had to order two months ago, and it was all very formal. It was going to be in the hall. This is something, we can have it in two days, so it's not really worth very much. And I get that impression from tutors and the students. (Boyle, 2007, p. 21)

Previous regulatory actions

QCA (2004a) describes the experience starting from 2002 of running on-demand testing with several awarding organisation partners for basic and key skills (BKS) qualifications. That report flags, among other things, the need for consistent delivery systems for on-screen testing, the variety of issues and tensions between awarding organisations and centres when dealing with new approaches to assessment and a set of proposed criteria to facilitate access to a central bank of BKS papers.

In 2004, the Chief Executive of the QCA issued a challenge to all those involved in assessment provision. He expressed a bullish vision of the future of assessment, including the following invocation:

... let's think about some clear goals and performance indicators. Why don't we, as the QCA, the awarding bodies and the learning providers set ourselves some objectives such as the following?

Within five years, say:

- all new qualifications, whether for schools or for adults in the workplace, will include an option for on-screen assessment
- all awarding bodies should be set up to accept and assess e-portfolios
- most GCSEs, AS and A2 examinations should be available optionally on-screen
- national curriculum tests should be available on-screen to those schools that want to use them
- on-demand assessments will begin to be a feature of GCSE. (Boston, 2004, p. 2)

One of the UK qualifications regulators' *regulatory principles for e-assessment* concerned on-demand testing, as follows:

9. Automatically generated on-demand tests

Awarding bodies must ensure that there is a sufficient volume of assessment items or questions to provide consistently secure, robust, balanced and unique on-demand tests, appropriate to the form of assessment.

9.1. Where awarding bodies use electronic assessment item banks to automatically generate on-demand tests they must ensure, by thorough testing, that there are sufficient assessment items to provide consistently robust, balanced and unique test papers for the assessment/test windows to be accommodated.

9.2. Where electronic assessment item banks are used to automatically generate individual on-demand tests, awarding bodies must make sure that the security of assessment items is not compromised by the level of use by ensuring that there are sufficient items available to accommodate the test window and candidate capacity.

9.3. Where electronic assessment items or question banks are used, awarding bodies must ensure that each item that contributes to tests is consistent and comparable with others over time for each session.

9.4. Where delivery of test items or questions is randomised, must have policies and procedures in place to analyse the possible impact of the

randomisation on candidates' performance and to ensure that question order does not bias results, for example by pre-testing.

9.5. Automatically generated on-demand tests must be appropriately designed to allow for equal choice for disabled learners. (QCA *et al*, 2007, p. 13)

In the development of the *regulatory principles*, the regulators had moved from conceiving of 'electronic item banks' to 'on-demand testing' (QCA, 2007b, para 9). That rewording represented a change of focus – from an internal test development perspective to a more 'public-facing' focus in which a wider range of issues affected test takers.

Proposals for future regulatory action

In one of its earliest investigations of e-assessment, QCA (2004b) recognised that the advent of on-demand testing would necessitate profound changes to codes of practice. In reviewing a contemporary code, QCA stated:

This section illustrates the fundamental reliance on experts in setting content and grades of exams. In an on-demand-instant feedback (or even on-demand only) e-assessment environment, even if parallel forms are used, the reliance has to move toward (*sic*) more use of statistical data. This is a massive shift, and changes almost everything in this section.

Wheadon *et al* proposed a set of 12 principles as a starting point for regulation of on-demand testing (2009, pp. 5–6 and pp. 50–51). The principles are organised under four headings:

- examination standards
- accessibility
- the burden of assessment
- communication.

It is interesting to compare and contrast proposed principles of Wheadon *et al* with Boyle's (2005) recommendations concerning the implications of the Tomlinson report for on-demand testing.

The table that follows shows recommendations from the two reports that address similar issues.² The recommendations have much in common; particularly in the

² The original numbering is retained in both sets of recommendations, giving the reader an impression of the extent of agreement between the two sets of recommendations.

areas of examinations standards and burden. The earlier recommendations do not address accessibility (perhaps reflecting that issue's increased salience in public policy and legislation in recent years³) and have less explicit focus on communications. The recommendations from Wheadon *et al* are expressed at a higher level of generality; while those from Boyle tend to focus on details of test development. Boyle's recommendations also reflect a certain scepticism to statements in the Tomlinson report that on-demand testing could or should be associated with a breaking down of the 'barriers' between summative and formative assessment.

