Centre for Education Research and Policy

ITEM BANKING WITH A TEST CONSTRUCTION INTERFACE: AN EVALUATION OF A PROTOTYPE

Claire Whitehouse, Qingping He & Christopher Wheadon

SUMMARY

A prototype item bank/test construction interface was developed and evaluated within the Research and Policy Analysis (RPA) department to assess how alterations to the current question paper setting process brought about by item banking may impact on the quality of tests. The use of a prototype avoided the costs and downtime usually associated with a staged implementation of major changes to operational processes. The evaluation of the prototype, particularly its capability for the semi-automated construction of tests from a calibrated item bank, took place in two stages. In Stage 1 three Principal Examiners (PExs) used the prototype to construct tests for their components: unit 1 of GCE Economics (ECN1/1), module 3 Foundation tier of GCSE Mathematics (43003F) and Entry Level Certificate (ELC) Adult Numeracy Entry Level 3 (EL3). The newly constructed tests (NCTs) were later part of a blind comparison with past question papers in Stage 2. The pairs of tests were evaluated and rated by subject administrators and senior examiners from cognate subjects.

The evaluation demonstrated that it is feasible for PExs to construct tests using the statistical information and information about the qualitative characteristics (metadata) of items that are stored with the items themselves in an item bank with a test construction interface. Whilst subject administrators and senior examiners from cognate subjects were able to differentiate the NCTs from the past question papers, the negative impact of the semi-automated test construction method was judged to be slight. This negative impact was apparent in the look and feel of the NCTs as represented by gradients of difficulty, the use of contexts in items and the number of items containing graphs or tables.

Three issues which need to be addressed for the successful implementation of test construction from item banks were identified. Firstly, the training of staff in the concepts underpinning item banking and test construction needs to be developed and implemented. Secondly, the metadata associated with items needs careful definition and expansion from the current pool of characteristics. Finally, there is enormous scope for simplifying assessment objectives and descriptions of content areas which would greatly assist the semi-automated construction of tests. Currently these features are multifaceted and difficult to separate out into individual skills for assessment.

BACKGROUND

An operational item bank stores items that have been tagged with pieces of information about their qualitative characteristics, e.g. assessment objectives and content areas (known as metadata) and contains information about the statistical performance of the items gained through some form of pre-testing. Information about item-test performance guides the construction of future tests. The resulting tests possess known characteristics and known grade boundary marks. These are the tests required of an assessment environment in which on-screen, on-demand is the order of the day.

The Assessment and Qualifications Alliance (AQA) is a company limited by guarantee registered in England and Wales (company number 3644723). Registered address: AQA, Devas Street, Manchester M15 6EX

An item bank containing appropriately tagged items offers flexibility of test construction. Provided sufficient items have been banked a test constructor can produce multiple versions of a test for test windows of varying duration and frequency. The Research and Policy Analysis (RPA) department has developed a prototype item bank/test construction interface for evaluation purposes. The evaluation described in this report sought to identify the operational parameters or business rules that may ensure flexibility of test construction.

A prototype serves as a model; the parameters of which can be adjusted without incurring the scale of costs and potential downtime that such adjustments would cause in an operational environment. Experimental work carried out on a prototype can identify key parameters and provide insight into ideal operating conditions. Technology specifications that are based in a working knowledge of the key parameters are more likely to incorporate achievable requirements and requests that are relevant to the purchasing organisation's operations. Moreover, the use of a prototype enables RPA to evaluate how any alterations to the question paper setting process could affect the quality of tests presented to centres and students.

This report presents the complete findings of the two-stage evaluation which was designed to address the following research questions:

1. Is it feasible for a Principal Examiner (PEx) to construct a test using a calibrated item bank?
2. If yes, what factors are important in contributing to the success of semi-automated test construction?

Stage 1 of the evaluation was concerned with the construction of tests by PExs using the prototype item bank/test construction interface and a comparison of their experience using the prototype with their usual test construction experience. Colleagues in subject administration and senior examiners in cognate subjects participated in Stage 2 of the evaluation which was based on a blind comparison of the tests constructed using the prototype and past question papers.

