PRINCIPLES AND PRACTICE OF ON-DEMAND TESTING

Christopher Wheadon, Claire Whitehouse, Victoria Spalding,
Katherine Tremain & Melody Charman

January 2009

Ofqual This report has been commissioned by the Office of the Qualifications and Examinations Regulator.

The Assessment and Qualifications Alliance (AQA) is a company limited by guarantee registered in England and Wales (company number 3644723) and a registered charity (registered charity number 1073334). Registered address: AQA, Devas Street, Manchester M15 6EX. *Dr Michael Cresswell,* Director General.

Table of Contents

A. EXECUTIVE SUMMARY

This research was commissioned by Ofqual to review how advances in computer technology were enabling on-demand testing in the UK and to consider the implications of these advances for high stakes general qualifications. Their intention was to deepen their understanding of the concerns of stakeholders in this area by looking at current practice in the UK and abroad. The findings will inform Ofqual's regulatory approach to on-demand testing.

The first section of this report is a review of the literature relevant to on-demand testing. This review suggests that on-demand testing ranges from the provision of more frequent test windows to anytime, anywhere testing. In its purest form, on-demand testing clearly supports the personalisation policy agenda and the desire to ensure all students achieve their potential. In its less pure forms the gains from on-demand testing include increased efficiency in the assessment system, with more timely results, and flexibility in scheduling that frees the timetabling of the curriculum from fixed, arbitrary examination dates. There are clearly risks inherent, however, in redesigning the assessment system. Every process from entries to results is affected; these processes are complex and interlinked, and not all in the direct control of the awarding bodies. However, no insurmountable technical difficulties were identified regarding issues such as the maintenance of standards over time and between test versions.

The second section of the report details the views of some key stakeholders in the assessment process: teachers, students and examiners. The teachers and pupils were generally sceptical about the idea of pure on-demand testing supporting a personalised learning programme. They felt that this model would require support in terms of smaller class sizes and greater individual attention from teachers, for example, which would never materialise. Furthermore, both teachers and pupils were wary of an on-demand system increasing exam pressure through competition between peers and parents. They did, however, recognise that more flexibility in choosing testing dates could alleviate some existing pressures as teachers would have greater control over the assessment timetable and therefore the delivery of the curriculum. The examiners welcomed the return to pre-testing that an on-demand system requires and were generally positive about the assessment models that could be used to deliver on-demand testing in a rigorous manner.

The third section reports on a survey of current practice in on-demand testing. Nine major test providers supplied information regarding their current practice, either via interview or by responding to a detailed survey. The scale of provision is impressive. Hundreds of thousands of tests are being delivered on-screen, on-demand every year in the vocational and higher education arenas. Sophisticated technological infrastructures have been developed in partnership with technology companies. These partnerships are yielding innovative assessment formats based on realistic task-based assessments. There are, however, few technology partners available, with eight out of the nine organisations surveyed sharing just two partners.

While it may seem that the major unitary awarding bodies are lagging behind in on-screen on-demand testing, the concerns they need to satisfy are more complex. Vocational bodies tend to use a strong criterion referencing standard setting approach which uses the judgement of experts to determine pass marks before tests are delivered. This can lead to large variations in pass rates over time as seemingly superficial aspects of difficulty in a test can affect how candidates perform on them. This situation would not be tolerated in high stakes national qualifications in England, not least because they are used as a benchmark for national performance. The awarding bodies would need more complex models of test-delivery, with statistical standard setting models integrated into the test-construction and test-delivery processes before they could countenance on-demand testing.

Finally, the report contains a draft set of principles for on-demand testing. These were initially drafted by the research team and presented to a group of technical experts who between them had substantial experience of working within UK awarding bodies and with on-demand systems operating outside the UK. Following their feedback the principles were revised. While the experts broadly reached consensus on these principles they do not claim to be final and absolute. Rather it is hoped that they will provoke discussion and debate and lead to a rigorous framework within which on-demand testing can be regulated. The principles are:

1. EXAMINATION STANDARDS

i. Decisions to move each syllabus to on-demand testing should be supported by a clear educational case. This case should have a sound theoretical basis and be supported by the teaching profession.

ii. On-demand testing should be underpinned by Item Response Theory methods of test-equating.

iii. Policies on item to test ratio, item re-use, pre-test procedures and evidence of coherence of scales should all be available.

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.

v. Systems should be in place to monitor and help explain changes in aggregate qualification outcomes over time.

vi. The reliability of tests should be such that there is little to gain from repeated re-sitting.

2. ACCESSIBILITY

vii. On-screen on-demand tests should provide greater accessibility than paper based tests through the use of assistive computer technology.

viii. Items in item-banks should be tagged according to accessibility requirements so that alternative items which cover the same area and test the same skills can be provided.

3. THE BURDEN OF ASSESSMENT

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.

x. Changes in the burden of assessment in the educational system as a whole, including additional pressures on teachers and candidates, should be monitored.

4. COMMUNICATION

xi. All stakeholders, including candidates, should be actively consulted during the redefinition of processes to support on-demand testing.

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.

B. INTRODUCTION

This research was commissioned by Ofqual to review how advances in computer technology were enabling on-demand testing in the UK and to consider the implications of these advances for high stakes general qualifications. Ofqual is keen to encourage innovation, but has a duty to ensure that this innovation delivers tests that are fit for purpose in terms of validity, reliability, comparability, security, authenticity and compliance with the law. The contract was undertaken by a team of researchers at the Assessment and Qualifications Alliance over the period September 2008 to January 2009.

Three strands of research were undertaken. The first was a literature review of on-demand testing which attempted to identify ways in which high-stakes national assessments (primarily GCSEs and A levels) could be delivered more frequently without compromising their quality. The second strand involved running focus groups with key stakeholders. These included current GCSE candidates, first year university students with recent experience of A levels, teachers at a selective state secondary school, a deputy head teacher at a special school, GCSE Science examiners and a group of technical experts who between them had substantial experience of working within UK awarding bodies and with on-demand systems operating outside the UK. The third strand of research involved a survey by questionnaire of five organisations and by interview of four organisations offering on-demand tests, mostly on-screen.

From these strands of research a set of broad principles for on-demand testing in high-stakes national assessments are suggested. These were initially drafted and presented to the expert focus group. The principles were then redrafted to take into account their comments. Consensus from this group was achieved on all of the principles, although they have not verified or ratified the final principles in any way. For this reason they remain the view of the research team rather than that of any wider body.

The researchers would like to thank all those who contributed to this research, including the staff at Ofqual, through direct participation or comment on versions of this report. It is hoped that the research is of interest and stimulates debate that will produce a robust set of principles that can be used to regulate this area. The views expressed herein are those of the research team alone.

C. LITERATURE REVIEW

1. AN ON-DEMAND WORLD

In '2020 Vision: Report of the Teaching and Learning in 2020 Review Group' Christine Gilbert presented the following vision of what personalised teaching and learning might look like in a 2020 school to the Secretary of State:

> *"Personalising learning means, in practical terms, focusing in a more structured way on each child's learning in order to enhance progress, achievement and participation. All children and young people have the right to receive support and challenge, tailored to their needs, interests and abilities. This demands active commitment from pupils, responsiveness from teachers and engagement from parents."* (Gilbert, 2006, p.3)

Teachers, according to the 2020 vision, are experts in the analysis of data, and use a mixture of formative and summative assessment to ensure that no student falls behind. All learners, regardless of socio-economic background, gender or ethnicity will achieve high standards, possess functional skills in English and mathematics and understand how to learn, think creatively, take risks and handle change. Teachers will operate a fast-response system to ensure learners do not fall off their upward trajectory and parents will become their child's co-educators. The vision is clearly aligned with the Every Child Matters: Change for Children policy agenda (http://www.everychildmatters.gov.uk) designed to protect children and young people from harm and help them achieve what they want in life. While the report did not touch on a different role for high-stakes assessment, the concept of learning tailored to individual needs presents an opportunity for high-stakes assessment to evolve.

The implications of this policy agenda for National Curriculum Tests (NCTs) were drawn out in the consultation document 'Making Good Progress' in which the Department for Education and Skills (DfES) set out the case for making NCTs available on a when-ready basis. The emphasis is placed on the engagement of all, and the progression of all:

> *"The model could be a powerful driver for progression, raising expectations for all pupils, motivating them, bringing a sharp focus on 'next steps' and perhaps especially benefiting those who start the key stage with lower attainment than their peers, or are currently making too little progress."* (DfES, 2007, p.13)

NCT data is obviously of little use to teachers once their students have moved from primary school to secondary; the assumption is that more frequent testing will offer more frequent appraisals of progress and prevent students falling behind. The model is now being trialled, but is as yet to report.

The purpose of this literature review is to consider what is currently known about the issues that will arise as high-stakes general qualifications (primarily GCSEs and A levels) evolve to meet the needs of the personalised classroom of 2020. While experience in the vocational world and the world of NCTs will be drawn on, the assessment issues for general qualifications can be quite different. General qualifications reward performance on a broad syllabus of study carried out over a substantial length of time, and therefore have complex aggregation and compensation models to support them. The uses of general qualification results can also be different to vocational qualifications, especially when they are used as a national benchmark for the performance of the education system. The report will also be limited to what could be termed tests rather than assessment. On-demand testing already exists for some subjects such as music, where a performance can be captured at any time and submitted by e-portfolio. The issues there are quite distinct from those involving more traditional conceptualisations of tests.

2. WHAT IS ON-DEMAND TESTING?

On-demand testing in its purest form is the provision of assessments whenever and wherever a customer (examination centre, teacher, student) wishes to take that assessment. In practice logistical constraints often mean that there are constraints on the flexibility of this delivery. For example the SAT®, a standardised test for college admissions in the United States, is currently available from a specified list of accredited centres on specified dates. While it is therefore not a pure on-demand model, the flexibility of location and dates it offers requires a quite different model of test construction and delivery to that currently used in the UK for high-stakes testing. Table 1 illustrates some of the models of on-demand that have been or could be implemented in the UK.

Table 1: Flavours of "on-demand"

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.

Perhaps the main conceptual difference between the US model behind the SAT® and the UK model behind the GCSE, for example, is the emphasis in the US on questions (usually termed items) rather than tests. In the UK subject experts construct tests; in the US psychometricians construct test versions from items written by the subject experts. Tests according to the UK definition are simply not flexible enough to support on-demand testing. In a system where candidates are likely to take any given examination on different days, and the answers are liable to appear on Facebook at the end of any test session, there will simply never be enough tests. When tests are automatically constructed from a bank of items, however, methods can be employed to draw items from that bank to minimise security threats while satisfying comparability requirements between test sessions.

While item-banks defined as a collection of test items that may be easily accessed for use in preparing examinations (Ward & Murray-Ward, 1994) have existed for many decades, on-demand testing requires a new conception of item-banks. In order to satisfy comparability concerns the difficulty of items in each test version must be known (calibrated). Once calibrated, multiple versions of tests can be produced by an automated or semi-automated procedure and delivered according to non-linear algorithms. This process requires a complex technological infrastructure through which the item-bank interfaces with test delivery, test production and test reporting facilities. An example of this kind of infrastructure is given in Figure 1. Item-banks are therefore only one component, albeit a key component, in a complex architecture that can deliver the vision of an on-demand future for assessment. That vision will now be assessed in some detail before the technicalities of the model are addressed.

3. THE VISION OF ON-DEMAND: MAJOR DRIVERS BEHIND ON-DEMAND TESTING

3.1 Efficiency

According to Bennett (2001) many in the private sector in the United States view education as a huge industry that produces mediocre results for a high cost. It is fundamentally inefficient. In England, there is a clear example of this inefficiency. Applications to Higher Education are based on predictions rather than results as these results are not available sufficiently early to be used in entrance procedures. As a consequence there is no high quality, timely information on achievement which students can use to target their applications effectively, and HE institutions can use to plan resources and the provision of financial assistance for study required for the forthcoming year (Department for Education and Skills, 2005). Similar inefficiencies affect the educational decision-making process at GCSE level. Decisions on whether and what students will continue to study after the age of sixteen can be delayed by a lengthy appeals process. This appeals process for a summer examination series does not end until the following spring. The earlier information is available to direct learners to those courses that will allow them to maximise their potential, the less the impact of mistakes and delays in the process. An on-demand model that delivers immediate results would increase efficiency; if the on-demand tests were multiple-choice they would obviate appeals and give timely access to fairly incontestable information.

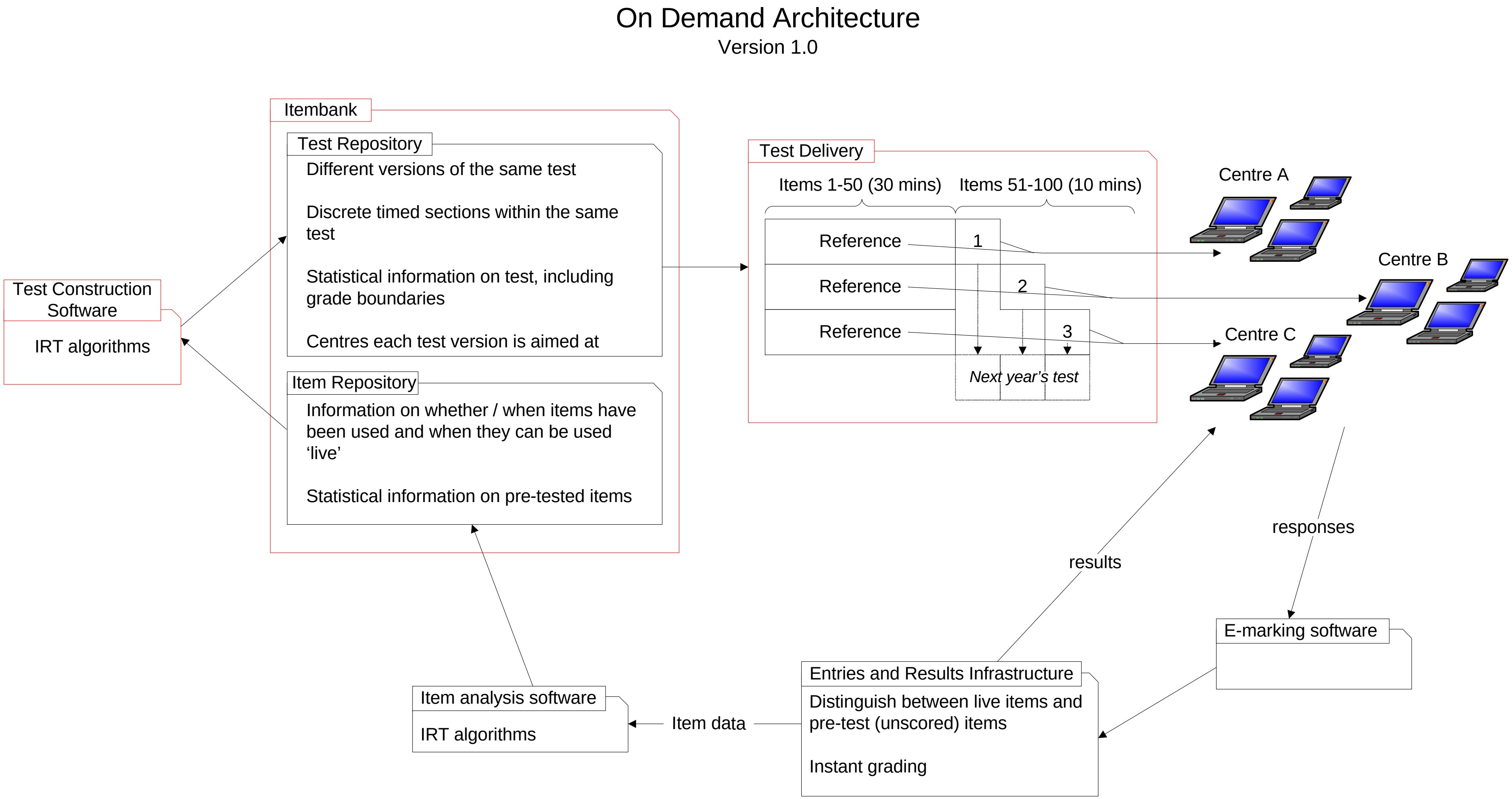

On Demand Architecture
Version 1.0
Test Construction Software
IRT algorithms
Itembank
Test Repository
Different versions of the same test
Discrete timed sections within the same test
Statistical information on test, including grade boundaries
Centres each test version is aimed at
Item Repository
Information on whether / when items have been used and when they can be used 'live'
Statistical information on pre-tested items
Test Delivery
Items 1-50 (30 mins)
Items 51-100 (10 mins)
Reference
1
Reference
2
Reference
3
Next year's test
Centre A
Centre B
Centre C
responses
results
E-marking software
Entries and Results Infrastructure
Distinguish between live items and pre-test (unscored) items
Instant grading
Item data
Item analysis software
IRT algorithms

Figure 1: On-demand architecture

Of course simply making assessments available earlier does not mean that learners will take them earlier. The experience from the US suggests that large scale achievement tests whose scores are only needed once a year are the worst suited to testing on demand (Wainer, 2000b). The most popular dates for the SAT®, which is available on demand, are a Saturday morning in December and in January. This is the latest date at which results are necessary for college admissions, giving students the most time available to study. High stakes achievement tests, therefore, seem most suited by their very nature to mass administration on certain dates. Low-stakes tests where item security is not an issue, licensing tests where results are required immediately and vocational tests which offer more realistic simulations of skills required are identified as better candidates for testing on demand (Wainer, 2000b).

The demand for timely information in the UK is likely to increase soon, given the introduction of more hurdle-based systems into UK assessment. Success in the Diploma, for example, requires passes in all three Functional Skills assessments. Learners will want to ensure that they have timely information on the status of their Functional Skills to avoid failure on these relatively minor components causing lengthy and potentially costly delays to their wider programmes of study.