It is worth noting two proposals from Boyle's paper that are not present in the later recommendations:

- insights from adaptive testing algorithm design should be adapted to inform the development of on-demand testing item banks (recommendation 11)
- the cost of on-demand testing should be monitored (recommendation 15).

³ Including the extension of the Disability Discrimination Act to cover general qualifications from September 2007.

Wheadon et al 2009 draft principles	Boyle 2005 recommendations
<p>ix. The impact of introducing on-demand testing on the education system as a whole from first teaching to entries through to results should be modelled from end-to-end.</p> <p>x. Changes in the burden of assessment in the educational system as a whole, including additional pressures on teachers and candidates, should be monitored.</p> <p>xi. All stakeholders, including candidates, should be actively consulted during the redefinition of processes to support on-demand testing.</p>	<p>4. A committee should be appointed to investigate logistical costs of moving to on demand.</p>
<p>ii. On-demand testing should be underpinned by Item Response Theory methods of test-equating.</p>	<p>5. Item banking should be used to facilitate on-demand testing, and this item banking should be based on Item Response Theory.</p>
<p>iii. Policies on item to test ratio, item re-use, pre-test procedures and evidence of coherence of scales should all be available.</p> <p>xii. Teachers and candidates should be informed exactly how items are pre-tested, how they are likely to be re-used, and how test versions will be equated.</p>	<p>6. An on-demand testing programme should be preceded, and supported on an ongoing basis by the pre-testing of items.</p>
<p>ix. The impact of introducing on-demand testing on the education system as a whole from first teaching to entries through to results should be modelled from end-to-end.</p> <p>x. Changes in the burden of assessment in the educational system as a whole, including additional pressures on teachers and candidates, should be monitored.</p>	<p>7. The use of pre-testing should be investigated by a body charged with prioritising data collections in education, such as the 'Star Chamber'.</p>
<p>iv. Where items are re-used, item parameters should be monitored for unexpected changes over time or between versions that may indicate security breaches, drift, over-use or changes in testing conditions such as reduced time available for question completion.</p> <p>xii. Teachers and candidates should be informed exactly how items are pre-tested, how they are likely to be re-used, and how test versions will be equated.</p>	<p>9. Tomlinson-reformed on-demand testing should follow the practice of current reputable testing organisations and produce tests that have shared items between versions.</p>
<p>vi. The reliability of tests should be such that there is little to gain from repeated re-sitting.</p>	<p>13. Where on-demand testing involves schools administering tests, the availability of tests should be limited; so that the legitimate needs of students to have more convenient access to the assessment system, and the unavoidable practical constraints on schools can be fairly balanced.</p>
<p>vi. The reliability of tests should be such that there is little to gain from repeated re-sitting.</p>	<p>14. Despite the explicit statement in the Tomlinson report to the contrary, the frequency with which students are able to re-sit a qualification should be subject to a reasonable limit, but the total number of permitted re-sits should not be subject to any limit.</p>

Focus groups

The following section reports findings from the focus groups. The findings are organised according to the core questions listed above. The findings should be interpreted in the light of the caveats and assumptions set out in the method section above.

On-demand testing as a driver of e-assessment uptake

The focus group discussions gave reasonable support to the suggestion reported in the review above that on-demand testing could be one driver of e-assessment uptake, although not necessarily the only or main one.

There were a number of reasons why on-demand testing was considered likely to be a driver of e-assessment. Several speakers suggested that on-demand testing could have 'motivational benefits'. They pointed out that the relatively recent innovation of modular testing in general qualifications was popular with students. They thought that on-demand testing would be similarly popular. This suggestion chimes with the observation in the review that on-demand testing was not necessarily a radical step-change in assessment practices, but rather could be conceived of as an evolution from recent trends.

Several speakers took the view that insofar as on-demand testing provided 'accurate and relevant quantitative feedback' to students, then it could be a facilitator of e-assessment. However, others cautioned that, sometimes, the learner voice in schools could be 'passive' and that e-assessment was something that was 'done to learners' rather than with them. This suggested that students' positive disposition towards on-demand testing might not necessarily lead to its rapid implementation.