DESIGN

The prototype item bank/test construction interface developed within AQA was capable of storing items and their mark schemes. Each item was tagged with its metadata in the form of assessment objectives and content areas. Item-test performance information was also stored with each item and was calculated from both Classical Test Theory (facility and discrimination indices) and Item Response Theory (difficulty and discrimination indices). Where items were used in operational tests, a set of facilities by grade were also available. The test design interface had the capability to search for items using one or more of four pieces of information (assessment objective, content area, facility and difficulty).

Stage 1: Semi-automated test construction

The operation of this prototype in the construction of tests was evaluated by three test constructors. This was a small sample size and whilst this alone did not make the results less valid they should be treated with some caution. They are not representative of all test constructors or all tests. Each test constructor was a PEx and therefore experienced in the drafting of question papers for a Question Paper Evaluation Committee (QPEC). Each evaluation session took place in the AQA office and consisted of the construction of a test using the prototype, the self-administration of an evaluation questionnaire (see Appendix A) and a structured interview conducted by a member of the research team. The interview was based on observations made by the facilitator (who also conducted the interview) during test construction and on the participant’s responses to the evaluation questionnaire. Questions asked during the interview covered topics such as the use and purpose of multi-part items and the order of presentation of items. The three components and subjects used in this evaluation were unit 1 of GCE Economics (ECN1/1), module 3 Foundation tier of GCSE Mathematics (43003F) and Entry Level Certificate (ELC) Adult Numeracy Entry Level 3 (EL3). ECN1/1 is comprised entirely of one mark, multiple-choice items, whilst the other two components utilise short free text response items; see Table 1. These item types are most appropriate for on-screen tests and thus on-demand testing. Only Adult Numeracy required the keying of candidates’ responses from scripts as there are large volumes of such data stored electronically for the other two components. This evaluation used items and data from nine examination series for ECN1/1, five examination series for 43003F and, from four of the current versions of Adult Numeracy. These latter two components use multi-part items with maximum marks ranging from one to nine. As there is usually a common stem or context running through these items they cannot be separated into independent items with lower maximum marks. Consequently, at just 44 large, multi-part items, Part A of module 3 Foundation tier of GCSE Mathematics (43003F) has the smallest item bank.

Table 1: Summary of components used to evaluate the prototype item bank/test construction interface

	GCE	**GCSE**	**ELC**
			Adult
	Economics	**Mathematics B**	**Numeracy**
	ECN1/1	**43003F Part A**	**Entry Level 3**
Multiple-choice (marks)	15	--	--
Short free text response (marks)	--	32	40
Maximum mark	15	32	40
Duration (minutes)	15	40	60
Source of items in prototype bank			
Initial examination series	January 2004	November 2006	---
Total number of tests	9	5	4
Number of items in prototype bank			
Previously used	135	44	102
Pre-tested	30	--	--
Maximum mark highest	1	9	6
Maximum mark lowest	1	2	1
Item-test performance information			
Sample size, pre-tested items	180 - 350	--	--
Sample size, previously used items	3,600 - 8,100	5,000 – 72,000	200 - 300
Source of item data	EPS	CMI+	keying of marks

ECN1/1 is the only component of the three which is compiled from a physical item bank and re-uses items that were released in past papers. It is not usual to re-use items in the other two subjects therefore this simulation of an environment in which semi-automated test construction took place with pre-tested items lacked some ecological validity. For 43003F and Adult Numeracy the calibration of items was non-ideal in that there were no linking items between tests and it was necessary to assume random equivalence of candidate ability between examination series or sittings. This was also the case for ECN1/1 due to there being too few linking items between tests; although the accepted exposure rate of items for this component is two years.