3.2 Bridging the gap with formative assessment

The frequency of testing implicit in an on-demand model should lead to the provision of better quality, more timely information in the education system. Does this mean that high-stakes testing can finally bridge the gap between formative and summative assessment, and make formative testing irrelevant? Will this lead to general educational gain?

Surprisingly the debate on whether an on-demand high-stakes assessment could play a significant formative role was played out in the UK in the 1980s when a paper-based on-demand system of mathematical modules known as the Graded Assessment in Mathematics Project (GAIM) was developed with the original intention of providing an alternative path to a GCSE (Brown, 1989). Based solely on coursework, even its critics agreed that it provided an interesting and excellent basis for curriculum development. The authors of the programme claimed that it was the continual flow of diagnostic information that delivered excellent outcomes: its critics attributed increased outcomes to a flawed equating model (Noss, Goldstein, & Hoyles, 1989). The technical argument is hard to resolve as the outcomes from different modes of assessment will always be difficult to equate. The argument against the theoretical standpoint that better outcomes were to be expected due to the diagnostic features of the GAIM assessment model is, however, worth repeating. Critics of these gains argued successfully (the GAIM model was never accepted for GCSE certification) that schemes that attempt to provide both grading and diagnostic information are fundamentally unviable and educationally unsound.

The evidence for educational gain through formative assessment comes from a particular model that prioritises dialogue and reflection which builds the self-esteem of the learner (Black & Wiliam, 1998). When diagnostic feedback is accompanied by a grade the feedback loses its worth: grading encourages the suppression of a student's weaknesses and a concentration on maximising assessment ratings or test scores (Noss et al., 1989). Grading dulls the message about what it means to improve, so summative assessment has limited use where teachers have little control over setting the assessment content or marking (Black, Harrison, Lee, Marshall, & Wiliam, 2003). As for the claims of greater student motivation, the wider psychological literature suggests that the provision of extrinsic rewards is likely to have a damaging effect on intrinsic motivation (Deci, 1975). The theoretical case for pedagogical gains from diagnostic information from summative assessment is therefore weak.

More worrying still, in the context of GAIM, Noss et al. argued that the provision of both grading and diagnostic information in a single scheme can be extremely damaging when a hierarchical model of learning is, without theoretical or empirical underpinning, turned into a recipe for curriculum sequencing. If on-demand testing leads to smaller, more carefully defined steps through a curriculum, the epistemological and psychological distortions produced by this didactical transposition should be made explicit. To manage this risk, much closer links with pedagogy will be required to avoid the potentially damaging consequences of ill-conceived modularisation. Nor will this process be simple, as there are no clearly agreed steps to learning that work for all children in all areas of learning. An examination that samples a curriculum after two years of learning, ignorant of the path which that learning has taken, clearly poses far fewer risks to pedagogy.

There are probably cheaper and more effective ways of delivering formative assessment than more regular high-stakes assessment. Any claims that the diagnostic nature of high-stakes tests has led to genuine gains in understanding that feed through to increased outcomes should be viewed with scepticism. Tests are unreliable instruments, and where the best results can be banked candidates are likely to improve their scores by re-taking simply through chance. As Black (2007) stated simply, “test again and again and again – standards will go up”. The costs of the on-demand system to schools should be monitored; as should the final outcomes.

3.3 Validity

The medical profession was quick to realise that a passive learning model with a series of one-off assessments was a poor way to assess the skills of surgeons who need to perform consistently over short intense periods of time and react to situations which are dynamic. As a result a great deal of work has been put into simulations that can be taken on a when-ready basis, that provide just-in-time information in gaming-type environments to assess decision making, and which provide feedback through objective metrics of performance (see Westwood, 2007 for a recent review of this area). As technology slowly changes the ways in which we assess, on-demand delivery may be a more suitable mechanism. While the issues involved with such a change are out of the scope of this report, it is interesting to note the difficulties of obtaining objective metrics from simulations. One study, for example, found that the time taken to complete intracorporal knot-tying was not a good proxy for proficiency in knot-tying as it ignored the quality aspect of the work (Ritter, McClusky, Gallagher, & Smith, 2005). Harnessing the data-streams from more complex assessment formats can present both reliability and validity challenges (Boyle, 2006; Richardson, Baird, Ridgeway, Ripley, Shorrocks Taylor, & Swan, 2002).

While virtual-realities hold future promise for a vocational world, an imminent threat to validity is present in a large investment in technology which delivers a limited range of item-types. Dominance of one mode of testing over all others increases the threat of content irrelevant variance so a mixture of standardised assessment instruments including tests, practical tasks and observations is recommended (Department for Education and Skills, 1988). On-demand testing should mean more variation in assessment instruments, not less.

3.4 Competitive advantage

Another driver for on-demand testing is the desire to achieve a competitive advantage. Registering online for the SAT® you are now promised not only flexibility of time and place but also direct approaches by universities (online registrations only!). This offer reveals an enviable efficiency in data-flows which the legacy systems of the unitary awarding bodies in the UK, designed to support large cohorts being tested in a single series once per year, will be hard pressed to replicate. With increased flexibility, however, come new markets. Just as newspapers can be downloaded on a wireless enabled train to an e-book reader there will be new opportunities for assessment. These may come from the enhanced services that the efficient processes that underlie on-demand testing enable, or from providing a model that is more suited to classroom timetabling, or from an increase in the amount of assessment that is taken.

Key to achieving a competitive advantage therefore, will be the desire to invest in a technological infrastructure to support on-demand testing. Bennett (2001) draws on the analogy of the Encyclopaedia Britannica to illustrate how technology may prove disruptive to assessment, requiring a rethink of processes. Too slow to embrace new technology the Encyclopaedia Britannica suffered a spectacular collapse in the 1990s as sales of printed encyclopaedias fell by eighty per cent (Wurster & Evans, 2000). It has now been reborn on the internet, a perfect medium for its delivery. Its traditional strength, its immense depth of scholarship, is perfectly suited to the richness and reach this technology provides. From paper, through CD ROM, to the internet, Encyclopaedia Britannica struggled with its legacy assets and its legacy mindsets to adapt to a new economics of information. Only a complete collapse and cannibalisation of its own assets ensured its survival.

Assessment has a different commercial model to publishing, however. While Britannica's market research showed that people only opened their encyclopaedias once per year, and purchasing decisions were made largely through guilt over a child's education (Wurster & Evans, 2000), assessment is much harder to do without. In fact a JISC report worried that the result of on-demand assessment would be more demand for tutorial services, preparation materials and an increase in parental pressure (Whitelock, 2006). While encyclopaedia sales in all formats remain at a tenth of their paper sales, this is unlikely to be the case for the new on-demand model of assessment. The modularisation of GCSE Science, providing short assessments in an attractive format with three testing opportunities per year, has led to a huge increase in the number of retakes, with one module offered by the Assessment and Qualifications Alliance attracting over forty per cent retakes. Without restrictions, on-demand testing is likely to lead to more retakes. Does this represent a decrease in the burden of assessment or an increase? Further research needs to be undertaken to assess the motivation for the retakes and whether on-demand would promote a form of teaching that is superficially geared to aspects of the tests rather than aspects of the curriculum. Against this threat, a more flexible approach may mean that, rather than rushing to complete a syllabus to meet an arbitrary test date, a syllabus can be completed in the time dictated by its educational content and breadth.

4. BARRIERS TO ON-DEMAND TESTING

4.1 Technological infrastructure

The most recent report on technological infrastructures for large-scale assessment systems concluded that there is no single solution that can currently be implemented, with existing tools covering only 32 per cent of requirements (Squire, Owen, Baines, & Byrne, 2007). A huge amount of groundwork has been done, however, on technical specifications which would facilitate on-demand testing and the interoperability standards required to ensure the infrastructure is agile (Sclater, 2004; Squire et al., 2007; Whitelock, 2006; Young, MacNeill, Adams, & McAlpine, 2005). The conceptual work on a shared infrastructure model (Sclater, 2004) appears to be bearing fruit for Higher Education and Further Education in the UK Collaboration for a Digital Repository (Squire et al., 2007). With so few technology partners available (Squire et al., 2007) it would seem inevitable that some form of sharing of elements of any system will need to occur, although, given the commercial pressures involved, it is unlikely that this sharing will be done on the open-source basis called for by McAlpine and Zanden (2006).

The commercial software model being adopted is unfortunate, as open source development is starting to deliver on its original promise, empowering educational projects such as the one laptop per child initiative (One Laptop per Child, 2008). Moodle is a particularly good example. An open source e-learning platform, its modular design allows a globally diffuse network of commercial and non-commercial users to contribute to its development. Harnessing a bewildering array of industry standards (IMS QTI, XML and XHTML IMS Content Packaging, SCORM, AICC), it has engendered a rich array of plug-ins that ensure it develops as technology and pedagogy develop. Immensely popular worldwide, Moodle is one of the technologies that drive the Open University's LearningSpace and LabSpace, where the open source concept is applied to the development and distribution of teaching and learning materials. As a business model, the open source model should certainly be considered as a serious alternative to the centralised model of development employed by a commercial organisation. The dialogue such an approach engenders can only help improve compliance to interoperability standards, which commercial organisations have less incentive to adhere to (Whitelock, 2006). Such an approach may lower barriers to entry for all, which is nationally desirable, but may conflict with individual commercial interests.

The technological infrastructure to support on-demand testing is not simply a question for the awarding bodies. Restructuring the entire system from entries to results will impact upon all aspects of data-flows: from school information systems used to submit entries to all forms of reporting of outcomes. The Sutherland Inquiry (2008) attributed some of the failure of the NCTs in 2008 to the lack of end-to-end testing. Timetabling, entries, standardisation, marking reviews and a myriad of other procedures will need to be streamlined to ensure good levels of service for candidates and all other educational professionals from examination officers and teachers through to markers and examiners.

4.2 Maintaining Standards

4.2.1 Item Response Theory

Unitary awarding bodies are charged with maintaining qualification standards over time and voluntarily monitor inter-awarding body standards. While vocational bodies tend to use criterion referencing, usually specifying the same pass mark over multiple versions of tests, a strict criterion referencing approach has never seriously been considered for general qualifications as it tends to lead to large variations in pass rates from year to year (Baird, 2007). The methods used instead were developed to inform the grading of large homogenous cohorts who are tested once per year. For GCSEs for example, the same percentage of candidates is expected to pass any given subject year-on-year (within a limited tolerance range) unless there is compelling evidence to doubt the stability of the cohort. Modularisation of the examination system has led to the development of new systems that take into account changes in the cohort, but these would be stretched to breaking point to accommodate on-demand testing. Any role of expert judgement in the maintenance of standards over time would have to be rethought as it simply wouldn't be possible to convene multiple committees to make judgements over the standards of multiple versions of tests.

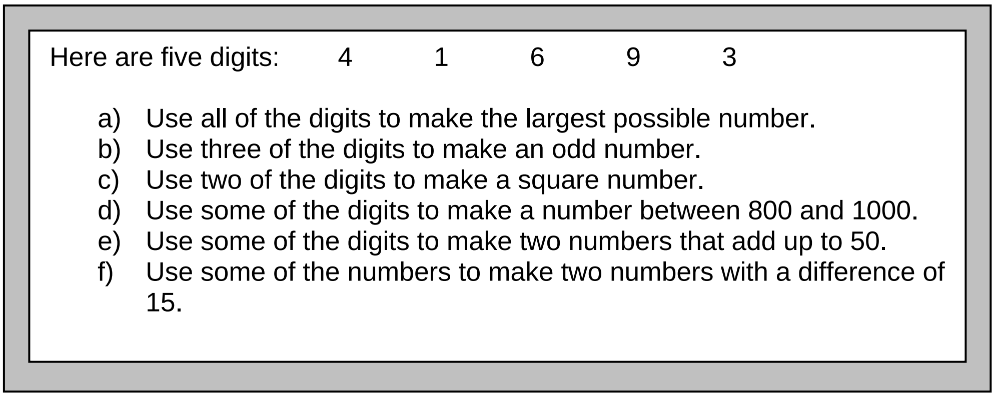
In the US the maintenance of standards over time has always been linked to performance on particular test questions (items) rather than performance of cohorts. Item Response Theory (IRT) models performance at an item level in order to separate the characteristics of the population taking that test from the characteristics of the items in that test (Lord, 1980). IRT models free the measurement of ability from dependence on a fixed set of items, and the measurement of item difficulty from dependence on a fixed population. Given the right conditions, therefore, and the acceptance of some strong statistical assumptions (see later) that do not hold precisely in real testing situations, IRT can be used to compensate for the variation in candidate performance that is due to the variation in difficulty of a test (Kolen & Brennan, 2004). In order to achieve this however, certain assumptions of the IRT model have to be accepted, and changes to the design of tests have to be made to incorporate test equating designs. The assumptions of the models have been subject to controversy in the past (e.g. Goldstein & Wood, 1989) while changes made to test delivery could undermine trust in the entire assessment system.

4.2.2 Violations of the IRT model

IRT developments in the UK came to a sudden halt in the late 1980s following a series of attacks on the IRT test equating methodology employed by the Assessment of Performance Unit (APU). This unit had had been charged with monitoring national standards over time, and, to do so, had developed a series of interlinked assessments that would be equated using IRT. The attacks on the method were backed by a high profile paper which expressed the view that the assumptions of IRT models would always be violated in practical testing situations in the UK, and that the assessments would have to be watered down to meet these requirements (Goldstein & Wood, 1989). It was generally agreed that the APU had lost the debate; shortly afterwards, it was closed down (Panayides, Robinson, & Tymms, In Press). When NCTs were introduced in the 1990s equating approaches were not publicised, were applied in a piecemeal fashion, and depended largely on the enthusiasm of a few key individuals (Bramley, 2006).

One of the most controversial aspects of the use of IRT models in assessment is the assumption of unidimensionality. Unidimensionality requires that one ability is measured in a test (Hambleton, Swaminathan, & Rogers, 1991); yet reality is multidimensional (Goldstein & Wood, 1989). Indeed the architect of modern IRT, Lord (1980) wondered whether chemistry tests that in part involved mathematical training or arithmetic skill and in part required knowledge of non-mathematical facts may not be suitable for IRT models. The predictions were dire: psychometrics may have limited applications (Guilford, 1954); redefinition of the achievement domains to meet IRT assumptions will torture validity (Anderson, 1972); achievement tests will become saturated with aptitude (Willingham, 1980); unidimensionality will be ignored and the statistical models underpinning test equating, item-banking and adaptive testing will be compromised (Goldstein & Wood, 1989).

A study of any test will reveal different dimensions. Figure 2, for example, shows a section of the GCSE Mathematics assessment that a Principal Components Analysis of Residuals consistently identifies as testing a separate construct from the rest of the examination. The wider question is whether this item, purportedly testing knowledge of the number system, is really testing Mathematics at all; but what is in the syllabus must be tested. Bejar (1983), however, provided a key clarification of the requirement for unidimensionality; that it is not necessary for a single latent trait to account for the performance of all the items in a test as long as a coherent scale can be constructed (see also Hambleton et al., 1991). IRT methods of test equating have elaborated on this premise, finding that where different dimensions have been found to exist, they appear to share the same equating function, as the same linear composite of latent traits underlies the item responses on both tests. The overwhelming consensus is that IRT methods of test equating are robust to violations of the assumption of unidimensionality within homogenous populations (Harris, 1993). Dimensionality, however, remains an empirical issue to be monitored; and less work has been done on the interactions between population sub-groups and violations of unidimensionality.



Here are five digits: 4 1 6 9 3

a) Use all of the digits to make the largest possible number.
b) Use three of the digits to make an odd number.
c) Use two of the digits to make a square number.
d) Use some of the digits to make a number between 800 and 1000.
e) Use some of the digits to make two numbers that add up to 50.
f) Use some of the numbers to make two numbers with a difference of 15.

Figure 2: A different dimension from GCSE Mathematics

A second assumption of IRT models that may be violated is that of local independence of item parameters. This requires that candidates’ responses to any question are statistically independent when the ability influencing their performance on the whole test is held constant. Figure 3 shows a question that clearly violates this assumption. Answers to the first question will lead to different chances of success on the second, all other factors being equal. This design is typical of UK assessments which tend to group questions around a context such as a passage or a diagram. The solution is simple: responses that are not conditionally independent should be aggregated. Aggregation of responses introduces a third consideration, which has been less examined. At what level of aggregation do IRT models cease to be useful? Long responses, for example, are marked on a number of criteria which are then implicitly or explicitly aggregated. Are IRT models appropriate in such cases? It is a little studied area, largely because assessments in the US and Australia tend to be multiple choice or short-answer. Some progress is being made in this area (He, 2008) but more is obviously needed.

(i) Name an English port (other than London) that has sea links to Europe. (1 mark)
(ii) For the port in (i) above name:
the main road link leading to/from the port
a port in Europe to which it is linked (2 marks)

Figure 3: **Conditionally dependent questions from GCSE Geography**

4.2.3 Test equating models for on-demand testing

Using IRT it is relatively straightforward to equate different tests as long as some proportion of the items from the tests to be equated are taken by a sample of the entire cohort. As a rule of thumb, around one-fifth of items in any one test should overlap with any other test; and sample sizes should be several hundred, although specific requirements may require more or less of either (Kolen & Brennan, 2004), depending on the IRT model used. Using this simple rule of thumb, an on-demand matrix of overlapping versions can be built which ensures that standards are comparable between versions (Figure 4). There are various designs which achieve this aim (see Beguin, 2000 for a full review of these designs), but they all rely on the premise that the tests to be equated should be built to exactly the same specification and measure the same construct, and where common items are employed, these should ideally represent a miniature of the entire test (Kolen & Brennan, 2004). The concept of a miniature may seem problematic given the tendency in national examinations to produce detailed specifications for the exact balance of assessment objectives, content and skills required by every assessment. The psychometric concern, however, is to ensure that equating produces stable results in a multidimensional context. This is a matter for empirical evaluation rather than a priori description.