The data on on-demand testing as a solution to the logistical challenges of e-assessment were somewhat inconsistent. Some speakers emphasised the difficulty of doing whole cohort testing using e-assessment. It followed therefore that on-demand testing could be the way forward for e-assessment. Further, it was felt likely that computer provision in schools (through initiatives such as Building Schools for the Future) was moving away from traditional computer suites. This implied more flexible IT resources and therefore a more flexible approach to testing. However, on-demand testing would cause logistical problems as well. Exams officers would be burdened by having to schedule very frequent test sessions, for example.

Some drivers of e-assessment other than on-demand testing were noted. These included there being a clear vision for the curriculum and assessment systems in the twenty-first century, and the use of e-learning increasing in quantity and quality.

In favour or against increased use of on-demand

The research participants tended to be in favour of the use of on-demand testing. This was unsurprising as many of them had a prior interest in e-assessment. However, that positive response was rather lukewarm. Participants had some reservations about on-demand testing as well.

When speakers were expressing positive attitudes towards on-demand testing, they mentioned its flexibility and repeated the comments summarised in the previous section about the popularity of modular GCSEs.

Some other comments reflected speakers' scepticism towards on-demand testing or at least a weariness as to whether conditions in schools were conducive to realising the perceived benefits of on-demand testing. Speakers emphasised the natural rhythms of the school year, and were not sure that schools would have developed strategies for occupying students if they had completed a test early in the year. One speaker said, 'personalisation has far to go.' Some commentators reported concerns around standards. Some such concerns were genuine – a speaker was worried that candidates sitting tests repeatedly would lower the reliability of outcomes (cf. Wheadon *et al's* proposed regulatory principle vi.). However, other concerns surrounded perceived, although not necessarily genuine, threats to standards. It was suggested that some observers would be cynical and consider any increase in students' ease and comfort in dealing with the testing system as a drop in standards.

What GCSEs and A levels could learn from other areas

Focus group members acknowledged that on-demand e-assessment was further down the line in areas such as vocational and professional qualifications (VQs & PQs) than in general qualifications (GQs). It was said that on-demand testing was associated with lower drop-out rates in vocational qualifications and Skills for Life; in these contexts, on-demand testing was reported to contribute positively by providing shorter milestones, and more frequent feedback and affirmation for learners.

Successful implementations in vocational and professional qualification contexts suggested that on-demand testing could port across successfully to the general qualification arena. In particular, the use of on-demand testing for professional qualifications – both in the UK and the United States – showed that on-demand testing could operate in a high-stakes environment.

As in previous questions, respondents expressed some reservations. In particular, speakers alleged that vocational qualifications made heavy use of multiple-choice questions, and that such questions could be viewed with suspicion by users of general qualifications.

Regulatory approaches to facilitate on-demand testing

Delegates spoke about regulatory arrangements with some passion. There was general support for an approach of high-level principles with detail provided by non-mandatory best practice guidelines and illustrative case studies. This view was broadly supportive of current documents and antagonistic to any proposals to develop detailed regulations to facilitate on-demand testing.

There were two findings that challenged this general acceptance of the status quo. Firstly, several focus groups followed the commentator cited above (QCA, 2004b – p. 15 in this document) in stating that on-demand testing would require radical changes to regulated measurement procedures, such as standards setting and maintenance, and the use of techniques such as pre-set pass marks. Further, such changes to the psychometric bases of awarding would be likely to lean heavily on approaches from IRT. If the current high-level, non-specific regulatory principles remained the sole regulatory documents governing e-assessment, then it was possible that stasis could set in and desirable developments could fail to take root.

Furthermore, some speakers noted potential weaknesses of a hybrid system. If e-assessment regulatory documents remained 'high level' and non-specific, it was possible that decisions could be taken elsewhere – in the development of other regulatory arrangements such as codes of practice – which would affect awarding organisations' ability to deliver e-assessment. In such a scenario, regulation could be piecemeal and e-assessment issues would not be given sufficient prominence.

The final comment that some group members expressed related to communication. They felt that communication from regulators had not always been great. This had reinforced tendencies referred to in the previous paragraph for e-assessment issues to be dealt with at a remove from the main regulatory process. This gave concern to those commenting that non-optimal solutions to e-assessment issues had sometimes been arrived at.