Stage 2: Evaluation of the newly constructed tests (NCTs)

The tests that were constructed by the PExs in Stage 1 were evaluated in a blind comparison with past question papers or, in the case of Adult Numeracy, a current operational version of the question paper. Text and formatting that might identify the tests were removed and they were referred to as “Test 1” and “Test 2”. Three colleagues in subject administration, one for each of the components listed in Table 1, and four senior examiners in cognate subjects participated in the blind comparisons. One of the senior examiners was unable to attend the evaluation session held in the office and therefore completed the blind comparison remotely. Another participant from a cognate subject was able to complete the Stage 2 evaluation for two components. Subject administrators completed the blind comparison at their desks.

The documents provided to each participant in Stage 2 of the evaluation are listed in Table 2. Responses from subject administrators to section one of the questionnaire (see Appendix B) were used to decide what additional documents should be provided to the participants from cognate subjects. This latter group of participants was allocated 30 minutes in which to familiarise themselves with the assessment objectives and content areas of the component for which they were requested to evaluate tests (see Appendix C for questionnaire and results).

Table 2: Summary of documents provided to participants in Stage 2 of the evaluation.[1]

Documents provided to participants in Stage 2 evaluation	**GCE Economics ECN1/1**	**GCSE Mathematics B 43003F Part A**	**ELC Adult Numeracy EL3**	**Provided to**
Newly constructed test (NCT)	✓	✓	✓	Subject administrators & cognate subject participants
Past question paper	✓	✓		
Operational version			✓	
Questionnaire	✓	✓	✓	
Guidance on completing questionnaire	✓	✓	✓	
Specification	✓	✓	✓	Cognate subject participants only
June 2008 question paper and mark scheme	✓			
Specimen question paper and mark scheme from AQA website		✓	✓	
Test construction grid	✓			
Adult Numeracy Core Curriculum (A1042). The Basic Skills Agency (2001)			✓	
GCSE mathematics criteria QCA/06/2901(October 2006)		✓		

In the blind comparisons participants were requested to evaluate the two question papers separately and then to respond to a questionnaire and indicate which question paper, in their opinion, was the NCT and which the past question paper. Section 1 of the questionnaire dealt with those characteristics of question papers that are recommended for consideration in AQA’s *Question Paper Preparation Procedure Guidance File* (2007) and in *Preparing question papers and mark schemes. Guidance for Chief/Principal Examiners, Revisers and Scrutineers*. Section 2 asked participants to consider the statistical information on each question paper, which included percentage weightings of assessment objectives, percentage weightings of

[1] Printed copies of the test, mark scheme and statistical information were provided for each test that was part of the blind comparison.

content areas, values of facility indices and difficulties for each item, and the estimated cut scores for each grade as a percentage of the maximum mark. Participants were provided with written guidance on what to look for in the statistical information and how to use it. Sections 3 and 4 asked participants to rate on a five point scale a series of statements concerning additional characteristics they may consider during the evaluation of question papers and the potential benefits of an item bank/test construction interface, respectively. With a total of seven participants, the same caveat applies to the results from Stage 2 of the evaluation as to those from Stage 1: the sample size was small and therefore it is not possible to generalise from them.

RESULTS AND DISCUSSION

Stage 1: The test construction experience

> "*I'm not, I'll just say this, I'm not a person who lives on IT but I soon found it very easy to use and I know if I came back to it tomorrow I could use it very easily straight away and I have a lot of problems with some IT….I think it passed the test in many ways.*"
>
> GCSE Mathematics test constructor

The responses of the test constructors during the evaluations of the prototype were, on the whole, strongly positive towards the new test construction experience, their constructed tests and the prototype application (see Appendix A). The test construction experience was rated highly for being informative, time efficient, enjoyable, focused and straightforward. It also received a low rating (1.67 out of a possible 5) for being confusing. The test constructors for GCSE Mathematics and Adult Numeracy reported finding themselves using statistical information more than they do in their usual test construction experiences. The ECN1/1 test constructor used this information to the same extent as usual, but found it was more readily available. The usual test construction experience for GCSE Mathematics and Adult Numeracy does not require test constructors to consider any statistical information associated with the items being included in tests, nor is any available as the items are not pre-tested. Examiners in GCSE Mathematics are provided with the statistical information for the last question paper taken, but as these items are not re-used the information is of no value in the item banking environment.