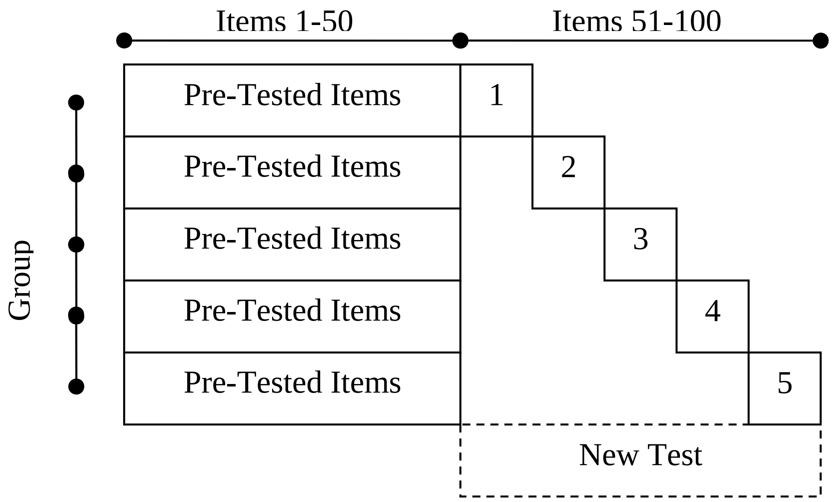


Figure 4: Pre-equating non-equivalent groups design (PENG)

While a matrix design offers some flexibility in the test-forms that can be delivered, and allows problems such as multi-dimensionality to be teased out, pure on-demand testing requires a full item-banking design. The aim of this design is to use items in a bank that have already been calibrated using a test equating design to calibrate new items as they are added to the item bank (Figure 5). One such design uses an algorithm to deliver a random uncalibrated item from the item bank at a specified anchor position within a test to each candidate. This allows a large number of items to be quickly calibrated while minimising their exposure and thus the security risk.

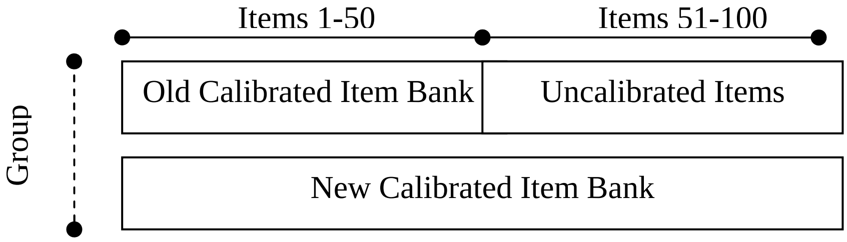


Figure 5: Common-item equating to a calibrated item bank

This design offers the most flexibility and the potential to deliver assessments on-demand. There are, however, practical issues that need to be considered as item banks are used over any length of time. Item difficulty can drift over time as content becomes dated or security becomes compromised. The maintenance of the item bank therefore requires continual care.

4.2.4 Evaluating the test equating designs

Choice of a test equating design requires the various claims of stakeholders to be balanced and evaluated. Pre-testing in live tests provides the highest level of quality assurance for those who set the tests and evaluate the quality of those tests; yet it could be seen as detrimental to the rights of the child and make testing less transparent to teachers and schools. The iniquity of one candidate sweating over a particularly difficult question while their neighbour breezes past an easy question in a different test version may be further compounded if those questions aren’t scored. The equity of current assessment procedures could equally, however, be challenged. As tests are currently delivered to candidates without a sound knowledge of the properties of the items in them, candidates who take a test in one session may be disadvantaged by a question which is contaminated by content irrelevant variance. An ideal situation would be to build a system that satisfies the needs of all stakeholders - pre-test then equate, for example – yet the national assessment system has a requirement to be delivered with efficiency.

One aspect of success of the system, therefore, is engagement with stakeholders. This engagement would aim to build an understanding of what exactly they value within the assessment system, and at what point they would feel disenfranchised by or lose confidence in the system. Loss of faith in the system will have universities fruitlessly reverting to their own tests (Stringer, 2008; Whitelock, 2006) and could cause a political crisis along the lines seen in the Scotland in 2000, England in 2002 (McCaig, 2003) and New Zealand in 2004 (Baird, 2007). All novel features of this system (item re-use, live seeding of pre-test items and different test versions in the same test sitting for example) should be tested against stakeholder perception. Where suitable, for example in the construction of anchor tests, stakeholders should be used in the redesigning the new processes.

On a technical level the challenge will be to ensure that standards are being maintained, that item parameters are stable over time and across sub-populations, that items are to some extent representative of the central construct being tested, and the security of those items is not being compromised. A useful cautionary tale in this regard is what has become known as the National Assessment of Educational Progress (NAEP) anomaly.

4.2.5 The National Assessment of Educational Progress (NAEP) anomaly: A cautionary tale

The cornerstone of IRT and its major difference from Classical Test Theory is the property of invariance of item and person parameters (Lord, 1980). This property implies that the parameters that characterise an item do not depend on the ability distribution of the examinees and the parameters that characterise an examinee do not depend on the set of items. When the IRT model fits the data, the same item parameters are obtained for the item regardless of the distribution of the ability in the group of examinees used to estimate the item parameters. An extension of this property is the assumption that item parameters are invariant across different test forms. Until 1986, the prevailing view was that item parameters are robust to changes in context. Following the National Assessment of Educational Progress (NAEP) anomaly in 1986, however, that view was substantially revised (Beaton & Zwick, 1990).

The NAEP is a relatively low-stakes congressionally mandated survey that is designed to measure trends in what students in American schools know and can do. As with all assessments that are designed to measure changes over time it suffers from the tension to keep its content relevant while following the well-rehearsed maxim that to measure change you should not change the measure. To compensate for changes in the measure deemed necessary to keep content relevant, an IRT test equating design was used. An anchor was constructed that was repeated over time, but following a major overhaul for the 1986 session the anchor items were administered in tests that differed in length, composition, timing and administration conditions. The result was catastrophic: the original analysis showed a dramatic decline in standards of 9- and 17-year old students, but an increase in performance of 13-year olds. Such anomalous results defied credibility and a major investigation was launched. The finding was that although many of the same items were used in both the 1984 and the 1986 assessments, student performance on these items differed substantially when the items were administered in different contexts. In particular, there was no assurance that the time available for the common items was held constant over administrations, and analysis showed that the percentages of candidates who failed to reach certain items were substantially different between administrations (Zwick, 1991). The warning signs were there in the original data as the item facilities had changed greatly, but only a carefully designed counter-balanced experimental design could tease out the proportion of the change that was due to the change of context of the items. IRT could not compensate for the changes in the assessment instrument.

The NAEP anomaly is clearly a cautionary tale. Under all test equating designs it is now common practice for anchors to be delivered as discrete blocks so that their administration and the time available for their completion can be standardised across different sessions. This approach would be suited to assessment designs that administer blocks of questions around specific stimuli such as a passage of text or a diagram. To accommodate this design e-assessment delivery should therefore be able to facilitate the delivery of discrete blocks within a test, each with its own time limit. It then becomes the key responsibility of the test agency to monitor the performance of items that are re-used over time for evidence of drift in any of their key parameters.

4.2.6 Monitoring outcomes

Although standards are to be maintained using performance on items, a key aspect of the validity process is establishing that the aggregation process of units and modules that make up a qualification is fair (Thyne, 1974). Aggregation throws up a great number of technical anomalies which affect pass rates at key grades. If candidates and schools continue to be judged by the achievement of key grades, new systems will need to be in place to monitor the aggregated outcomes at key grades over time and between awarding bodies to ensure that the aggregation and weighting processes are fair, consistent and transparent.

4.3 The skills deficit

It is apparent from the previous section that a different model for maintaining standards between test versions and over time will require quite different skills. The UK possesses very few awarding body researchers trained in IRT, yet no-one is systematically dealing with this skills-gap in educational assessment in the UK. Rather it is the fields of Health, Dentistry and Optometry that are leading in this area with the psychometrics centre at Cambridge (http://www.psychometrics.sps.cam.ac.uk) and The Psychometric Laboratory for Health Sciences (http://home.btconnect.com/Psylab_at_Leeds), based at the University of Leeds in the UK. This is a skills deficit that needs to be addressed if on-demand testing is to be implemented and regulated with rigour.

Test construction from items with known statistical properties will furthermore require different skills to those currently used by examiners in test setting. Awarding bodies are putting into place some elements of this training, including item writing training and the interpretation of item statistics, training that The Chartered Institute of Educational Assessors is seeking to formalise and recognise (Chartered Institute of Educational Assessors, 2008). The vision of large teams of qualified item writers filling item banks is still some way off, however.

4.4 Equity: The digital divide

While political pledges are notoriously fickle (Parkinson, 2008) the £300m pledged recently for a Home Access programme to help low-income, computer-less households does hold out the promise that social inclusion will continue to become less of an issue. There are those who maintain that until assessment breaks out of mainstream delivery mechanisms and finds its way onto gaming consoles or mobile phones that the infrastructure of testing will remain a social barrier (Brown-Martin, 2008). Political initiatives will serve to create different digital divides based on, for example, how good your laptop is. Indeed, if you can play high-stakes poker on your mobile phone then why shouldn’t you be able to interact with a foreign language examiner on X-Box live? As there is evidence that the mode of examinations affects participation (Chamberlain, 2008) the impact of changes in mode brought about by on-demand testing on participation rates should be monitored.

A second digital divide that on-demand testing must be careful not to exacerbate is for those who currently request specific access arrangements. Advances in computer technology hold great promise in this area. The computer technology exists to allow candidates to customise the display of tests on-screen, to change the font size, style and colour and background colours to suit their individual needs. This approach will be more flexible than currently exists for modified question papers which are only available in four standard formats.

It should also be possible to convert on-screen text into Braille format or synthesised speech obviating the use of readers. Similarly voice recognition software will allow candidates’ verbal responses to be captured as text instead of using a scribe. Assistive computer technology exists for people with mobility impairments: keyboard actions can be used instead of mouse control for those with manipulation problems; puff-and sip devices allow a user to move the mouse pointer without using his or her hands by puffing air into a tube; single switch devices allow users to interact with a computer by using slight body movements.

Ultimately on-screen tests, whether or not they are on-demand tests, can provide greater accessibility than paper based tests. In on-demand terms a carefully authored and edited item bank will provide an opportunity to identify and adapt any items which may have barriers for specific disabilities such as graphics for those with visual impairments or sound clips for hearing impaired candidates. In most cases alternative items could then be designed which cover the same area and test the same skills without the barriers. This would allow an on-demand test to be requested that is suitable for a visually or hearing impaired candidate.

4.5 Test construction models

On-demand testing will require a radically different testing model for national assessment. For general qualifications the specification sets out a programme of study and states the assessment, content and skills sampling that is expected of any one test. GCSE Science for example has three assessment objectives given particular weightings in any one examination paper. In addition there must be a certain number of 'How Science Works' questions that criss-cross the assessment objectives. The more constraints put upon a test construction algorithm, the larger the item-bank will need to be, and the more complex the algorithms needed to assess whether any form of automated test-construction is successful (Linden, 2005). Automated or semi-automated test construction is only feasible if the descriptions of the assessment objectives and content areas to be covered are clear and concise (Whitehouse, He & Wheadon, 2008). Tests are produced to serve a purpose, however, not a mode: objectives and content areas shouldn't be simplified merely to suit the needs of on-demand.

5. BEYOND LINEAR TESTING

Items in calibrated item banks can be used to design IRT-based tests which can be administered the same way as conventional tests. The advantage of IRT-based tests over conventional tests is that the standards (represented by grade boundaries or pass and fail marks) are set once a test has been designed rather than after it has been sat.

5.1 Computer Adaptive Testing

Once an item-bank has been established new possibilities in testing open up, not least of which is Computer Adaptive Testing (CAT). CAT was conceived by Lord (1980) as a way of providing an individually tailored test that could be mass–administered. An adaptive test is one that adapts the difficulty of the questions offered to candidates to suit their ability as illustrated by their response pattern. Thus, if a candidate fails to answer a question correctly an easier question is presented. If this question is answered correctly a more difficult question is presented. This process continues until the candidate's ability is measured to a predetermined degree of accuracy. Green (1983) outlined the major advantages expected of CAT as improved test security and an appropriate level of challenge for all candidates. Improved test security was expected as any one candidate would only see a small proportion of the total questions in the test pool: if this pool is large then learning the pool would be analogous to learning the subject (Wainer, 2000a). An appropriate level of challenge would ensure that time was not wasted on questions that were too easy or too hard for candidates: the brightest would be challenged while the weakest would not be discouraged.

The appropriate level of challenge has indeed proved a popular feature of CATs. In the United Kingdom the largest operational CAT is the Computer Adaptive Baseline Test (CABT) offered by the Curriculum Evaluation and Management Centre at Durham University. In 2005 over 100,000 adaptive tests of mathematics and vocabulary were delivered to 11 to 16 year olds using a Rasch-based adaptive algorithm. The tests have proved reliable psychometrically with a test-retest reliability above 0.9 and been welcomed by teachers as improving the testing experience of students (personal communication, Coe). Used as a baseline test, however, the CABT has the advantage of being delivered in a low-stakes environment.

In a high-stakes environment CATs have proved to have significant security flaws. The problem with CATs is that item selection algorithms do not choose all items with equal likelihood and a very small proportion of the item pool accounts for a large amount of the items administered (Wainer, 2000a). A common finding is that between 15 and 20 percent of the item pool accounts for more than 50 percent of the test items being administered. This occurs even when the distribution of difficulty of items in the item pool matches the ability distribution in the population. The result is that the tests delivered overlap considerably, especially for the most able students. These able students are precisely the students who can reproduce the items they are asked most accurately. According to Wainer (2000b) Kaplan Educational Centres were able to exploit this flaw to methodically steal a large proportion of the itembank being delivered by Educational Testing Services (ETS) for the American high-stakes test, the Graduate Record Examinations, the largest operational CAT in the world. Modelling how this was possible Mcleod (1999) found that by asking 8 candidates to memorise the difficult items they received a low-scoring examinee could use this information to increase their score by three standard deviations.

Although the case against Kaplan was never proven, and ETS denied that the security of the GRE had been compromised (Frantz & Nordheimer, 1997), item exposure, which models ways in which item pools can be more effectively utilised and test overlap limited, has become a major field of study. ETS withdrew the CAT version of the GRE from 2007 in favour of linear tests, citing security concerns as the main reason (ETS, 2006). While there are still many successful CATs, these are generally employed in fast-moving technology fields which require detailed knowledge that can be easily varied. It would take a brave assessment agency to ignore ETS's withdrawal from CAT.

5.2 Multistage testing

In addition to the security concerns surrounding adaptive testing, the need for item-level adaptive tests to be constructed “on-the-fly” using some form of automated test assembly presents limitations to their use. Complex specifications may need to be relaxed for their use as the sequential test assembly is not optimal, while design flaws can cause unintended test assembly issues. Some examinations may also have content requirements that are difficult to quantify or implement as rules. To deal with these concerns, and to ensure that stakeholders have sufficient input into the test construction process, an alternative design known as Multistage Testing (MST) has been implemented (Mead, 2006).

Multi-stage testing has the same aim as CAT: to shorten the test length while optimising discrimination. While CATs require complex algorithms to be built into test players to decide the selection of the subsequent item on every case, MSTs have built-in paths that lead candidates through a series of testlets. Depending on a candidate’s score on a particular testlet, they are directed to a subsequent testlet. Figure 6 illustrates a 1-3-3 module computer adaptive sequential test (CAST) configuration (Luecht, Brumfield, & Breithaupt, 2006). The possible routes through the seven testlets are indicated by the solid and dashed lines. Most examinees are expected to follow the solid pathways; the dashed lines compensate for unexpected performance. Some pathways, for example from 2E to 3H are precluded. The seven testlets and the associated routing rules are packaged together in units called panels. Figure 6 depicts multiple panels which can be assigned to examinees just like multiple test forms.

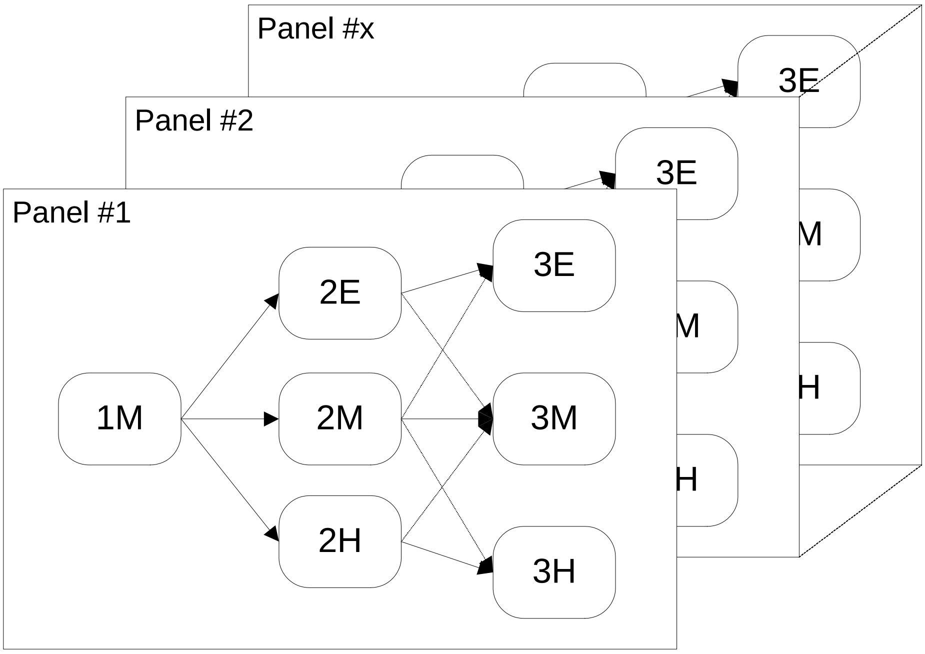


Figure 6: Design for a 1-3-3 computer adaptive sequential test configuration with multiple panels. E=relatively easy; M=moderately difficult; H=relatively hard.