Key challenges in doing on-demand testing

Several of the group had enumerated the majority of the key challenges in response to previous questions. Challenges that groups felt had not been mentioned earlier or which they thought were sufficiently prominent to merit re-stating included:

- Lead time is an issue. The process of accrediting qualifications meant that specifications could be 'locked in' for up to five years, and this could inhibit innovation.
- Confidence was a key predictor of uptake. Confidence was needed from centres, awarding organisations and candidates.

- It was important for awarding organisations to provide sufficient and appropriate supporting materials to centres. Schools and colleges liked to see past papers, or databases of test questions for candidates to practise on. If items were being reused in on-demand tests, then alternative sources would need to be found for such practice materials.

The final comment to note in this section came from an awarding organisation group member. He said that moving to a radically different system, while maintaining confidence in current arrangements, was like trying to 'service a plane in mid-flight'.

Discussion and next steps

UK qualifications regulators have taken a keen interest in e-assessment for some time. At times they have been very bullish and emphasised the potential for rapid progress (Boston's 2004 speech is probably the high-water mark for this phenomenon). But at other times, regulators' approaches have been alleged to have changed; they have been said to have 'soft-pedalled' on the subject.

Regulatory organisations and the bases upon which they work have been subject to substantial change in recent years. Regulatory functions have been separated from curriculum and qualification development, devolved to separate organisations and duties and powers have been defined by new legislation. As such, it is reasonable at the very least that regulators should pause and consider whether and how their prior approaches to e-assessment could be affected by new statutory powers and duties.

Perhaps a more interesting avenue – rather than discussing the perceived institutional orientations of regulators – is to consider why on-demand testing has been such a prominent focus for regulators. For instance, the aversion of UK assessment stakeholders to automated marking is well-documented. Wouldn't it have been useful for regulators to have taken a lead and demanded that awarding organisations conduct research to improve the automated marking of essays and short-answer questions, and so to make e-assessment more in line with UK stakeholders' perceptions of validity in assessment?

The answer to this question has several strands: the first is that when regulators ask, stakeholders reply that on-demand testing is important. The awarding body officers talking in 2006 (reported in Boyle, 2007) said that on-demand testing could 'unblock the logjam' of e-assessment. Unitary awarding body Chief Executives speaking to Hedra (2008) cited it as an important issue. Not everyone is in favour of on-demand testing; some focus group respondents in the current research did not believe that it was suitable for schools as they are currently constituted. However, it is a matter that keeps coming up whenever regulators ask about e-assessment.

Also, on-demand testing is properly a regulatory matter and not one of test development. Issues such as the development of high quality assessment instruments and marking approaches (including electronic marking of human language) are test development concerns. As such, they are an area in which awarding organisations would be best equipped to take a lead. In contrast, much of the substance of the on-demand testing issue concerns standards and comparability, as well as national arrangements for examinations (timetables, etc.) – core foci of qualifications regulators.

It is important to note how the understanding of the concept has evolved in recent years. The shift of terminology from 'item banking' to 'on-demand testing' is

instructive, the former implies a focus on internal awarding organisation processes, the latter is broader and is made up of a combination of complex and difficult-to-solve standards issues, public perceptions and corresponding communications imperatives as well as real concerns about the practical burden on centres. Any approach to on-demand testing would need to balance such strands.

A further change that can be noted in the concept of on-demand testing concerns the way that it needs to be regulated. When QCA reviewed its then current code of practice (QCA, 2004b), it found that assumptions arising from traditional paper-and-pencil approaches to assessment were 'hard coded' into the regulations. What was needed was a regulatory document governing e-assessment that worked at a higher level of generality and which did not set in stone a particular technology or way of working. The response to this has been the *regulatory principles for e-assessment*, a set of regulatory requirements that mandates desired outcomes rather than concreting awarding organisations into a single time-bound model of test development.