When asked about the tests they had constructed during the evaluation, the test constructors rated them highly for most of the criteria considered to be important in AQA's question paper setting process. The highest ratings were for appropriate assessment objective weightings, appropriate content area coverage and the capability of the constructed test to discriminate across the ability range. One of the outcomes of Stage 1 of the evaluation is that it is feasible for a PEx to construct a test using a calibrated item bank. A second outcome is the identification of a number of themes which may illuminate the new test construction experience, using the prototype, and the reasons why it received high ratings from the participants. These themes are discussed below[2].

Theme 1: Test constructor engagement

> "*The other word I'd choose is 'reassuring' because you've got the statistics there you know you haven't left a big chunk of something*

[2] The newly constructed tests are available to view on request to the authors. The tests for ECN1/1 and Adult Numeracy are considered to be live papers in terms of security.

> *out, or the paper's too hard or it's not a comparable version. You can check that very quickly with the facilities on there.*"
>
> Adult Numeracy test constructor

The complex statistical data incorporated into the item-test performance information were represented in easily accessible graphical and tabular forms in the prototype. Initially test constructors opted to use the graphical forms to assist them in ordering items within the test in terms of their relative facilities (or difficulties). The graphical forms also enabled constructors to pursue the first of two types of a 'filling the gaps' strategy, with the aim of constructing tests that showed a progression of difficulty. From about half-way through test construction the tabular forms of representing the statistical information were used with increasing frequency to check on the weightings of assessment objectives and for a balanced coverage of content areas. Two of the test constructors were able to avoid potential grade compression by using the feature which shows facilities by grade.

Given the resources to engage with the statistical information, the participants were able to move away from a holistic approach to question paper setting and adopt a more fragmented approach, although the ECN1/1 test constructor already works in this manner. The holistic approach is applied particularly to tests consisting of free text response questions and the question paper is a complete entity almost from the start of the question paper setting process. The item banking approach to test construction is more fragmented in comparison and forces the test constructor to build the test from blocks, the items, by using metadata and statistical information about items as aids in a decision making process.

This evaluation demonstrated that the participants were able to construct tests, with which they were generally satisfied, by using a fragmented approach. This approach enabled the test constructor for Adult Numeracy to build a test which he considered to be better, in terms of unpredictability, than those tests that are in use currently. This may not be surprising as the same six tests have been in use for a number of years. However, the ability to view and select banked items allowed the test constructor to compile items into a non-standard format whilst still retaining a close match to the weightings of assessment objectives in the current tests.

Theme 2: Visuals matter

The visual representations of the statistical information, using graphs and tables, captured the test constructors' attention. The participants sought to extend this visual interaction during the building of the test and the reviewing and revising stages. For example, the capability of viewing a group of items selected by a combinatorial search whilst building the test was requested. The facility to preview a test as it would look to candidates, on screen and in printed form, was also considered crucial as it allowed test constructors to get the 'flavour' of the tests they had constructed: "*Downside – not able to see paper building up visually*" (ECN1/1 test constructor). Previewing in this case meant being able to see which items were on facing and following pages, in the case of paper-based tests, and being able to replicate how candidates were likely to interact with on-screen tests.

Theme 3: Combinatorial searching

At the start of test construction the participants used assessment objective, facility and content area: a combinatorial search. This is equivalent to the initial searching and extraction a test constructor carries out on a physical item bank, but much faster. Provided the test constructor had a starting point in mind for the test, combinatorial searching acted as a fine meshed filter.

For these participants the starting point was the qualitative characteristics of the first few items in the tests.

> "*…you just can't find what you need erm to fill in those little gaps at the end where you've only got one or two marks at a grade left or a particular topic left that you want to ask on then it does, I mean the bank isn't as big as you'd want…*"
>
> GCSE Mathematics test constructor

Unfortunately, the usefulness of combinatorial searching started to decline somewhere between two-thirds to three-quarters of the way through test construction. One test constructor felt he was "relying too much" on searching by facility and instead opted to search by content area only and then to scan the contents of the item bank item by item. By this stage of the test construction the constructors were using a second type of a 'filling in the gaps' strategy. The key pieces of information for selecting an item were now ciphers for how well the potential item would fit with the items already in the test.