This design alleviates many of the problems found with CAT: MST developers should never get back their results and find that 20 per cent of the items in the pool have been used on 80 per cent of examinees' tests as is common in CAT settings (Wainer, 2000b). All tests can be subject to the same quality review procedures as are currently in place for general qualifications, and the test delivery software does not have to handle complex scoring and item selection algorithms. This approach would seem particularly suited to general qualifications which have struggled with the problem of differentiation since they were re-launched in 1988 with the brief to emphasise positive achievement while retaining optimal discrimination (Good & Cresswell, 1988). The current approach, tiering, has significant technical flaws (Wheadon & Beguin, In Press) and has been criticised for the need to allocate candidates early on in their course of study to a level (or tier). MST leaves the decision until the last possible moment, and makes the judgement on objective information available at the time of testing. The decision to use such models concerns more than just the tests themselves: whether or not candidates of all levels of ability should follow the same syllabus, an assumption to some extent implicit in this model, has wide-ranging implications for education.

6. CONCLUSION

Four major areas of concern emerge from this review. Broadly they relate to the monitoring of systems, outcomes, validity and participation. The Sutherland Inquiry (2008) highlighted the need for all systems from entries to results to be integrated so they ensure a high level of service for everyone involved in the business of assessment. Some of these are under the direct control of the awarding bodies, while others such as the entries procedures and their interfaces with Management Information Systems are not. Interdependencies need to be clear, critical paths mapped and contingencies available.

The monitoring of outcomes and the procedures that underpin those outcomes will need new regulatory methods to ensure that results between versions, over time and between awarding bodies are transparent and fair. Professional judgement of standards will play a much reduced role in an on-demand world. Instead of ensuring that procedures are being followed, the current regulatory model, test-equating designs and item statistics will need to be scrutinised to ensure that standards are not being allowed to drift and candidates are not being disadvantaged.

Validity evidence will continue to play a vital role in an on-demand world. Syllabuses should not be compromised in order to achieve on-demand testing. Given the substantial investment in technology required to deliver on-demand testing there will be a temptation to achieve economies of scale without due consideration of which syllabuses lend themselves to on-demand testing. Although there will inevitably be some interplay between assessment methods and the ways in which a syllabus is taught, the syllabus should determine the testing mode, not the mode the syllabus.

Lastly it would seem sensible to monitor whether the move to on-demand is really delivering the personalised vision of 2020. This is perhaps the hardest task of all. Participation rates can be monitored and shifts in the achievement of demographic sub-groups mapped over time. Whether genuine educational gain is being achieved will be far harder to monitor. Claims related to better results being achieved due to the enhanced validity of the assessment process or as a result of better diagnostic information should be subject to close scrutiny as the theoretical case is weak. Only well designed research experiments will be able to tease out some of the reality of educational gain or loss from the confounding factors that abound.

D. STAKEHOLDER PERSPECTIVES ON ON-DEMAND TESTING

1. INTRODUCTION

The authors of the report to the Teaching and Learning in 2020 Review Group (see Gilbert, 2006) are clear that deep and complex changes will need to occur in schools if its vision of personalised learning is to be realised. Schools will have to innovate if they are to meet this challenge: change will have to be embraced. There may be little disagreement with the concept of testing when ready, but as the Chair of the National Governors' Association at the Westminster Education Forum emphasised, the details of the system are critically important:

> *"You need a system that is physically accessible, i.e. there where the teacher is, not somewhere in another part of the school. It's also got to be easy to navigate. It's got to give teachers support in their teaching and planning and assessment, but it mustn't be overburdening, it's got to be manageable, it's got to be well-integrated and not cumbersome and an add-on and I think that's essential because governors of course are very interested in the well-being of children, but we are rather interested in the well-being of our staff as well and we don't want them to sink under another burden and a different kind of pressure."* (Bennett, 2008, p. 33)

Although the above statement recognises the need for high-quality assessment data and for teachers trained in the use of that data, the risks of overburdening and increasing pressure on teachers are clearly highlighted. Seeking specific feedback at this early stage on how an on-demand system might work is of limited use because stakeholders will have limited practical experience of on-demand testing. Nevertheless, as the Sutherland Inquiry (2008) emphasised, the community of assessment must be consulted during the process of change if that change is to be successfully implemented.

Therefore, three separate focus groups were conducted with teachers, students and examiners. A flexible approach was adopted, so that issues not explicitly considered by the researchers could be indentified and discussed in the focus groups. Similarly, definitions of on-demand testing were kept vague, so that the broadest spectrum of conceptualisations of on-demand testing could be explored.

2. METHOD

2.1 Participants

Given the time constraints on the project, participants were recruited on an opportunistic basis. Two focus groups were conducted with students. The first focus group consisted of two male and two female participants, who were first year undergraduate students at the University of Surrey. The second focus group consisted of three male and three female year 11 GCSE Science students from a selective state grammar school. University and GCSE Science students were selected because they had insight into modular exams, which are the current form of assessment most akin to on-demand. A £10 book token was offered to the university and school students for taking part in the study. The teacher focus group consisted of one male and two female GCSE Science teachers from the same selective state grammar school. Additionally, on a separate occasion, a deputy head teacher of a special school for students with behavioural and emotional problems was interviewed. The examiner focus group was made up of three examiners, one each from GCSE Biology, Chemistry and Physics, all with substantial experience in examining stretching back to the 1970s.

2.2 Procedure

The student discussions lasted for half an hour and were based on the session plan shown in Appendix 1. A loose topical approach was used, as opposed to specific questions, to allow for the discussion of issues which had not been identified by the researchers. The university student group were also provided with stimulus materials (see Appendix 2) to re-familiarise the participants with GCSE examinations. Alternative materials (see Appendix 3) were used to facilitate discussion in the GCSE student group.

The teacher discussions lasted for half an hour and were based on the session plan shown in Appendix 4. A structured question approach was used to elicit teachers' views on specific aspects of on-demand testing. No stimulus material was used, but a technical expert on on-demand testing was available to answer any specific queries on how on-demand testing may work in practice should they arise. To stimulate the discussion with the examiners, they were presented with two hypothetical visions of on-demand testing, and asked to think through and discuss the implications of each (see Appendix 4). Again, a technical expert on on-demand testing was on hand.

All participants, including the students were asked to give their explicit written consent to their participation in the study after having been told of its aims and outputs. The participants were informed that sessions were being voice-recorded and that all of their comments would be kept anonymous; all participants' names have been changed. The qualitative data were transcribed from the audio recordings and then analysed using dominant themes analysis. While focus group data do provide a rich source of qualitative data, the findings are limited by the small sample size. The data received from the focus groups reflect the views of very small sub-sets of the populations they represent and thus may not reflect the views of those populations generally; however, they can highlight some of the issues regarding on-demand testing that concern teachers and students.

3. STUDENT FINDINGS: DOMINANT THEMES

Five dominant themes emerged from the analysis: exam pressure, exam integrity, frequency of exams, the effect of on-demand testing on schools, and the effect of learning factors on on-demand testing. This section outlines and discusses each of the themes, using quotes from participants where appropriate.

3.1 Theme 1: Exam pressure

Several factors influence the degree of exam pressure experienced by candidates. The participants identified: frequency of exams, amount of and type of revision, peer competition, parental pressure, availability of re-sits, location of exam, and group support as factors affecting the amount of stress they felt over exams.

On-demand testing can provide flexibility as to when candidates take exams, and it is hoped that allowing candidates to take their exams when they feel ready will reduce exam pressure. The school students' comments about the pressure of exams were focused around the amount of revision preceding the exams. Many school and university students felt that on-demand testing would allow them to spread out their exams, alleviating the pressure of doing all of their revision at once. Additionally, some of the school students felt that having the facility to take an exam when they felt ready would relieve such pressure:

> *"Yeah, spread it out more…so you don't have to revise more at the end of the year, you get all this revision at the end which is quite stressful."* (Paul, school student)

However, university students felt that on-demand testing would create a competitive environment which would contribute to exam pressure. They felt that weaker students may be adversely affected by stronger students forging ahead:

> *"I think it puts more pressure on you to be ready at a certain time because everybody else has already taken them and you're like – ah, I'm not ready yet."* (Sarah, University student)

The school students did not talk about competitiveness among students; however, a minority were concerned that there may be parental pressure on students regarding when they should take the exams:

> *"I think they'd [parents] probably think that you weren't doing it, like if you said 'oh I'm leaving it right 'til the end of the year then they might be a bit like, 'you're not doing anything, you're just sitting there, you're waiting for it all to come to you' sort of thing."* (Vicky, school student)

Several factors which would reduce exam pressure were identified. Firstly, both groups felt that re-sits significantly reduced the pressure they felt when taking exams:

> *"It's [the re-sit] more like a safety net. It's a good thing to have, just so you know that if there is some sort of mitigating circumstances than you can always do it again some other time, rather than just completely fail. I'm more comfortable knowing that I can do re-sits."* (Nick, university student)

Secondly, university students felt that working through past papers reduced their nerves, although school students did not talk about past paper practice. Thirdly, the university students felt that taking the exams in a familiar environment reduced exam pressure; however, the school students were unconcerned by the prospect of taking their exams at a test centre, claiming it would make little difference in terms of exam stress for them. Lastly, both groups felt strongly that taking the exams with their friends greatly reduced exam pressure. The university and school students reported that working with friends reduced pressure in a variety of ways; through supporting each other, preventing isolation, creating a sense of solidarity, sharing experiences of the exam paper, and avoiding a competitive atmosphere. Both university and school students were concerned that the individualised study that on-demand testing facilitates would have a negative impact upon the comfort and help they found working in a class atmosphere:

> *"It's still like scary to take it all together but you know like everybody's in the same boat, you're all doing it at the same time. So it's like, less nerve racking."* (Jane, university student)

> *"I don't think I'd like doing it on my own, cause I'd feel like, I'd be like looking around and thinking oh no, no one else is doing it, you'd rather be with people in the same situation."* (Ilona, school student)

> *"You'd wanna do it when your friends do it and if they're not ready you'd be like 'oh I don't really wanna do it then'."* (Paul, school student)

Both university and school students were also concerned that on-demand testing would highlight differences in ability. If the brighter students take their exams earlier, weaker students are likely to feel disheartened as they are 'left behind':

> *"When other people go ahead I'd be left behind I'd feel like worse they're going on and doing exams and I'm not ready for it yet. So it's better to do it all at once."* (Jane, university student)

The school students also felt that they were more confident that their responses would not be lost if they sat their exams in groups, and if the worst were to happen they would be reassured if their friends were in a similar situation.

3.2 Theme 2: Exam integrity

The participants were concerned about the comparability of on-demand exams, opportunities for cheating, and the implications of unlimited re-sits. Some of the university and school students were alarmed that, in any given year, there could be several different exam papers which candidates could take. These students were concerned that a standard for all these different papers could not be maintained with on-demand testing:

> *"If at the end of it, say, me and my friend did a different exam but she got like, much higher marks than what I did, I'd probably be like 'oh I had a harder exam', probably try and blame it on the exam."* (Vicky, school student)

Others, however, showed a greater understanding of how grades are awarded and felt that it was little different to the existing modular A-Levels:

> *"It's the same at A-Levels when you took another module, and when you re-took. There are different questions there."* (Jane, university student)

> *"I think that if like questions were equally difficult or if they were like worded differently or different situations of the question, then it would probably be ok."* (Sam, school student)

The university students also felt that the re-use of questions, on which on-demand testing relies for test-equating, would be taken advantage of by cheats:

> *"Yeah you could just whack [the questions] on Facebook or something like that."* (John, university student)

One student suggested that any repetition of questions would be noticed by centres and would encourage teaching to the test:

> *"But quite often don't schools keep the paper anyway? So if there are any repetitions of questions then I think that you're going to get people noticing that and try to exploit it."* (Nick, university student)

The school students did not discuss cheating.

Pure on-demand testing would vastly increase the opportunities for candidates to re-sit examinations. While one university student jokingly requested unlimited re-sits, there was a general consensus among both groups that there should be a minimum period between attempts, or a limited number of attempts, to prevent candidates re-sitting repeatedly in a short period of time:

> *"I think that you should be allowed to get as many in by a set date so like if you say right everyone's got to have done this GCSE by whenever and then if you sort of fail it then you can keep trying to get it done by that date."* (Paul, school student)

> *"I think there should be like a certain amount so like you're only allowed three or four or something."* (Alex, school student)

3.3 Theme 3: Frequency of exams

As mentioned in theme one, participants were in favour of exams being spread across the year, as they felt that this reduced exam pressure.

> *"I'd prefer if every subject was split up into modules, then you have the exams when you have the modules so that it's split up into sections, 'cause then, at the end of the year, you've got to put in loads and loads of revision for like every subject cause none of them do exams like before."* (Paul, school student)

The school students were in favour of the modular approach; however, they had concerns regarding the total number of exams this may result in them taking and the possibility of taking exams before they were properly prepared. The university students had mixed feelings with regard to modular exams. Whilst they felt that they would like to be able to take an exam immediately after the end of a topic as it would ease revision, they also felt that, in some subjects, the end of the course would be better, as they would have had more practice and gained a deeper understanding of the subject overall. There was no consensus as to how frequent exam windows should be to suit their learning styles:

> *"Its good to have it in modules because then you'll immediately…have that information, whereas GCSEs you have to look back on two years worth of work and get to the start of stuff and like, 'I did this?', whereas A-Levels are split into smaller modules so that you know you have only got a small chunk to revise but then when you get to the end it is a lot easier to do some stuff so, it's a bit of both."* (Sarah, university student)

The alternative approach to more frequent exams is to offer windows at more frequent intervals. In this scenario, students could still only take one exam for their GCSE but would have the choice to take it at any point in the year. In this way candidates could stagger their exams. The school students wanted to stagger their exams:

> *"I prefer fixed dates but not all at the end of the year, throughout the year, so then you can concentrate on one thing at a time so then there's more chance of passing it 'cause you haven't got to worry about everything else."* (Ilona, school student)

However the school students were concerned that they would not cover all the work necessary for the exam if they took it early and would be at a disadvantage compared to those who took the exam later:

> *"You'd have to try and cram in more and learn in a shorter space of time."* (Vicky, school student)

The university students felt that a high frequency of exams would not be appropriate. They felt that every week, for example, would be too frequent and questioned how this would benefit candidates. Nevertheless, the university students felt that an interim period of a month between exam windows was suitable.

3.4 Theme 4: The effect of on-demand testing on schools

The participants raised several concerns regarding how on-demand testing might impact on the educational system. There was discussion in both groups about how on-demand testing could be timetabled within their school and how schemes of work would be covered. Concerns were raised over how teachers would cope with on-demand testing, and how this would affect their revision material.

Neither student group thought that a pure on-demand system could be practically implemented in schools. The participants did not feel that schools could provide the ultimate flexibility of a pure on-demand system. They were concerned that schools would not be able to teach candidates at differing rates and that, in attempting to do so, teaching quality would suffer:

> *"Yeah, so if someone, if you've got different people in the group doing the exams earlier then, especially in smaller schools where it might not be easy to split them up, then you're going to end up with kind of people at different stages which is going to make the teaching less efficient."* (Nick, university student)

> *"You couldn't do it literally individually because then the teachers would have to keep up with every pupil and it would get really confusing."* (Jane, university student)

> *"When we have it at the end of the year we sort of know we've got to try and get through all this by the end of the year. If you pick when you wanna do it and people pick it at different times, then some people have got to learn different stuff by a different time so I'm not sure how that would work."* (Vicky, school student)

The school students discussed how they would fit the work in for a subject if they took the exam earlier, expressing concern about the fairness to those candidates opting to take the exam early. One potential benefit of on-demand testing is allowing candidates to enter for more exams. The school students felt that, while in theory this was good, there was still the difficulty in timetabling in the extra work required for doing more GCSE subjects:

> *"I think if they split them up then there would be probably more time to do more exams, but then still you would have to fit extra lessons into your timetable. Yeah, so then you wouldn't have like so many periods on each lesson, so you wouldn't be able to learn as much in like a week or whatever."* (Paul, school student)

University students did not talk about the logistics of covering the schemes of work. Having experienced a Microsoft Office on-demand assessment system (MOS), the school students were concerned as to how the exams would be timetabled into their day. Several students in their school had to stay after school hours to do the MOS exams. School students were also concerned that on-demand exams would affect the current mock exam system and revision sessions held by teachers. While the university students did not talk about mock exams or revision sessions, some did indicate that past papers were an important part of their revision. The school students were unconcerned about the potential unavailability of past papers.