The *regulatory principles for e-assessment* remain a relatively new departure. It is incumbent on regulators and the regulated to continue to evaluate their functioning and to consider the implications of this new kind of document. In particular, neither regulators nor the regulated should see 'higher-level regulatory instruments' as a 'soft touch' or as simply a way of avoiding controversy. Although it is desirable that e-assessment regulation does not bind awarding organisations to a particular implementation of technology, the use of non-specific instruments will bring its own problems. Specificity and technical detail are intrinsic elements both of regulation and of qualifications development and awarding. To pretend otherwise is to perpetuate a myth. If e-assessment principles must necessarily be non-detailed, it just means that detail is exported elsewhere. And this implies that e-assessment regulation (and perhaps all regulation in future) will be a hybrid system. This brings its own problems; it may mean that e-assessment is a minor concern in the regulation of new qualifications, when – looking to the future of our education systems – it ought to be highly prominent. It may also risk e-assessment development proceeding in an ad hoc and arbitrary manner, with developments in different suites of qualifications not being sufficiently co-ordinated. Such problems are avoidable if regulators and the regulated can see the bigger picture, but they are the type of issues that may well occur in a hybrid regulatory system.

Next steps can be envisaged. Some matters are relatively straightforward to implement, others are more diffuse and not wholly under regulators' control. Regulators ought to be able to design well-written regulatory documents. Wheadon *et al* (2009) have provided a useful set of principles for regulating on-demand testing. These may need to be supplemented for practical implementation (for instance, Boyle, 2005, may be the source of useful recommendations), but they are a good starting point. Insofar as it is desirable to amend the *regulatory principles for e-*

assessment, it is probably preferable that the revision is comprehensive rather than to address the issue of on-demand testing in isolation. Similarly, it should be possible for regulators to synthesise findings from Wheadon *et al* (2009) and provide a more detailed model of steps that would be necessary to maintain standards in an on-demand world.

More abstruse factors would require collaboration with outside agencies and probably a longer timeframe. For example, concerns that on-demand testing would impose unreasonable burdens on schools and colleges would need to be tracked – for instance with bodies such as the ‘Star Chamber’. Also, pedagogic matters, such as the maturity of personalised learning in schools and colleges, would need to be discussed with relevant bodies such as Ofsted and Becta. Account should also be taken of e-assessment’s potential to provide assessment materials in two languages in bilingual nations such as Wales.

Many of the matters outlined above are difficult to solve and may give rise to controversy. However, part of the role of regulators is provide leadership on tricky matters of public policy. It is suggested that regulators can continue to show leadership in the sphere of on-demand testing.

Appendix 1 References

Boston, K. (2004) *Speech to e-assessment conference*, Tuesday, 20 April 2004, The Royal Festival Hall. Available online:

http://www.qcda.gov.uk/libraryAssets/media/6998_kb_speech_20040420_e-assessment.rtf.

Boyle, A. (2005) *On-demand testing: implications for the Tomlinson reforms – a discussion paper for the 14 – 19 Strategic Planning Group*. Unpublished QCA report.

Boyle, A. (2006) *Evaluation of the 2006 pilot of the key stage 3 ICT test*. Available online at:

http://testsandexams.qcda.gov.uk/libraryAssets/media/2006_pilot_evaluation_final_version.pdf.

Boyle, A. (2007) *Initial research into the regulation of e-assessment: internal draft for discussion*. Unpublished QCA report.

Hedra (2008) *Encouraging technology related innovation in the general qualifications market?* Available online: http://www.ofqual.gov.uk/files/QCA_E-regulation_Report_Final_V3.pdf.

Qualifications and Curriculum Authority (QCA) (2004a) *The basic and key skills (BKS) e-assessment experience report*. Available online: <http://www.ofqual.gov.uk/1569.aspx>.

Qualifications and Curriculum Authority (QCA) (2004b) *Extract from GCSE, GCSE in vocational subjects, GCE, VCE, GNVQ and AEA code of practice 2004/5*. Unpublished QCA report.

Qualifications and Curriculum Authority (QCA) (2007a) *Functional skills: essential for life, learning and work*. Available online: http://www.ofqual.gov.uk/files/qca-07-3170_functional_skills.pdf.

Qualifications and Curriculum Authority (QCA) (2007b) *Consultation report on the regulatory principles for e-assessment: 16 January 2007 – 2 March 2007*. Unpublished QCA report.