Theme 4: Speed of test construction

The participants' opinion of the speediness of test construction using the prototype may be summed up by the test constructor for GCSE Mathematics: "*Much quicker way of producing a paper which is balanced and appropriate*". All of the test constructors agreed, in response to the questionnaire, that the use of the prototype led them to construct their tests faster than they would usually (see Table A2). They also gave a high mean rating (4.7 out of a possible 5) to the phrase 'Time efficient' when used to describe their test construction experience with the prototype (see Table A3).

It took the ECN1/1 test constructor approximately 75 minutes to construct a test. This compares to three to four hours to draft a test using the item bank consisting of index cards; a potential time saving of approximately 60% for the construction of a multiple-choice test.

Theme 5: The limitations of multi-part items

Once items are accepted into an item bank, they should remain unaltered during storage and subsequent use. However, multi-part items present test constructors with a problem: not all parts of a multi-part item may be suitable for the test under construction. This became increasingly apparent towards the end of test construction when the constructors were searching for items to extend the coverage of content areas. Test constructors preferred to take one of two routes, both of which are barred to them under the constraints imposed by item banking. The first route is to extract those parts of the multi-part item which are considered to be useful to the test under construction. The second route is to edit the item. An alternative route which is appropriate for item banking is to write items asking only one question.

A test constructor's urge to edit items became most apparent with items that can be cloned, especially when the constructor's knowledge of operational tests told him/her that by changing a number or word the question would be converted from stale to fresh, presenting a new challenge to the candidates. This mindset may be overcome by preparing sufficiently large item banks that contain a range of pre-tested as well as exposed items. As the test constructor for Adult Numeracy observed: "*I guess it's human nature to want to change things, to make them your own*".

Stage 2: The newly constructed tests (NCTs)

Comparison of statistical information

The statistical information for the newly constructed tests (NCTs) and the past question papers used in the blind comparison (Table 3 and the tables in Appendix D) demonstrated that the NCTs were able to at least match, if not show some improvement over, the past question papers in all aspects except the weightings of content areas and also the weightings of assessment objectives for GCSE Mathematics. However, a more considered analysis of these weightings for Adult Numeracy and GCSE Mathematics showed the NCTs contained only one less content area than did the past question papers (see Tables D3 and D5), which may explain why the test constructors were satisfied with their NCTs. Weightings of content areas (and assessment objectives) are easily calculated and monitored in an electronic item bank during and after test construction. So long as these features of an assessment are clearly defined and the test specification is precise, there is no scope for subjective interpretations of content.

Table 3: A summary of statistical information derived from the quantitative information in Appendix D for the NCTs and the past question papers in the blind comparison.

Component	NCT and past question paper match of:	
	assessment objective weightings	content area weightings
GCE Economics ECN1/1	✓	✓
ELC Adult Numeracy EL3	✓	✗
GCSE Mathematics 43003F Part A	✗	✗

Blind comparison of tests

The NCTs constructed by the PExs in Stage 1 of this evaluation were identified easily, with all subject administrators and 80% of cognate subject examiners correctly selecting the NCTs (see Tables B3 and C3). Unfortunately, the blind comparison was probably confounded by the memory of past question papers, whether conscious or not, retained by the participants in the subject administration group. This may also have been a factor in the group of cognate subject examiners, although the questionnaire did not contain any items to test their memory or working knowledge of question papers in their cognate subject.

Though the participants in the blind comparison were able to identify the NCTs, there is very little to separate the NCTs from the past question papers when participants' ratings of the 15 characteristics listed in Table 4 are considered. The ordering of the characteristics (highest to lowest rating) is similar for both tests (columns 5 and 6 in Table 4). The four highest rated and the four lowest rated characteristics are identical for both NCTs and past papers. Interestingly, in an operational item banking system, four of the top five rated characteristics are decided before an item is accepted into the bank for use in live tests (indicated by ✍ in column 2 of Table 4).