3.5 Theme 5: The effect of learning factors on on-demand testing

The university and school students identified a range of factors which mediate candidates' learning, including self-discipline, the need for goals and structure, and maturity level. In an on-demand system where students choose when they are ready, candidates will need to have self-discipline. Although the school students thought they would like to decide when they are ready to take an exam, both student groups felt that if candidates were left to choose that they would put the exams off until late in the course:

> *"I think it would be better but then, if it was me, I'd try and like delay them as much as possible 'cause I'd be really nervous and then I'd have them all at the end, all built-up."* (Ilona, school student)

Some students felt that the choice as to when to take an exam would give rise to added stress and complications for candidates:

> *"I think having to choose a date is just more stress than there needs to be; a fixed date you've got something to aim for."* (Sam, school student)

It was for these reasons that both groups decided that it would be best for the teachers to decide when a candidate is ready for an exam, but that candidates should be consulted.

The need for goals was identified as key to most of the participants' learning styles. Both groups said that the current GCSE exam windows gave a clear deadline to prepare for. They liked the structure of current classroom teaching and said that it helped motivate them to prepare for exams and prevented them 'putting off' work. When the school students were given the choice between being tested when ready and a set test date, the majority chose the set date to work towards.

The need for clear goals for school students ties in with the issue of maturity. The university students felt that GCSE students would need guidance and were not mature enough to set their own goals and exam deadlines. Additionally, they felt that if candidates were able to choose their own path through their GCSEs, pick what exams they took and when they would take them, this would result in added stress for the candidates:

> *"I think that's too soon to be that individualised, too soon to be that much like independent really, I don't think many people would do it to the best of their abilities, I think you need a lot more guidance."* (Sarah, university student)

> *"If you've got the freedom to choose it yourself then there are a load of issues to deal with, like if you've decided to timetable something wrong."* (Nick, university student)

4. TEACHER FINDINGS: DOMINANT THEMES

Four dominant themes emerged from the teacher analysis; exam integrity, the effect of on-demand testing on schools, requirements for on-demand testing, and the effect of learning factors on on-demand testing. While the majority of the themes identified in the analysis were common with the university and school students, requirements for on-demand testing being the only exception, teachers had a different perspective on the issues. This section outlines and discusses each of the dominant themes, using quotes from participants where appropriate.

4.1 Theme 1: Exam integrity

The participants discussed comparability, cheating and re-sits in relation to exam integrity. All of the participants shared the students' concerns over the comparability of on-demand exams. The teachers were concerned that on-demand tests would not be comparable, and some felt that the parents and the students would not consider the exams fair:

> *"And it would be seen as 'oh you got the easy paper that's why you got better marks'. It would give us another thing for parents to hit us over the head with."* (Brian, Science teacher)

Neil, the deputy head teacher of the special school, however, felt that so long as it was ensured that the questions were of the same difficulty, he would be comfortable with the new system. Regarding questions on fractions, for example, he commented:

> *"A third add a half is not that dissimilar to three quarters add a fifth. If you're talking about seventeen fiftieths, it's not comparable."* (Neil, deputy head teacher of special school)

All participants felt that cheating was problematic for any assessment system, but they did not consider on-demand testing to be at particular risk of cheating:

> *"The security of an exam paper will always be compromised if somebody wants to compromise it."* (Neil, deputy head teacher of special school)

Some of the teachers felt that unlimited re-sits would damage the integrity of the exams and there was a general consensus that a limit should be imposed on re-sits. Additionally, the participants were asked how they felt about questions being included in the exam which were not marked in order to pre-test the items. All the participants said that they were unconcerned by this so long as only a small minority of the items on an exam were unmarked:

> *"I wouldn't have a problem with that, no…no we would be happy with that, as long as the kids weren't aware that it wasn't being marked, we'd keep that very quiet."* (Brian, Science teacher)

Neil raised concerns that the large bank of questions required for an on-demand system would push the limits of exam writers' creativity.

4.2 Theme 2: The effect of on-demand testing on schools

The teachers' comments on the effect of on-demand testing on schools echoed those found from the students' focus groups. The participants were concerned about how to timetable lessons in an on-demand world, how schemes of work could be covered, the amount of administration on-demand tests would generate, and the effect of on-demand testing on their students.

The participants had strong concerns about how they would implement a pure on-demand system in their school. The teachers felt that they would not be able to timetable classes if their students were taking tests at different times:

> *"I think it would be an absolute nightmare, I really do. What I'm looking at is if I'm looking at a class that I take, and half of them have had an exam and the other half haven't, what do I do with the half that have? 'Cause they now switch off and then. I think it would have to be a whole class at a time, or, we can't have... I know where you're going with the flexibility but I don't think it's going to work."* (Brian, Science teacher)

Even with a high teacher-to-student ratio, Neil still felt that he would still struggle with a purely on-demand system. He felt that, as well as timetabling, additional issues may arise such as resentment from his students if some are put forward for an exam while others are held back as they are not ready. Equally, the teachers felt that this competitiveness would be problematic in their school as parents would put more pressure on the teachers:

> *"It could be an absolute nightmare, absolute nightmare. I mean, I'd say we are quite fortunate, we have 6 children in year 11; it's manageable. I don't know how manageable that would be in another type of school. Ah, people turning up on the wrong day for exams, people think they are going to enter for an exam when they are not because they're not ready. And this issue of 'why's he ready, I'm not?' can cause resentment, ah, between pupils. 'He's already done two exams; I haven't done any - why?', 'Cause you're different'. 'That's not good enough; I want to do an exam'."* (Neil, deputy head teacher of special school)

The teachers did not want more frequent exams. They felt that the current four exam windows were already disruptive to the students' learning. Additionally, the teachers felt that even more exams would result in an increase in the amount of administration they would have, which would be impossible to cope with. They felt that the only way in which schools could use on-demand testing would be to provide more frequent opportunities to enter their classes into exams, allowing them more flexibility in planning the year:

> *"As I say, if the exams were more often, we would decide on a different time but it wouldn't be so flexible as when they are ready or not or something like that. We would decide for our next two years which dates we are going to choose."* (Claire, Science teacher)

Neil, felt that more frequent exams would be beneficial as they would provide a means of documenting the progress of the students.

4.3 Theme 3: Requirements for on-demand testing

The third theme to emerge from the analysis is closely related to theme 2. The participants discussed teachers' requirements for an on-demand examination system. The participants considered support materials, exam feedback, online availability, school facilities, class sizes, and planned flexibility in the future.

The participants stressed the importance of past papers as preparation materials for their schools. All the participants felt that specimen material was a poor replacement for the past papers and were concerned that past papers may not be available in an on-demand system. The teachers wanted a secure internet area where they could access past papers for the students to practice and to use for mock exams:

> *"I mean the mocks could have sort of could be sent to us on a disk and we could run them internally on the virtual learning environment. That doesn't have to be online as such; it could be within the school but could actually recall which kids tackled which parts of the paper."* (Brian, Science teacher)

Neil felt that new schemes of work would need to be provided if on-demand testing were to be implemented, as the current schemes of work would not fit with on-demand exams.

The teachers were very keen on statistical feedback on the exams. They felt that this helped them direct their teaching and improve their students' understanding of where they went wrong on the exam paper. The teachers were keen on online implementation of exams as they felt this would result in faster processing of results:

> *"…more feedback and perhaps a bit more analysis of where the areas of strengths and weaknesses are in the multi-choice, because we know we're teaching some things well and we know we're teaching some things badly, but we're not quite sure which is which at the moment 'cause we don't get anything like it."* (Brian, Science teacher)

Whereas the teachers felt that on-demand would increase the amount of administration, they felt that online testing would significantly reduce the amount of administration that the school would need to do. Some teachers were concerned, however, as to the amount of strain on-demand testing would put on their computer resources, despite the school having excellent computer facilities.

The teachers felt that class sizes would need to be dramatically reduced in order for on-demand testing to be done in an individualised fashion. One teacher suggested class sizes of 10, although, as previously mentioned, Neil thought it would be difficult even in his class of 6 students.

All the participants valued goals to work towards. They felt that no school system would be able to function on a test-when-ready basis. They liked the flexibility which on-demand testing offers, however, they felt that this needed to be planned:

> *“I think the current number actually focuses you down. You realise you’ve got a date and you work towards it.”* (Brian, Science teacher)

> *“I mean we have a pretty rigid scheme of work, because there's so much content that it's not a matter of 'they're ready this month or they're ready in two months’, I'm teaching every topic to a timetable. So, you know, I know exactly when I should have finished unit 1 and unit 2 and so on.”* (Claire, Science teacher)

> *“Schools work better with dates in diaries… And a bigger school, the cogs move a lot slower, I think there has to be, ‘we have assessment week’ and then if you want to do any assessments in your subject, it is in that week. Because it doesn’t work otherwise ‘cause you’ve planned to do this and, guess what…they’ve all gone to Colchester Zoo. So, yeah, flexibility is great, but it needs to be planned flexibility for it to work, because there are always lots of other things going on, as well.”* (Neil, deputy head teacher of special school)

4.4 Theme 4: The effect of learning factors on on-demand testing

The teachers did not talk about the effect of their students’ learning factors upon on-demand testing in as much detail as the students had in their discussions. However the teachers’ comments did fall under a distinct category, hence its inclusion as a theme. The participants considered variation in readiness, maturity of their students, and the potential benefits for other types of schools.

The teachers felt that there was little variation in their students’ ability; as it was a highly selective school, they felt that most were ready for their exams at the same time. They felt that an individualised learning approach would be of little benefit to their students, as they are all learning at approximately the same rate.

> *“But we teach a fairly narrow ability range here. And we keep them all, we have the same expectations broadly from everyone.”* (Claire, Science teacher)

They did, however, consider that on-demand testing could be of benefit in comprehensive schools and schools where streaming occurs. Neil thought that a test-when-ready approach would be beneficial to many of his students as they tend to plateau in performance and drift away from their study in the latter years of their education. However, Neil did consider this to be due to the troubled nature of his students rather than a characteristic of most GCSE students.

The teachers considered the benefits of on-demand testing for their students. They felt that some of the more able students may consider broadening their exam range but the majority would not be affected. Whilst they described their students as ‘self-starting’, some of the teachers obviously doubted their students’ ability to motivate themselves and organise their study independently.

5. SUMMARY AND CONCLUSIONS FROM TEACHERS AND STUDENTS

On the basis of these findings, it appears that, rather than desiring a purely on-demand system, teachers and students would prefer a modular system that simply provides more windows in which schools can enter candidates for exams than current modular specifications provide. Neither students nor teachers thought that a personalised approach to learning and assessment is currently feasible or desirable. They value goals to work towards, the support that working together as a class provides, and worry about the competitive element of individual exam entries. They liked the flexibility that on-demand testing offers but both groups felt it should be planned flexibility with dates entered in diaries early on in the process.

The comparability of multiple versions of on-demand tests was clearly a concern for students and teachers. In the event that on-demand tests are introduced, it will be necessary to persuade test users of the rigour of the test-equating techniques employed. Students considered on-demand testing to be at greater risk from cheating than existing exams, although teachers did not concur.

It may be worth repeating this study to include different types of centres such as comprehensive schools or further education institutions as they may have differing attitudes to the grammar school who took part in this project. The views presented, however, present an opportunity to stop and consider for whose benefit on-demand testing is being introduced and revaluate whether all the necessary support for such a far-reaching and complex change can actually be provided.

6. EXAMINER FINDINGS: DOMINANT THEMES

The examiners were presented with two hypothetical visions of on-demand testing in order to stimulate the discussion. In scenario 1, tests would contain a built-in ‘anchor’ of around ten items, which would be repeated across series in order to maintain standards. Multiple anchors would be in use, and the candidates would not know which items comprised the anchor. In scenario 2, tests would contain one or more randomly allocated pre-test items which would not be marked, but which would be re-used in a following session, serving as an anchor. Different candidates could be allocated different items.

6.1 Theme 1: The importance of pre-testing

All three of examiners agreed that the statistics supplied for pre-tested items were highly useful in deciding which items to include in a new test, as they highlighted those items that had performed poorly. Referring to his past experience as an examiner in the 70s during a period when pre-testing was the norm, one examiner commented:

> *“After we’d had pre-tests we’d have a review, and with statistics similar to yours but more visual, one could see exactly in most cases why a particular item within a set had performed badly and one could then do something about it. Items like that were then re-tested and then used…if you asked me for my ideal world I’d construct tests entirely from pre-tested…or reused items.”* (Mike, Science examiner)

The examiners without direct experience of constructing tests using pre-tested items also agreed that it “makes a lot of sense to do so” (Paul & Tony, Science examiners).

6.2 Theme 2: Implications of on-demand testing for item re-use, release and security

Pre-testing necessarily involves some re-use of items. Again, the examiners drew on their past experience of working with pre-tested items when discussing the implications of item re-use. One examiner commented that during the 70s, standard procedure was to wait three years before re-using an item, as allowing this period to elapse should ensure that the majority of candidates likely to re-sit a paper would have done so before the item was re-used:

> *"If you assume that candidates who took it in November 2007 would have until June 2009 [to re-sit] you would not repeat any items from November 2007 until after 2009…obviously you could have a candidate taking a re-sit…three years on but it doesn't matter about the odd one…so basically no candidate would see the same question twice."* (Mike, Science examiner)

The same examiner then went on to say that at one point in his experience, it was mandatory that between thirty and fifty percent of items on a paper were pre-used, in order to compare populations over time, a practice which the examiners all agreed was reasonable. Secure pre-testing in the 1970s facilitated this approach, with the caveat that certain assumptions have to be made about the sample of candidates being tested (Wilmut, 1975).

When asked about the release of items, the same examiner commented that items were released to the public during this period and that, as far as he could see, "there was no evidence that any candidate could have learned the answers to the questions to significantly affect performance" (Mike, Science examiner). The other examiners agreed that candidates "will still learn the science anyway; …they can't remember the sequence of letters" (Tony, Science examiner). A further point was made about the usefulness to teachers of mock exams, a sentiment also expressed by the teachers:

> *"The fact that tests are available to schools to use as mocks etc will increase your market share because teachers can use them and love using them…and that is a big consideration."* (Paul, Science examiner)

The examiners agreed that it would be very difficult at this point in time to return to a situation in which schools were unable to use past papers for mock examinations.

The question of item re-use and release leads on to that of security, and all of the examiners agreed that they saw no danger of an increase in security risks with the (re-)introduction of pre-testing. The examiners reported that in the 70s, teachers were allowed to look over the papers to see what direction the board was taking, but not permitted to remove the papers from the examination room, and that there was "no evidence at all that candidates gained from that". The examiners went on to emphasise the importance of placing trust in teachers:

> *"It doesn't matter what system you operate, if a teacher's determined to be a rogue…he or she will be a rogue…you've got to base any system on the fact that teachers are professionals, once you take away that assumption you might as well give up."* (Mike, Science examiner)

6.3 Theme 3: Other implications of on-demand testing

The discussion touched on two other more specific possible consequences of introducing on-demand testing, that of an increase in test length, and the implication of candidates within a cohort taking different items. An increase in test length is more of an issue in scenario 1, where an anchor of ten questions, or twenty percent of the test length is built into a test. The examiners felt that this was not a problem in the case of individual half-hour tests, however when multiplied for all six Science A tests it would "significantly increase the assessment time" (Tony, Science examiner). One examiner also pointed out that when the new Science specifications were written, QCA had stipulated that the assessment time must not be greater than that in the previous specification, and that "politically, with over-assessment at the moment, any move towards increasing the length of assessment time I think would be met with a big n-o" (Paul, Science examiner). It was also noted that the problem of increasing assessment length would be avoided by adopting scenario 2. When asked how they felt about candidates being allocated different pre-test questions, as in scenario 2, the examiners could not foresee any problems: "If it's only one item per test and if it's the only way we're gonna get pre-testing then I'd be happy…I can't see any real snags with doing that" (Mike, Science examiner).

7. SUMMARY AND CONCLUSIONS FROM EXAMINERS

The examiners featured in the focus group were all open to the idea of on-demand testing, particularly as some of them had experienced some elements, namely pre-testing, in the past. This meant that they had few concerns about the possible implications of introducing on-demand testing such as item re-use and release, and instead were able to offer an insight into how such issues could be, and indeed have been managed. They do represent, however, a limited sample of examiners who are numerate and who are used to working with objective test formats. Other examiners, even from within similar disciplines, may offer different perspectives.

E. THE PROVIDERS' PERSPECTIVE

1. INTRODUCTION

The purpose of this strand of the research project was to survey current practice in the UK in on-demand testing so that aspects of that practice which are relevant to the administration of on-demand tests for high stakes, national qualifications could be identified. In total nine organisations were surveyed (Table 2). With the notable exception of Cambridge ESOL, who declined to participate, they represent the testing agencies offering the most advanced models of on-demand testing in the UK. Five organisations responded to a questionnaire, three test providers took part in face-to-face interviews and one participated in a telephone interview. The structure of the interviews was based on that of the questionnaire (see Appendix 5) but allowed for more in-depth questioning in the areas of interest (standard-setting, generation and use of statistical information, and security in particular). Data from the interviews and questionnaires have been supplemented by information provided on websites and in ancillary documentation from the participating organisations.

2. PROFILES OF ORGANISATIONS

2.1 Universities Medical Assessment Partnership (UMAP)

UMAP does not fall within the remit of Ofqual because it is not an awarding body and operates only within the higher education sector, specifically with medical schools. It also fits the profile of an item broker (Chelu & Elton, 1977; Sclater, 2004). It provided either all or some of the items for 50 summative assessments delivered by its 15 partner medical schools in the 2007 – 2008 academic year. Assessments are compiled and administered by partner schools. As an example of scale, one partner school provides tests on one day to all of its year groups, giving a total of 25,000 entries. Partner schools may draw out items from the bank at any time, although May, June and December have the greatest number of withdrawals.

The brokerage system operates by charging partner schools a fee and expecting each partner to equitably share the burden of item writing and reviewing. UMAP manages an intensive annual round of item writing workshops and item review sessions which act as quality assurance of the approved items in the bank. The bank contains currently approximately 7,000 items. UMAP aims to increase this to 15,000 to meet the learning outcomes set out in *Tomorrow's Doctors* by the General Medical Council (General Medical Council, 2003).