Qualifications and Curriculum Authority (QCA), Department for Children, Education, Lifelong Learning and Skills (DCELLS), Council for the Curriculum, Examinations and Assessments (CCEA) & Scottish Qualifications Authority (SQA). 2007. *Regulatory principles for e-assessment*. Available online at: http://www.ofqual.gov.uk/files/Final_regulatory_principles_document_-_PRINTED.pdf.

Qualifications and Curriculum Development Agency (QCDA) (undated) *On-demand assessment*. Available online: <http://www.qcda.gov.uk/4539.aspx>.

Thomson (2005) *Drivers and barriers to the adoption of e-Assessment for UK Awarding Bodies*. Available online: http://www.qcda.gov.uk/libraryAssets/media/pdf-05-1671_Barriersanddriverstoeassessment.pdf.

Thomson (2006) *Acceptance and usage of e-assessment for UK awarding bodies*, Available online: <http://tinyurl.com/yl2jfu>.

Wheadon, C., Whitehouse, C., Spalding, V., Tremain, K. & Charman, M. (2009) *Principles and practice of on-demand testing*. Available online: www.ofqual.gov.uk/files/2009-01-principles-practice-on-demand-testing.pdf.

All web links were active on 24 February 2010.

Appendix 2 Focus groups discussion guide

Introduction

Welcome to the afternoon focus groups. My name is ... , I'm the facilitator of this session. Would you please briefly introduce yourselves. ...

We have five groups of questions. That means each one should last about 15 minutes – in chairing this session I'll try to let the conversation flow and to be spontaneous as far as possible; I won't necessarily stick to the letter of what is written on my discussion guide.

My colleague, ... , is going to act as a scribe. S/he will try to get the main points of discussion down and we will then type them up and analyse them.

In these groups we would like to operate the 'Chatham House Rule'. By that, we mean that we plan to make a note of what is said, and we would like to use statements made in the group – either to inform our policy, or in any reports or other documents that we might publish. However, if we do use a statement that any group participant has made, we will **not** attribute it to the speaker – either as an individual or as an organisation.

We would ask all participants to observe the same rule.

The participants at this event today are very diverse; there are representatives from government departments, general and vocational qualifications awarding organisations, educational technologists and a student representative. Given these participants; understanding diversity of opinion is a key aim from these afternoon sessions.

So, we would like to ask participants to help us to draw out the range of opinions that exist in the group, and, of course, to treat difference of opinion as something to be valued.

Are there any comments or questions on those points?

Is everyone happy to proceed with the group?

Right, here goes ...

Questions

1. Previous research participants have told us that on-demand testing will be an important driver of e-assessment uptake.
Do you agree?

Why (not)?

Are there any other important drivers of e-assessment uptake?

What are they?

How might they cause e-assessment to be taken up more widely?

2. Broadly speaking, how would you feel about increased use of on-demand testing in qualifications?

Would you be in favour or against?

How does your role or perspective affect your answer?

Is there anything in your fellow group members' answers that surprises you or which you differ from?

3. What could on-demand testing in GCSEs and A levels learn from on-demand testing in other areas (e.g. vocational, professional or licensure qualifications)? Does experience of on-demand testing in another sector make an organisation more or less likely to take it up in general qualifications?

How? Why?

Are there any senses in which general qualifications are different from other qualifications, and how does this affect the use of on-demand testing?

4. The regulators are proposing to develop a regulatory approach to facilitate on-demand testing.

What is your view of this?

Should they be doing this?

Why (not)?

What sort of principles should inform the regulatory approach?

What are the key issues that the regulators need to address?

How do participants' perspectives and viewpoints differ? How might differing perspectives be reconciled?

5. What are the key challenges (difficulties) in doing on-demand testing? How do participants' perspectives and viewpoints differ? How might differing perspectives be reconciled?

How might any difficulties be overcome?

Ofqual wishes to make its publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by The Office of Qualifications and Examinations Regulation in 2010.

© Qualifications and Curriculum Authority 2010

Ofqual is part of the Qualifications and Curriculum Authority (QCA). QCA is an exempt charity under Schedule 2 of the Charities Act 1993.

Office of Qualifications and Examinations Regulation
Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

Telephone 0300 303 3344

Textphone 0300 303 3345

Helpline 0300 303 3346

www.ofqual.gov.uk