Did participants' ratings show they were of the opinion that one test was better than the other? Weighted mean ratings across all 15 characteristics of 3.8 for NCTs and 4.0 for past papers (out of a maximum rating of 5.0) suggest perhaps they considered the past papers to be better, but only by a small margin. Relative ratings (column 7 in Table 4, by which the table is ordered) of

Table 4: Characteristics of the NCTS and the past question papers in the blind comparison sorted by relative rating in descending order and whether there was agreement between subject administrators and senior examiners in cognate subjects (ratings of all participants used)

1	2 Timing of decision-making	3 Rating for NCT	4 Rating for past paper	5 Order by NCT rating	6 Order by past paper rating	7 Relative rating (column 3 – column 4)	8 Agreement between two groups	9 Total N
Free from bias	✍	3.75	3.57	5	= 8	NCT rated higher	✓	8
Able to discriminate amongst candidates of varying ability	💻	3.57	3.43	= 7	= 11		✓	7
Not predictable in terms of questions and topics covered	💻	2.88	2.75	15	15		×	8
Can be answered satisfactorily in the time allowed	✍	4.63	4.63	1	2	Same	✓	8
Provides a suitably demanding challenge for higher-achieving candidates	💻	3.57	3.57	= 7	= 8		✓	7
Would allow an average candidate to gain around 60% of the marks	💻	3.50	3.57	9	= 8	Past question paper rated higher	×	*7
Similar to past question papers or current versions	✍	3.88	4.00	= 3	4		×	8
Contains appropriate content area (topic) weightings	💻	3.00	3.25	14	14		✓	8
Presents questions in the best order	💻	3.67	4.00	6	5		✓	6
Contains a balanced coverage of content areas	💻	3.00	3.38	12	13		✓	8
Contains content area (topic) coverage that is comparable with past papers or current versions	✍	4.43	4.86	2	1		✓	7
Assesses the full range of skills and abilities as defined by the assessment objectives	💻	3.00	3.43	13	= 11		✓	7
Contains appropriate assessment objective weightings	💻	3.25	3.75	11	7		✓	8
Contains an appropriate gradient of difficulty	💻	3.29	3.86	10	6		✓	7
Comparable in demand with past question papers or current versions	💻	3.88	4.50	= 3	3		✓	8

* N = 6 for NCTs ✍ = decision-making process takes place before test construction 💻 = decision-making process takes place during test construction

the characteristics indicate there were 5 out of the 15 on which the participants considered the NCTs performed better than or the same as the past papers. Two of the characteristics on which participants felt the NCTs performed better than the past papers deal with discrimination, as may be expected given the statistical information available in their construction. Once the properties of tests are measurable guidelines can be developed for best practice in test construction. The following scenarios show how the result of the measurement of tests can be incorporated into test construction.

Simplifying and identifying to measure accurately

Participants perceived sufficient difference between the NCTs and past question papers to rate the past question papers higher for their capability to assess the full range of skills and abilities. The assessment objectives for the three components considered in this evaluation, particularly for Adult Numeracy and GCSE Mathematics, are dense, multifaceted and overlapping. Within these formal descriptions lie requirements for assessment that are not fully articulated. The two examples provided below serve to illustrate this problem.

Scenario 1

GCE Economics ECN1/1: AO2: apply knowledge and critical understanding to economic problems & issues

This requires candidates to use and interpret graphs or tabulated data or the information contained in the context of a question. Therefore, the proportions of items containing graphs, tables or information in a block of text become a measure of the quality of a question paper.

> "*In economics whether the question requires any calculations or not because some students find those difficult, some find them easy. So it might have the same facility as an average word based answer but there er should be numerical questions on an economics unit 1 paper.*"
>
> GCE Economics cognate subject examiner

Scenario 2

GCSE Mathematics 43003F Part A: AO2-NA-4: Solving Numerical Problems

This often requires items to be set in a context, leading to one of the criteria in the evaluation of question papers being the variety of contexts used and whether there is overlap with items in past question papers or within the question paper under consideration.