Partner schools may draw from the bank up to 200 extended matching question items (EMQs) and up to 250 multiple-choice question items (MCQs) annually. They are expected to send a copy of the assessment(s), candidates' item level responses and item level scores to the broker after the assessment(s) has been administered. These latter two data transfers are electronic, as is the item bank. However, partner schools usually deliver assessments in the paper-based mode.

UMAP has been self-funding since January 2006 after starting up with a grant from the Higher Education Funding Council for England (HEFCE). It places a premium on developing items that possess high levels of validity within the medical context.

Table 2: Current position in the on-demand testing arena of those organisations that participated in the survey.

	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 6 months
Scottish Qualifications Authority (SQA)	×	✓	✓	Mid	Selection to higher education & employment	800 approx.	Continuous	Not applicable
Ifs School of Finance	✓	×	✓	High/Low	License to practice within industry sector/life skills	40 - 50	Continuous	1 day minimum
CEM Centre	×	×	✓	Low	Formative & diagnostic assessment in primary & secondary schools	8	Variable: 1 per year of 7 months to continuous	None to 1 month
EDI	✓	×	✓	Mid	Diagnostic & summative	350 approx	Continuous	4 weeks
Driving Standards Agency	×	×	✓	Low	License to practice a life skill	11	Continuous	10 day minimum
Organisation I	×	×	✓	High	License to practice within industry sector	2	Test 1: 6 Test 2: 1 of 30 days	Test 1: 1 week Test 2: 4 weeks
Organisation II	✓	×	✓	Mid/High	License to practice within industry sector	20	5 days per week, except 2 shut down periods	7 days to 3 months

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

2.2 City & Guilds

A pioneer of e-migration in the UK since the turn of the century, City & Guilds offers internationally a wide range of qualifications from entry level up to level 8 in the Vocational Qualifications arena. Slightly more than 20% of the assessments on offer from this provider are on-screen, on-demand. Due to the highly specialist nature of some of the subjects offered by this awarding body under its charter of social responsibility the number of candidates entering for some assessments can be in single figures, others in their hundreds/thousands. In the first year of offering on-demand tests online, City & Guilds tested 65,000 candidates; by the start of the sixth year of on-demand provision this number had increased by a factor of 13 to 862,000 candidates. According to published information there are, on average, 16,000 candidates per week logging in to City & Guilds' on-demand test system.

The on-demand tests are available through registered centres and a registered centre may be a college, a workplace or an e-assessor with a laptop and mobile access (offline testing). Once a candidate is registered for a qualification with this provider a test may be scheduled for the candidate up to six months in advance or instantly. The candidate may sit the test at any time within an eight hour window centred around the time scheduled by the centre. Tests may be scheduled for any time or day providing an invigilator is present.

The item bank system that supports the on-demand tests is hosted by a third party supplier. In partnership with other third party suppliers City & Guilds is developing additional assessment packages for the e-environment. These include a learning support portal and an e-portfolio platform; a strategy with the potential to provide a modular solution to the learning and assessment needs of various providers of vocational education and training. There is an emphasis on partnerships, both technological and educational, within this organisation.

2.3 Scottish Qualifications Authority (SQA)

The Scottish Qualifications Authority (SQA) offers a wide range of qualifications in a number of subjects to students in schools, colleges and the workplace. Approximately 1,500 registered centres deliver the thousands of units that make up these qualifications. In starting the process of embedding a culture of on-demand testing into the centres it serves, this assessment provider is working initially with colleges that deliver its vocational qualifications. The reasons for this are (i) the ICT infrastructure in colleges tends to be more robust than in schools; and, (ii) the qualifications, whilst still serving as measures for selection to higher education and employment, are not regarded as high stakes in the national assessment context. Thus, items in the banks for each of the approximately 800 units available for on-demand tests may be used for formative, diagnostic and summative purposes.

The assessment environment in which this test provider operates grants colleges independence in the creation, delivery and marking of assessments; processes which SQA moderates. In creating e-assessments utilising SQA's online item bank system teaching staff may choose to use only items from the provider's item banks, only items that they themselves have written or to mix the provider's items with their own. The advantage to using items from the provider's item banks is that these have been quality assured by SQA through its development and validation processes. Training in the development of e-assessments is offered to teaching staff by SQA, which is then able to populate its item banks by commissioning these subject teachers to write items. Gradually, with greater use of the item banks, the focus will shift from the post-assessment moderation to the item writing and test construction that take place prior to assessment. This change in focus will allow SQA to standardise assessments within its vocational qualifications whilst maintaining colleges' independence in the setting of assessments.

2.4 Ifs School of Finance

Ifs School of Finance is a professional body regulated by Ofqual. The organisation's aim is to educate professionals in the financial services sector and the general public in personal finance. To facilitate this aim it administers regulatory and personal finance qualifications at levels 1 to 3 in the National Qualifications Framework plus degrees through the universities it has developed educational partnerships with. In the academic year 2006 - 2007 the personal finance qualifications were delivered by 250 schools and colleges to over 11,000 students. Ifs also provides learning support for its qualifications through a web portal.

Of the 21 regulatory and personal finance qualifications, 18 may be assessed entirely in the e-environment. Across the range of units that constitute the qualifications between 40 and 50 units are delivered on-screen, on-demand. Most tests are created and delivered electronically through a third party supplier's delivery system and network of test centres. In addition to the qualifications, the organisation also delivers tailored awards to companies in the financial sector using this delivery method. Some of the non-regulatory tests are delivered to centres through the organisation's bespoke item bank system that was developed in-house.

Ifs School of Finance emphasises the development, writing and validation of the items that make up its tests. To this end it manages a number of subject experts external to the organisation to ensure item banks are populated with validated items and that tests are constructed to published specifications which are based on qualitative criteria.

2.5 CEM Centre (CEM)

Centre for Evaluation and Monitoring, based in Durham University in North-East England. CEM provides indicator systems to schools and colleges; the confidentiality of these systems renders them unique. Established in 1983, the Centre works with schools, colleges, education authorities and government agencies to provide high quality information through scientifically grounded research. CEM is the home of a widely used family of monitoring systems including ALIS, Yellis, MidYIS and PIPS.

2.6 EDI

EDI is an accredited Awarding Body and leading international education company with a wide range of products and services including vocational and professional qualifications both within the UK and internationally through LCCI, Goal online assessments for schools, approved training programmes for employers, an electronic assessment delivery system, electronic portfolio package and specialist business broadband service.

2.7 The Driving Standards Agency (DSA)

DSA's vision is "Safe Driving for Life". Their overall mission is to contribute to the public service agreement objective to achieve 40% reduction in riders and drivers killed or seriously injured in road accidents, in the age group up to 24 years, by 2010 compared with the average for 1994-98. DSA employs 2,653 staff, of which some 1,911 are driving examiners. In 2006/07 the Agency conducted over 1.8 million practical tests for car drivers, 101,000 vocational tests and over 83,000 motorcycle rider tests. Over 1.5 million theory tests were carried out at 158 centres. At the end of the year there were 41,507 people on the Register of Approved Driving Instructors.

3. CURRENT PRACTICE IN ON-DEMAND TESTING

3.1 The infrastructure of on-demand testing

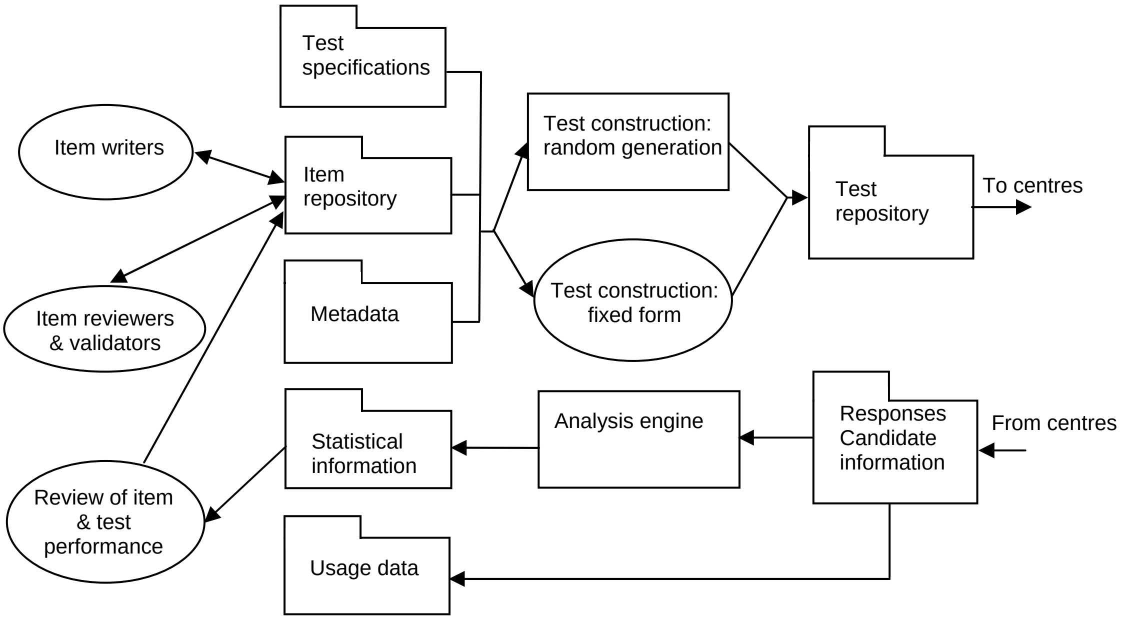

Test specifications
Item writers
Item repository
Item reviewers & validators
Metadata
Test construction: random generation
Test construction: fixed form
Test repository
To centres
Analysis engine
Responses Candidate information
From centres
Statistical information
Review of item & test performance
Usage data

Figure 7: Schematic diagram of a common item bank system

Most of the item bank systems considered in the survey have a similar design based around a relational database that stores the items; see Figure 7. In addition there are storage facilities for metadata, statistical information, usage data and test specifications. Item writing may be supported by an online authoring module, but just as often item writers deliver items to subject managers who key the items into the database of the item repository. Modules or processes of test construction utilise the item repository, metadata and test specifications, but rarely do they use statistical information or usage data. Tests are stored in a test repository prior to delivery to centres. Three topics emerged as being of major interest in terms of managing an item bank and constructing tests for on-demand delivery. These topics are standard setting and maintenance, items and tests, and technology.

3.2 Standard setting and maintenance

With one exception, grade boundaries are set in advance using an Angoff procedure or strong criterion referencing. Only CEM equates tests to ensure equivalence; applying the Rasch model to its adaptive tests. All of the other participating organisations rely on front-end processes of development, item writing and validation (or reviewing) to quality assure items.

"Where we tend to [ensure forms are parallel] is kind of front-end as much as we can.

We do all the analysis and research and evaluation up front and then we inform that by the statistical analysis of how these items perform within forms so then we can review and validate the work that was done up front.

But it's the experience of item writing and, and subject experts." (Interview 2)

There was awareness of IRT models in other organisations, and SQA were clear that, although they do not use IRT for the mid to low stakes tests they currently offer on-demand, it would offer a robust and transparent method of equating between tests, determining grade boundaries and calibrating questions for use in item banks for their high-stakes examinations.

Other than the adaptive tests of CEM, the construction of forms offered by the participating organisations is not based in any part on statistical information. This is despite the fact that the test delivery systems used all have the ability to store and transfer candidates' item level responses and there is thus a large quantity of raw data available for analysis using IRT.

"… but the thing is there's rafts of stuff that [the item-banking system] does record, but it's a separate issue about whether we report it or use it…….. The fact that we don't use it within reporting or as rules to deliver assessment, is probably the best way of putting that." (Interview 3)

There appear to be two reasons why IRT is not applied to the raw data generated by on-screen on-demand testing. Firstly, the expertise in IRT located in the UK does not appear to have reached the critical mass necessary for IRT analyses to be routinely applied in operational tests. Two organisations commented on the need for "upskilling of staff" in this area. Secondly, there is a perception that for those tests where entry numbers are low, an IRT approach would not be viable.

"Our numbers are pretty small and IRT would prove challenging." (Questionnaire 3)

In these situations IRT may not be useful in post-hoc evaluations, but there is no reason why an equating design could not establish a priori comparability given the right experimental design.

Two organisations routinely pre-test their on-screen on-demand items. CEM randomly seeds items in specified positions within their adaptive tests to determine difficulty and item quality parameters. These items do not contribute to the candidates' scores. One other provider, working in a high-stakes environment, pre-tests new items in their live tests. On logging on to the test platform the candidates involved (a high proportion of the total cohort) are informed that they may be asked to pre-test new questions. After completing the live test they are asked whether they wish to continue to the pre-test questions, which will not contribute to their final score. The pre-test data is quality controlled, and the facility and discrimination indices of the items analysed to inform inclusion in future tests.

The advice offered internally at one other provider is to pre-test new items, but this is not always possible to accomplish as the writing of items is part of the development of a specification, and on many occasions there was insufficient time to attend to this task. One provider that already pre-tests new items for those tests offered in a paper-based mode intends to insert a block of new items at the end of operational tests to pre-test in its on-demand tests.

3.3 Items and tests

The item banks in this review contain a variety of formats including multiple-choice items, matching questions, True/False and short free text. The more experienced providers of on-demand tests are designing and piloting more innovative item types. For example, City & Guilds is piloting the use of audio files in its English for Speakers of Other Languages (ESOL) tests. SQA appears to be some way ahead of other providers in that it prepares a full range of item types including short and long free text, sequencing, pie and bar chart creation and items that require candidates to enter mathematical formulae. Only the Driving Standards Agency uses simulations as part of its on-screen on-demand tests.

Participants reported varying population levels in their item banks from 50 per level [of test] to 20,000 online items for all on-demand tests offered by that provider. The ratio of banked items to number of items in an operational test (where the link could be clearly identified) varied between 4, as the minimum recommended before implementing a new specification, and 39. Most ratios were below 25, but only one organisation had a recommended operational ratio of 8. The rate at which approved new items are banked varies between organisations, as do the reasons for preparing new items. For those item banks for which values were available, replenishment occurs annually at between 4% and 20% of the bank population. The reasons given for replenishment were: to reach an optimum recommended ratio of banked items to number of items in test; to extend coverage of relevant topic areas; and to replace items that were retired.

The writing, validation and reviewing of items are part of the front-end processes which start with the development of a specification and continue throughout the lifetime of that specification with varying levels of intensity. These processes serve as quality assurance of the items. All of the organisations interviewed reported sophisticated procedures for training item writers and then either accepting or commissioning items from them based on the qualitative requirements of a specification, i.e. assessment objectives, topic areas and, sometimes, level of demand. Once accepted in draft form into an item bank the items begin a validation process. Currently, the validation process underpins the standards set in on-demand tests, whilst also checking for face and content validity. It tends to be a manual batch process, relying heavily on the involvement of subject experts to check, for example, that an item is appropriate for the specification, assesses the skills or knowledge it needs to, is correctly written, has a correct response, has an explanation for the purposes of feedback, and, in some cases, to assign estimated values for the item parameters.

Once items have been approved for use in on-demand tests, they are compiled into forms. With the exception of CEM's adaptive tests, there appear to be two methods used to create forms: fixed forms and random generation of forms. Fixed forms tend to be created as a batch of, quite commonly, four forms by one or more subject specialists. They search the item bank using metadata tags for appropriate items and select items for individual forms. The selection of items is made according to the 'rules of test construction' set out in the specification. Such specifications stipulate the number of items that will appear in a form for each topic area or learning objective, the maximum mark and therefore the maximum number of items. The aim is to include all of the content of a specification in any given form. Another round of validation takes place for the tests and again this is mediated by subject experts. After which the items are no longer just in banks but in forms which are effectively in a bank of forms waiting to be allocated randomly to a candidate as and when the centre registers the candidate for that particular test.

"…it is using the subject experts, the experts to create that form … using the technology to help them."
"… we tend to do it as one job because again that's all about feeding into the, getting the, the standards between them more similar than if we're doing them six months apart." (Interview 2)

"Four fixed forms are available at any given time. These forms are constructed and launched at the same time. New items are written by trained industry experts, edited by panel. A test is constructed and reviewed by panel. The forms go through a standard setting meeting together. Higher level and item level statistics are reviewed on a quarterly basis and compared to historical data." (Questionnaire 3)

The alternative procedure in use is the random generation of forms. Forms are created by an algorithm which uses rules embodied in the specification. This is done when a centre registers a candidate for a test, no matter how far in advance the test is scheduled. The item bank system records the item ids required for the randomly generated form. Just before the candidate sits the test, the form is compiled. Currently, the algorithm for form creation does not use any statistical information to perform its task. Nor does the creation of forms generally take account of the previous exposures of items, which may lead to candidates re-sitting forms containing one of more items they have responded to in earlier attempts. This is another example of item-banking systems recording and storing a lot of data (which items have been seen by which candidates), but not converting them into information that may be used in the form creation process.

In an on-demand testing system the item level responses of candidates are being uploaded continually. These data are generally analysed using CTT and the item parameters facility and discrimination are routinely reviewed with a frequency that varies between quarterly, six monthly and annually. The reviews consider the performance of items in live tests. Those items that do not perform statistically as expected, that is, they deviated from the opinion of the subject experts, are assessed as to whether they should remain in the item bank or not. This represents the item retirement strategy of most organisations. One interviewed organisation's item bank system automatically checks tests for whether a retired item is part of a test. If it is, an alert is issued for human intervention in making a decision as to whether a new item should be commissioned or if existing ones are appropriate for substitution for the retired item.

There is an enormous resource available in terms of knowledge and experience of the writing and validation of items to populate banks for one or more specifications. This resource is comprised primarily of subject experts and subject managers. However, there is a dearth of expertise in the use of statistical information to construct tests or forms that are of a quantitatively measured level of difficulty.

3.4 Technology

The technology for the operation of item bank systems is available and maturing. The technology in use ranges from bespoke, in-house solutions involving Access databases and Excel spreadsheets to integrated modules each performing an item-banking function (item authoring, item storage, metadata storage, test construction, test delivery, storage of responses, analysis). From this survey it may be concluded that there is no one-size-fits-all technology solution to item-banking. Whether developed in-house or provided by a third party, most participants were confident that their systems comply with the IMS Global Consortium's Question and Test Interoperability (QTI) specification in terms of items being transferrable between different item banks and delivery systems. However, there was less awareness of standards for the transfer of metadata.

Eight of the nine participating organisations use third party suppliers or technology partners to provide at least one part of their item-banking system that supports on-demand tests. The minimum level of involvement is to use a third party supplier's test delivery system (this also includes immediate scoring and the return of candidate level item responses) and network of test centres. At the highest level of involvement the third party supplier either hosts all modules of the item-banking system or the on-demand test providing organisation licenses a fully integrated item-banking system. Two technology partners, BTL and Pearson VUE, provide one or more modules of the item-banking systems to eight of the studied on-demand test providers.

3.5 Security

Traditional invigilation procedures are used by all the providers, with guidance offered on issues such as the minimum distance between computers for example. Most test delivery platforms ensure that all other software applications are locked during the test.

4. CONCLUSIONS

The scale of the on-demand testing operations that currently exist in the UK is impressive, and the technology clearly developing. Technology partnerships seem to be a successful and preferred method of delivering on-demand solutions, although there does appear to be a limited number of technology partners available. These partnerships are delivering high volumes of on-screen on-demand tests and increasing the validity of these tests through provision of more realistic assessment tasks.

While the technology is developing, however, there seems to be little innovation in assessment models which could enhance the rigour of on-demand testing. Only two agencies seem to be taking advantage of the increased efficiency in pre-testing procedures that on-screen assessment offers, and only one of those is using that information to ensure the comparability of test versions. The traditional skills of the subject experts and the assessment administrators have been adapted to operate in an on-screen on-demand environment, but these traditional skills have not yet been supplemented by psychometric skills. This is clearly of concern to some of the organisations surveyed, and at least one is actively trying to train its research staff in this area. This situation is all the more vexing given the wealth of data that on-demand testing is producing.

F. PRINCIPLES OF ON-DEMAND TESTING

From the literature review, the findings from the visits to current providers of on-demand testing, and the focus groups with teachers, students and examiners a set of draft principles were drawn up. These were then presented to a group of five technical experts in a focus group. The participants comprised technical experts with many years of working within UK awarding bodies, psychometricians with experience of on-demand systems abroad, and a retired deputy head teacher. Following their feedback the principles were revised, but given time restraints on the project, the participants have not been given a chance to comment on the revised principles. For this reason the principles represent the views of the research team, although only those principles on which there was general consensus from the group have been included.

1. EXAMINATION STANDARDS

i. Decisions to move each syllabus to on-demand testing should be supported by a clear educational case. This case should have a sound theoretical basis and be supported by the teaching profession.

ii. On-demand testing should be underpinned by Item Response Theory methods of test-equating.

iii. Policies on item to test ratio, item re-use, pre-test procedures and evidence of coherence of scales should all be available.

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.

v. Systems should be in place to monitor and help explain changes in aggregate qualification outcomes over time.

vi. The reliability of tests should be such that there is little to gain from repeated re-sitting.

2. ACCESSIBILITY

vii. On-screen on-demand tests should provide greater accessibility than paper based tests through the use of assistive computer technology.

viii. Items in item-banks should be tagged according to accessibility requirements so that alternative items which cover the same area and test the same skills can be provided.

3. THE BURDEN OF ASSESSMENT

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.

x. Changes in the burden of assessment in the educational system as a whole, including additional pressures on teachers and candidates, should be monitored.

4. COMMUNICATION

xi. All stakeholders, including candidates, should be actively consulted during the redefinition of processes to support on-demand testing.

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.

G. CONCLUSION

This report has considered what is known about the issues that will arise in any move to providing high-stakes national assessments on-demand; stakeholder views on this proposed change; and the current state of practice in this field in other sectors of assessment in the UK. The literature review certainly concludes that there are no insurmountable technical reasons why a high stakes qualification in the UK cannot be delivered on-demand, although this does not underestimate the complexity and the investment that will be required to deliver a seamless end-to-end assessment system which provides timely high quality information. The movement to on-screen, on-demand testing may be slow at first, and limited in scope, but as technology partners work with awarding bodies there is every reason to believe that the scope and range will increase.

The visits to current providers show that technology is currently being harnessed to deliver hundreds of thousands of on-screen on-demand tests within the UK and that assessment agencies are learning to work successfully with technology partners to deliver flexible models of assessment. With the exception of the innovative adaptive testing occurring at one agency, however, the assessment models that are being delivered have not evolved. The emphasis on face validity and post-hoc evaluation is not a model that would serve high-stakes testing well. It is not worth entertaining a situation in which awarding bodies deliver on-demand tests without a sound knowledge of the statistical properties of those tests. Comparability between versions is a key concern of stakeholders, and the examiners would welcome a return to pre-testing. There are psychometricians in the UK working in health and optometry, and there are certainly talented researchers working in the awarding bodies. Their expertise needs to be harnessed in developing models that will deliver tests that are guaranteed to be reliable and valid before they are delivered to candidates.

The views of teachers and students highlight the extent of the challenge that lies ahead if assessment is to support the personalised learning of 2020. If the deputy head teacher of the special school with a class size of 6 cannot envisage how the system will work it may be time to return to the drawing board. Nevertheless, the prospect of any accredited assessor with a networked laptop being able to conduct an assessment at any time, which is the reality for one testing agency in the UK, does open up new possibilities for inclusion. Given such an opportunity it would surely not be long before innovative schools and colleges started to develop new models of teaching and learning that could exploit this flexibility. The competitive market for assessment provision in the UK can only be of benefit in this regard, rewarding those awarding bodies who work closely with schools and colleges to develop innovative products.

Finally, the regulator needs to be aware of the steps awarding bodies are taking towards on-demand testing. Awarding bodies making this move are likely to be asking themselves the following questions:

1. If a technology partner is going to be employed, will the procurement process ensure that this partner is committed to ensuring the integrity of the assessment system?
2. Is the systems architecture robust and secure?
3. Has accessibility been explicitly considered at the design stage?
4. Does the integration and migration plan ensure a seamless transition between or integration of legacy and new systems?

5. Has each module of the architecture been tested, and do all modules work together? What are the contingency plans for failure at any point in the process? How secure is each stage of the process?
6. Are all support systems in place? To what extent are the support systems coherent with current practice?

At any one of these stages the awarding bodies may seek reassurance from the regulator that what is proposed will meet its Code of Practice (QCA, 2008). For each product that is then developed for on-demand testing the following issues arise:

1. What is the educational case for providing this syllabus on-demand? Is there good theoretical evidence and backing from classroom practitioners that the provision of the assessment on-demand will deliver educational benefit?
2. Is the test-equating model that is being proposed feasible? Can the tests that are being equated be mapped onto coherent scales?
3. What is the item to test ratio being proposed for the syllabus? What are the policies on item re-use and pre-testing?
4. What is the reliability of the tests that are being delivered?
5. What systems are in place to ensure that the equating models used deliver stable and defensible qualification outcomes?

If good evidence is provided on these issues then it is unlikely that the awarding bodies or their regulator, or the multitudes of educational practitioners who are dependent to some extent on the assessment system, will be caught unawares by unintended consequences of the move to on-demand testing.

Acknowledgements

The researchers would like to thank Anton Béguin of CITO for his insightful comments as Project Reviewer.

H. REFERENCES

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research, 42*, 145-170.

Baird, J.-A. (2007). Alternative conceptions of comparability. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Beaton, A. E., & Zwick, R. (1990). *The Effect of Changes in the National Assessment: Disentangling the NAEP Reading Anomaly*. Princeton: National Assessment of Educational Progress.

Beguin, A. A. (2000). *Robustness of Equating High-Stakes Tests.* University of Twente.

Bejar I, I. (1983). *Achievement Testing: Recent Advances*. Beverley Hills: CA: Sage.

Bennett, J. (2008). *Assessment and the challenges of personalisation*. Paper presented at the Westminster Education Forum Keynote Seminar: Testing Times – what is the future for assessment and learning?

Bennett, R. E. (2001). How the Internet Will Help Large-Scale Assessment Reinvent Itself. *Education Policy Analysis Archives 9*(5).

Black, P. J. (2007). *Can we design a supportive assessment system?* Paper presented at the Chartered Institute of Educational Assessors.

Black, P. J., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for Learning: Putting it Into Practice*: McGraw-Hill International.

Black, P. J., & Wiliam, D. (1998). *Inside the black box : raising standards through classroom assessment*. London: King's College.

Boyle, A. (2006). *Evaluation of the 2006 pilot of the key stage 3 ICT test*: Qualifications and Curriculum Authority.

Bramley, T. (2006). *Equating methods used in KS3 Science and English*. Paper presented at the NAA technical seminar. from http://www.cambridgeassessment.org.uk/ca/digitalAssets/114007_Equating_methods_KS3_TB.pdf.

Brown-Martin, G. (2008). Home Access - Who Benefits? Retrieved 4th November, 2008, from http://www.handheldlearning.co.uk/content/view/57/ (Archived by WebCite® at http://www.webcitation.org/5c6Mf9Ua2)

Brown, M. (1989). Graded Assessment and Learning Hierarchies in Mathematics: An Alternative View. *British Educational Research Journal, 15*(2), 121-128.

Chamberlain, S. (2008). Sleeping in or Selecting Out? Candidates' Absences from GCSE examinations. *Research in Education, 79*(1), 53-66.

Chartered Institute of Educational Assessors. (2008). Retrieved 4th November, 2008, from http://www.ciea.org.uk/news_and_events/press_releases/Top_UK_Universities_to_offer_masters.aspx (Archived by WebCite® at http://www.webcitation.org/5c6LA8GYf)

Chelu, C. J., & Elton, L. R. B. (1977) An item bank for multiple-choice questions. Physics Education, *5*, 263-266.

City & Guilds. (n.d.). Global online assessments available 24/7. Retrieved 2008/12/17, 2008, from http://www.cityandguilds.com/cps/rde/xchg/SID-A0E63ED9-21D58322/cgonline/hs.xsl/2478.html

Cross, R. (2004) Review of existing item banks. In N. Sclater (Ed.), IBIS Item Banks Infrastructure Study (pp. 17-34): Joint Information Systems Committee (JISC), Higher Education Funding Council for England (HEFCE).

Deci, E. L. (1975). *Intrinsic Motivation*. New York: Plenum Press.

Department for Education and Skills. (2005). *Improving the HE Applications Process: Equality Impact Assessment*.

Department for Education and Skills. (1988). *Task Group on Assessment and Testing: A Report*. London: Department of Education and Science and the Welsh Office.

Department for Education and Skills. (2007). *Making Good Progress*. London.

ETS. (2006). Revised GRE General Test. Retrieved 15th November 2006, from http://www.ets.org

Frantz, D. & Nordheimer, J. (1997, September 28). Giant of Exam Business Keeps Quiet on Cheating. *New York Times.* Retrieved January 26, 2009. URL:http://query.nytimes.com/gst/fullpage.html?res=9400E3D6153AF93BA1575AC0A961958260&sec=&spon=&pagewanted=7. Accessed: 2009-01-26. (Archived by WebCite® at http://www.webcitation.org/5e78yqhcW)

General Medical Council. (2003). *Tomorrow's doctors*. London and Manchester: General Medical Council (GMC).

Gilbert, C. (2006). *2020 Vision: Report of the Teaching and Learning in 2020 Review Group*. Nottingham: Department for Education and Skills.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42*, 139-167.

Good, F. J., & Cresswell, M. J. (1988). *Grading the GCSE*. London: Secondary Examinations Council.

Green, B. F. J. (1983). Notes on the efficacy of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of Modern Psychological Measurement*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Guilford, J. P. (1954). *Psychometric Methods*. New York: McGraw Hill.

Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of Item Response Theory*. Newbury Park, California: Sage.

Harris. (1993). *Practical Issues in Equating*. Paper presented at the The Annual Meeting of the American Educational Research Association.

Hayes, M. (2008). *Equating Key Stage Tests*. Paper presented at the Third UK Rasch Day.

He, Q. (2008). *Using the Rasch Model to Analyse Dichotomous and Polytomous Items in GCSE Grading*. Guildford: Assessment And Qualifications Alliance.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking*. New York: Springer.

Linden, W. J. v. d. (2005). *Linear Models of Optimal Test Design*. New York: Springer.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.

Luecht, R., Brumfield, T., & Breithaupt, K. (2006). A Testlet Assembly Design for Adaptive Multistage Tests. *Applied Measurement in Education, 19*(3), 189-202.

McAlpine, M., & Zanden, L. v. d. (2006). *Itembanking infrastructure: a proposal for a decoupled architecture.* Paper presented at the 10th CAA International Computer Assisted Assessment Conference, Loughborough University.

McCaig, C (2003) The Politics of Exam Crises, paper to the Conference of the International Association for Educational Assessment, Manchester, 9 October 2003.

McLeod, L. D., & Schnipke, D. L. (1999). *Detecting items that have been memorized in the computerized adaptive testing environment.* Paper presented at the Annual Meeting of National Council on Measurement in Education, Montreal, Canada.

Mead, A. D. (2006). An Introduction to Multistage Testing. *Applied Measurement in Education, 19*(3), 185-187.

Noss, R., Goldstein, H., & Hoyles, C. (1989). Graded Assessment and Learning Hierarchies in Mathematics. *British Educational Research Journal, 15*(2), 109-120.

One Laptop per Child. (2008). One Laptop per Child. Retrieved 4th November, 2008, from http://laptop.org/en/laptop/software/index.shtml (Archived by WebCite® at http://www.webcitation.org/5c4sPuzfn)

Panayides, P., Robinson, C., & Tymms, P. (In Press). The Assessment Revolution that has passed England by: Rasch Measurement. *British Educational Research Journal*.

Parkinson, J. (2008). Pledge Watch: Laptops for all. Retrieved 4th November, 2008, from http://news.bbc.co.uk/1/hi/uk_politics/7686357.stm (Archived by WebCite® at http://www.webcitation.org/5c6LsedhP)

Pearson Driving Assessments Ltd. (2008). *Partnering with you for results that count*. Retrieved 2008/12/17, 2008, from http://www.pearsonvue.co.uk/Pages/default.aspx

Qualifications and Curriculum Authority (2008). *Code of Practice.*

Richardson, M., Baird, J.-A., Ridgeway, J., Ripley, M., Shorrocks Taylor, D., & Swan, M. (2002). Challenging minds? Students' perceptions of computer-based World Class Tests of problem solving. *Computers in human behaviour, 18*(6), 633-649.

Ritter, E. M., McClusky, D. A., Gallagher, A. G., & Smith, C. D. (2005). Real-Time Objective Assessment of Knot Quality with a Portable Tensiometer is Superior to Execution Time for Assessment of Laparoscopic Knot-Tying Performance. *SURGICAL INNOVATION, 12*(3), 233-238.

Sclater, N. (Ed.). (2004). *IBIS Item Banks Infrastructure Study*: Joint Information Systems Council, Higher Education Funding Council for England (HEFCE).

Shorrocks-Taylor, D., Curry, J., Swinnerton, B., & Nelson, N. (2003). National Curriculum Mathematics Tests in England at Key Stage 2: Weights and Measures? *Oxford Review of Education, 29*(1), 51-66.

Squire, J., Owen, A., Baines, D., & Byrne, G. (2007). *UK Collaboration for a Digital Repository: Final Report*: University of Manchester.

Stringer, N. (2008). Aptitude tests versus school exams as selection tools for higher education and the case for assessing educational achievement in context. *Research Papers in Education, 23*(1), 53-68.

Lord Sutherland of Houndwood. (2008). *The Sutherland Inquiry: An independent Inquiry into the delivery of National Curriculum tests in 2008*. London: The Stationery Office.

Thyne, J. M. (1974). *Principles of examining*. London: University of London Press.

Ufi Ltd. (n.d.). *Life in the UK tests*. Retrieved 2008/12/17, 2008, from http://www.ufi.com/home/section1/6_projects/lifeintheuk.asp

Wainer, H. (2000a). *CATs: Whither and whence*: Educational Testing Service.

Wainer, H. (2000b). *Computerized Adaptive Testing*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ward, A. W., & Murray-Ward, M. (1994). *Guidelines for the Development of Item Banks*: An NCME Instructional Module.

Westwood, J. D. (2007). *Medicine Meets Virtual Reality 15*. Amsterdam: IOS Press.

Wheadon, C. B., & Beguin, A. A. (In Press). *Fear for tiers: are candidates being appropriately rewarded for their performance on tiered examinations?* Assessment in Education.

Whitehouse, C., He, Q. & Wheadon, C. (2008). *Item banking with a test construction interface: an evaluation of a prototype*. Assessment And Qualifications Alliance: Guildford. RPA_08_CW_RP_064b

Whitelock, D. (2006). *Roadmap for e-assessment*: The Joint Information Systems Committee.

Willingham, W. W. (1980). *New methods and directions in achievement measurement.* Paper presented at the New directions for testing and measurement, ETC invitational conference on testing problems.

Wilmut, J. (1975). *Setting Objective Test Items*. Aldershot: The Associated Examining Board.