> "*Well we have some questions in context and some in clearly not context on exactly the same topic and it changes the facility. Sometimes it cuts both ways, sometimes the context helps, sometimes the context just throws them altogether. We've never done paint brushes. We put comments on the papers like 'We've never done paint brushes in maths'. Ask a question on paint brushes*."
>
> GCSE Mathematics cognate subject examiner

Assessment objectives and content areas for assessments identified as suitable for on-demand need to be drafted precisely, to contain one and only one objective or area and to be relevant to the assessment. They should also be considered along with additional criteria which would become metadata in an item bank. When asked about additional characteristics of question papers that they consider (Tables B6 and C6 for results from Section 3 of the questionnaire), both the senior examiners in cognate subjects and the subject administrators rated the following most highly.

1. The contexts of questions (see above)
2. The accuracy of the mark scheme
3. The stimulus materials used.

> "*An overall sense of balance across the paper considering length of questions, demand, context/no context.*"
>
> GCSE Mathematics subject administrator

The characteristics that received the highest ratings for each group represented their greatest concerns. The examiners in cognate subjects, who are themselves involved in setting and evaluating question papers, rated contexts of questions as being most important. Searching for new and relevant contexts can take up a great deal of a question paper setter's time and the setting of an item within a context is a non-trivial task. This may offer an additional explanation for the reluctance of question paper setters to move from the holistic test construction approach involving the use of multi-part items to the more fragmented approach of writing independent items which may require their own contexts. Such reluctance may be overcome by offering greater support in the searching for contexts and acknowledging that not all items need to be standalone.

The subject administrators were most concerned with the production of an accurate mark scheme. Again, this relates directly to their role in the question paper setting process. It is notoriously difficult to start to write a mark scheme part way through the process and the preferred method is to write the mark scheme in parallel with the question paper. An item bank that contains items with their mark schemes and any commentary is capable of producing automatically an entire mark scheme for items compiled into tests. Therefore, the effort shifts from the current QPEC process to ensuring item writing is completed fully before an item is banked.

Perceived benefits of item banking and semi-automated test construction

Both groups of participants in Stage 2 of the evaluation agreed on the following potential benefits of item banking with a test construction interface by giving them the top three ratings (see Tables B8 and C8 for the results from Section 4 of the questionnaire).

1. Construction of tests with known and recorded difficulty/facility and discrimination
2. Automatic production of mark schemes and grids showing the coverage of content areas.

These potential benefits can only accrue from the use of item response theory to calibrate and equate tests and from the implementation of an operational item bank that uses semi-automated test construction. They are key to the creation of improved tests in an efficient manner.

Curiously, the highest rated potential benefits from each of the two groups of participants revealed a disparity in the importance given to different aspects of semi-automated test construction. The senior examiners in cognate subjects responded with the automatic production of assessment grids being most important (see Table C8) – a task that is usually undertaken by subject administrators. Meanwhile the subject administrators gave their highest rating to the separation of test construction from item writing (see Table B8) – a change in emphasis which may have profound consequences for the way in which question paper setters work. At this point it cannot be emphasised too much that the sample sizes in this evaluation were small and therefore may not be considered to be representative.

CONCLUSIONS

This study has shown that it is feasible for a PEx to use a test construction interface connected to an item bank containing items that have been calibrated and tagged with metadata to construct tests that are similar in many respects to past question papers.

Issues for future item banking and test construction

Three issues that should be addressed for the successful implementation of item banking and semi-automated test construction within AQA have been identified as a result of this evaluation.

1. Training

This evaluation was perhaps fortunate in that its participants were all numerate. The statistical concepts underpinning item banking are complex. Staff who will be creating tests using this method need clear explanations of these concepts. Training that engages even the most innumerate in the use of item-test performance information, as well as the software of the item bank, is a prerequisite to implementing the semi-automated construction of tests. Additional training is also required in the operation of item banks. This should address issues around the immutability of items once banked and the use of metadata.