Wurster, T. S., & Evans, P. (2000). *Blown to Bits: How the New Economics of Information Transforms Strategy*. Harvard: Harvard Business School Press.

Young, R., MacNeill, S., Adams, D., & McAlpine, M. (2005). SPAID: Storage and Packaging of Assessment Item Data. Retrieved 4th November, 2008, from http://www.jisc.ac.uk/whatwedo/programmes/elearningframework/toolkitstrath1.aspx (Archived by WebCite® at http://www.webcitation.org/5c4qYxOu0)

Yunus, A., Kasa, Z., Asmuni, A., Samah, B., Napis, S., Yusoff, M., et al. (2006). Use of webcasting technology in teaching higher education. *International Education Journal, 7*(7), 916-923.

Zwick, R. (1991). Effects of Item Order and Context on Estimation of NAEP Reading Proficiency. *Educational Measurement: Issues and Practice, 10*(3), 10-16.

APPENDIX 1

STUDENT FOCUS GROUP SESSION PLAN

Stage 1 and 2a Topical areas for Group Discussion

Introduction

- State who we are
- State purpose of study – for Ofqual report
- Briefly describe on-demand testing
- State ground rules of discussion – everyone talk, interested in all opinions, don't talk over anyone
- Get everyone to say their name (makes transcription easier)

Stimulus material

Views on on-demand testing

- What do you think is good about on-demand testing?
- What do you think is bad about on-demand testing?

General Topical areas for more detailed discussion

Test Pressure

- Frequent vs. end of year?
- Re-sits?
- Who chooses when student is ready?
- Parents will force children to take examinations
- Students will be over pressurised by the new regime

Re-Sits

- How many?
- Should there be a limit?

Revision

- Do you prefer to revise in groups?
- What if only a few of you were taking the test?
- What if there were no past papers?
- If you could would you help each other in the test?

Specific Topical areas for more detailed discussion

End of year vs. on-demand

- Is it easier to remember a topic just after you have learnt it?
- Do you think you will have acquired all the skills you need if you take the test earlier?
- Are there any subjects that you don't think this will work for?
- What would you prefer?
- When would you choose?

Different papers for different people

- How would you feel about taking a unique test?
- Are you worried that the test your friend takes may be easier than the one you take?

Location

- Would you be happy taking your test in test centres?
- Would you prefer to take your test in a room at school?
- Would you prefer there to be 'test days' e.g. every Friday?

Security

- Do you think that some students may cheat? How do you think they could do this?
- Are you concerned that your paper could be lost?

Tailored route through education

- If you could take tests at any time how do you think school systems might work?
- How would you feel about choosing when you studied your different subjects and when you took your tests?
- If you have the choice, would you prefer individual study or classroom study with everyone?
- More able students will want to broaden their examination range
- More students will seek extra tuition from commercial tutorial companies

(*Extra questions from JISC)*

Students will be over pressurised by the new regime
Parents will force children to take examinations
More students will seek extra tuition from commercial tutorial companies
More able students will want to broaden their examination range

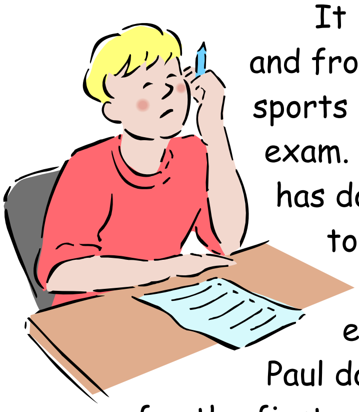
APPENDIX 2

UNIVERSITY STUDENTS' FOCUS GROUP STIMULUS MATERIAL

Scenario A

Becca is waiting to enter the school hall to take her English GCSE exam. It's her last exam of the week and is worth about 60 *per cent* of her final grade. The other 40 *per cent* was coursework, where Becca wrote an essay analysing the themes in the book her class had studied; *Lord of the Flies*. Everybody from Becca's year is here, including all her friends, standing outside waiting to be let in to the school gym, which has been cleared of its usual clutter of basketballs and sports bibs. In their place is row upon row of identical chairs and tables, a clock at the front, and the exam invigilator's table, where all the exam papers are sat. It's nearly time for the exam and Becca checks once more that she has all her pens and stationery. The door to the hall is opened and Becca and the rest of her year file in and take their seats while the invigilator writes the start and finish times on the whiteboard at the front.

Scenario B



It is mid-November and it's just starting to get cold and frosty outside. Paul and his friends wait outside the sports hall to go into their second GCSE Maths modular exam. Paul feels a bit nervous but reminds himself that he has done well on his first modular exam and thinks back to all the practice papers he did. As his friends have pointed out, this exam is shorter than end of year exams so there is less to remember. All the same, Paul doesn't feel as confident about this exam as he did for the first one and wonders whether he'll have to re-sit this one when he takes the third and fourth module exams. It would mean more revision in March, when the next set of modular exams take place. He thinks it would be more stressful to revise for two at once, especially as he doesn't get a break or a holiday before going back to class to start thinking about the next set of tests, but he supposes it might be worth it for a better mark.

Scenario C

Jodie and Callum are waiting to take their GCSE Science exam. It doesn't feel like exams they've taken before; normally the whole year group is waiting with them but today it's just Jodie, Callum and a couple of kids from Mr Gregory's Science class. Their teachers entered them for the exam because they felt that they were ready to take it. Some people took the exam a few weeks ago, when they were ready. Some people still haven't taken it. Jodie likes the idea that she can take the test when she feels she is able to – it's nice to have it out of the way. Callum though, feels worried. He likes to plan and spread his revision out over the year, and he feels anxious that he hasn't had enough time to prepare. He worries vaguely that he might be in the wrong place; they aren't in the sports hall as usual because there are so few of them. He tells Jodie his fears, who reminds him that he can always re-sit later on, and that they have covered this topic fairly recently in class and he did well then.

APPENDIX 3

SCHOOL STUDENTS' FOCUS GROUP STIMULUS MATERIAL

How do you feel about exams?

**We run GCSE and A levels,
and we'd like to know what you
think of them.**

Circle the answer closest to how you feel in each case.

1. How nervous do you get about exams?
   a. I show cucumbers the meaning of cool
   b. A little bit – generally I'm fairly calm
   c. Pencil-chewing, hair-tearing-out kind of nervous
   d. Quite nervous
   e. It depends on the exam

2. What do you do when you leave the exam hall?
   a. Meet up with all my friends and discuss how it went
   b. Go home and watch TV
   c. Go out and have fun!
   d. Start revising for the next one…

3. How would you like to revise for your exams?
   a. Past papers – and lots of them!
   b. Write out notes from the textbook
   c. Reading and rereading the textbook
   d. Having revision sessions with your friends
   e. Revision? I'm not sure I understand the question…

4. What kind of exam would you prefer?
   a. Short answer
   b. Long answer
   c. Multiple-choice
   d. Practical/oral

Please turn over…

5. How many re-sits would you like?
    a. Unlimited – the more the merrier
    b. As many as you can before the end of school
    c. 3 strikes and you're out
    d. Just the one
    e. None at all

6. What type of assessment would you prefer?
    a. More coursework, less exam
    b. More exam, less coursework
    c. Half coursework, half exam
    d. All exam
    e. All coursework

7. When would you prefer to take your exams?
    a. All at the end of the school year
    b. Spread throughout the school year
    c. Every Friday
    d. When I feel I've revised enough

8. How would you prefer to be taught?
    a. In class, with the same exam for everyone in the school hall
    b. In class, doing mini-exams at the end of each topic
    c. In small groups, taking exams when the teacher thinks you're ready
    d. By a tutor, taking exams when you choose to

Thank you for taking the time to fill this in.

APPENDIX 4

TEACHERS' FOCUS GROUP SESSION PLAN

Stage 2b – Focus Group

Introduction

- Introduce researchers
- We are running a project reviewing on-demand testing
- We would like your views
- Everyone introduce themselves
- Ground rules

Opening Questions

- How many times a year can your students enter their GCSE exams (e.g. just in June, or in November and March as well?)
- Are some students more ready than others at these set test windows?
- Do you find that the current 3 test windows a year are enough?

Describe On-Demand testing

Transition Questions

- What do you think about exams being more frequent and covering smaller chunks of the syllabus?

- If test windows were more frequent do you think that students would benefit?
- Do you think more able students will want to broaden their examination range?

- How would you decide when a student was ready?
- Do you think parents might force children to take examinations?

- What are the implications of students being ready at different times?
- It will become compulsory for children to stay in education until they are 18. If students were to take their exams earlier, what would you do with them after their exams?

- Do you think more able students will seek extra tuition from commercial tutorial companies

Key Questions

- Questions will need to be repeated over time– how do you feel about that?
  - o Do you think this will be a security risk?

(May need to explain live test –pre-test)

- How would you feel about questions being included in the exam which are not going to be marked?
- How would you feel about past papers not being released, only one specimen paper or a set of specimen questions?
- How would you feel if your class took a different paper from other centres but was graded independently?
- What support would you want from AQA?

Serendipitous Questions

Ending Questions

- Give a summary - Is this an accurate summary?
- Have we missed anything?

APPENDIX 5

QUESTIONNAIRE

SECTION 1: ON-DEMAND TESTS

1.1 Please list the assessments which are supported by an item bank and which your organisation may describe as "on-demand". If it is more convenient to provide a list on an additional piece of paper, please do so.

Queries under following headings:

Assessment /Duration of a test window / Number of test windows offered per year / Period of notice from candidates or centres / Primary delivery format (on-screen/paper) / Intended use of assessment (formative/summative/diagnostic/selection) / Intended users

SECTION 2: ITEM BANK

If you listed more than one item bank in your response to question 1.1, you may want to select one or more on which to base your responses to the following questions. If this is the case, please would you indicate which item banks your responses are based on.

2.1 What types of items are stored in the item bank system?

Multiple-choice questions / True/false questions / Multiple-response questions / Matching questions / Short free text entry / Essay prompts / Hotspots / Hotlines / Sliders / Simulations / Other

2.2 What functions does the item bank system carry out? Please tick those that apply.

Functions listed: Storage of metadata / Storage of statistical information / Calibration of items / Equating of tests / Automatic construction of tests / Semi-automatic construction of tests / Other

Please state other functions.

Respondents requested to indicate location of function with three options:

In same storage facility as item / Stored in separate storage facility from item / Not a function of the item bank system

2.3 Who is responsible for providing the components of your item-banking system?

Components listed: Item storage / Metadata storage / Storage of statistical information / Storage of usage data / Calibration of items / Equating of tests / Construction of tests / Storage of tests before delivery / Storage of user access data

Respondents requested to indicate location of responsibility with five options:

In-house, using own software / In-house using proprietary software / Outsourced to a third party / Brokerage system / Other

2.4 If you ticked "Other" for one or more of the components in question 2.3, please would you describe who has responsibility in the space below.

2.5 How many items are stored in each item bank?

2.6 How old is each item bank?

2.7 Is (are) the item bank(s) based on a relational database? Yes / No

2.8 Please describe briefly the design of your item bank system.

SECTION 3: MANAGING AN ITEM BANK SYSTEM

3.1 Who uses or accesses the item bank system?

Users listed: item writers / Test constructors / Item bank managers / Administrators of qualifications / Managers of delivery system / Other, please state

Respondents requested to indicate frequency of use with four options:

Daily / Weekly / Monthly / Other, please state

3.2 How often are items exposed?

Information requested divided into:

Number of times an item is exposed / Over what period of time / No item usage strategy in place

3.3 Does the item bank system have an item retirement strategy? If "yes", please outline the strategy and what information about an item the strategy requires.

3.4 In what format are the items stored within the item bank?

3.5 How many new items are validated per year?

3.6 Do the item banks adhere to the IMS Global Consortium's Question and Test (QTI) specification? Yes / No

3.7 Are any additional materials stored with an item other than the question? If "yes", please state what these additional materials are. Yes / No

SECTION 4: SECURITY ISSUES

4.1 Are any of the following encrypted during either storage or transfer or both?

Objects for respondents' consideration:

Items / Metadata / Statistical information / Candidates' responses / Candidates' personal information

4.2 If encryption is used, what protocol is followed?

4.3 What safeguards are in place to ensure that the items in tests that are delivered electronically are secure during and after the test is taken?

4.4 How is the integrity of items that will be re-used ensured?

SECTION 5: PRE-TESTING

5.1 Are items pre-tested? Yes / No

If you answered "yes", please go to question 5.2, if you answered "no", please go to question 6.1.

5.2 Are **all** new items pre-tested? Yes / No

5.3 What is the sample size for pre-testing?

5.4 What is the frequency of pre-testing?

5.5 Is pre-testing done by inserting non-scoring items into live tests? Yes / No

5.6 Please describe the methodology used for pre-testing of items.

5.7 Please describe what measures are taken during pre-testing to ensure the security of items.

SECTION 6: ITEM CALIBRATION

6.1 Which pieces of statistical information about item performance are retained, would it be desirable to retain and, if not currently retained, may be retained in the future? Please tick.

Pieces of statistical information: Mean mark / Standard deviation / Sample size / Difficulty / Facility / Discrimination / Distractor statistics / Bias statistics / Item usage / Maximal error information / Standard error / Quintile performance

Respondents requested to indicate use of statistic with three options:

Statistic retained / Statistic desirable to retain / Statistic will be retained in the future

6.2 Are any other pieces of statistical information about items retained?

6.3 Which pieces of statistical information about test performance are retained, would it be desirable to retain and, if not currently retained, may be retained in the future? Please tick.

Pieces of statistical information: Mean mark / Standard deviation / Sample size

Respondents requested to indicate use of statistic with three options:

Statistic retained / Statistic desirable to retain / Statistic will be retained in the future

6.4 If any other pieces of statistical information about tests are retained, please describe them below.

6.5 What is statistical information about items used for?

Uses listed: Next exposure of item / Item retirement / Item alteration / Future item writing / Checking for bias / Checking for candidates with prior knowledge of items / Other

Please state other uses.

Respondents requested to indicate use of statistical information with three options:

Currently used for this purpose / Plan to use for this purpose in the future / Statistic and critical range of values

6.6 How frequently are items calibrated or re-calibrated?

6.7 What method is used to calibrate items? Please tick all those that apply.

Classical Test Theory / Rasch modelling / Item Response Theory: 2 parameter / Item Response Theory: 3 parameter / Other / Items not calibrated

If you ticked "Other", please state what method of calibration is used.

SECTION 7: METADATA AND TEST CONSTRUCTION

7.1 Does the item bank system have the capability of producing metadata in a format that conform to the IEEE Learning Object Metadata Standard?

7.2 How are items initially organised in the item bank?

As individual items / In item pools / Aggregated into tests / Under themes / Other

If you ticked "Other", please describe the first level of storage in the item bank.

7.3 Please indicate which of the following you consider to be metadata, statistical information, or usage data.

Objects listed: Globally Unique Identifier (GUID) / Time taken to attempt an item / Instructions to candidate / Specific skill / Discrimination / Name of author of item / Last exposure of item / Content or topic area / Difficulty/facility / National learning objectives / Next exposure opportunity for item / Name of validator/reviewer of item / Item type

7.4 What are the purposes of metadata in the item bank?

7.5 How are items selected for inclusion in an on-demand test?

7.6 What criteria are used to select an item for inclusion in an on-demand test?

7.7 If there is a test specification, what criteria does it focus on to construct a suitable on-demand test?

7.8 To what extent is the construction of on-demand tests an automated process?

7.9 Who is responsible for test construction?

7.10 Where and by whom is test construction carried out?

7.11 When in the assessment cycle is test construction carried out?

7.12 How is the equivalence of on-demand tests ensured?

7.13 Please describe any reviewing or validation process of on-demand tests that is carried out prior to the administration of the tests.

7.14 If parallel or alternate forms of on-demand tests are offered, why is this done?

7.15 Please describe how the comparability of electronically delivered tests is measured, monitored and maintained?

SECTION 8: STANDARDS

8.1 Based on your knowledge of the assessments listed in Section 1, please describe the procedures used to ensure standards are maintained from one administration of an on-demand test to another administration.

8.2 If a particular test design is used to enable test equating (horizontal equating) over time, please describe it below.

8.3 Is an internal or external set of anchor items used?

8.4 If an anchor is used, what percentage of the items in the entire test make up the set of anchor items?

8.5 If the partial credit model is used, please describe the circumstances of its application.

8.6 What precautions are taken to account for drift of the values of difficulty and discrimination over time?

SECTION 9: RESPONSES AND RESULTS

9.1 Where are the responses to questions in an on-demand test stored?

9.2 For how long are the responses stored?

9.3 In what format are the responses stored?

9.4 Are on-demand tests scored immediately after the tests are sat by the candidates using software installed on centres' hardware?

9.5 How soon after sitting an on-demand test do candidates receive notification of their results?

9.6 In what form do candidates receive their results (e.g. a mark, a scaled score, a grade, or a level)?

9.7 What mode of delivery is used to send results to candidates?

9.8 How soon after sitting an on-demand test do candidates receive their certificates?

9.9 Is there any guidance offered to candidates on re-taking on-demand tests? If "yes", please describe the guidance on re-taking.

9.10 Who has responsibility for analysing the responses from candidates?

SECTION 10: THE FUTURE

10.1 What do you think are the areas in which development is needed in the provision of on-demand testing?

10.2 What do you think are the areas in which development is needed in item-banking?

First published by The Office of the Qualifications and Examinations Regulator in 2009.

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

The Assessment and Qualifications Alliance (AQA) is a company limited by guarantee registered in England and Wales (company number 3644723) and a registered charity (registered charity number 1073334). Registered address: AQA, Devas Street, Manchester M15 6EX. *Dr Michael Cresswell,* Director General.