2. Additional metadata tags specific to components

The use of metadata in tagging items within an item bank when those metadata will be used in the construction of tests needs to be considered carefully. Obvious pieces of metadata are assessment objectives and content area. However, there appear to be other pieces of information about both items and tests which are relevant to constructing a test. Does the item contain a graph? What is the maximum mark? Does the item contain information in a table? Which precise skill is the item expecting the candidate to demonstrate? What is the context within which the question is placed? Further work is needed to identify these pieces of information and establish their usefulness, or otherwise, within the context of item banking.

3. Simplifying assessment objectives and content areas

The assessment objectives and content areas of the three components considered in this evaluation caused great difficulty. The descriptions of the assessment objectives and content areas, particularly for GCSE Mathematics B and Adult Numeracy, are broad and deep. Each can cover a number of skills. Because of this it is possible to have items in the bank which are assigned to more than one assessment objective or content area. These items may assess multiple assessment objectives or content areas simultaneously or different objectives or areas depending on where they best fit into the weightings grids.

The simplification of assessment objectives and content areas is a necessity for the construction of tests that are capable of assessing specific skills and that can also be demonstrated to assess these skills. To this end AQA needs to engage with the regulators, Ofqual and QCA, and other stakeholders within educational assessment as to the purposes of assessments which are likely to be entering the on-demand environment.

ACKNOWLEDGMENTS

Thanks are due to Tony Daly, Andrew Spreckley and Anne Trant for reviewing and constructively commenting on questionnaires and guidance documents. The contributions of colleagues from senior examining teams and in the Qualifications, Development and Support division as participants in this evaluation are greatly appreciated.

Claire Whitehouse	Senior Research Officer
Qingping He	Senior Research Officer
Chris Wheadon	Principal Research Manager

17 November 2008

APPENDIX A
Results of survey by questionnaire for Stage 1: Test construction

Table A1: Comparison by test constructors of their test construction experience using the prototype with their usual test construction experience (percentage all test constructors)

	Percentage responses (%)				Total N
	Less	More	Neither	No response	
Referred to the specification	66.7	0.0	33.3	0.00	3
Referred to past question papers	33.3	33.3	33.3	0.00	3
Referred to past mark schemes	33.3	33.3	33.3	0.00	3
Referred to past assessment grids	33.3	33.3	33.3	0.00	3
Revised the question paper based on information about percentage weighting of assessment objectives or content area	33.3	33.3	33.3	0.00	3
Revised the question paper based on statistical information about items (facility, difficulty, item-test correlation, discrimination)	66.7	33.3	0.0	0.00	3
Referred to statistical information about item performance	0.0	66.7	33.3	0.00	3
Wanted to re-write items/questions	0.0	66.7	33.3	0.00	3
Had difficulty deciding on the choice of topics to cover	33.3	0.0	66.7	0.00	3
Accessed information about assessment objectives and content areas (topics)	33.3	33.3	33.3	0.00	3
Accessed information about item facility (CTT)	33.3	33.3	33.3	0.00	3
Accessed information about item difficulty (IRT)*	0.0	33.3	33.3	33.3	3

* The test constructor for GCE Economics would have used this if it was available. Due to an error in data transfer the information was unavailable within the prototype during the evaluation.

Table A2: Speed of test construction using the prototype compared with usual test construction experience (percentage all test constructors)

	Percentage responses (%)			Total N
	Faster	Slower	Neither	
Speed of decision making	66.7	0.0	33.3	3
The time you spent drafting the question paper	100.0	0.0	0.0	3
The time you spent searching for particular items	66.7	0.0	33.3	3

Table A3: Test constructors' perception of their test construction experience using the prototype in comparison with their usual experience (mean rating all test constructors: 1 = less, 5 = more)

	Mean rating	Total N
Informative	4.67	3
Time efficient	4.67	3
Enjoyable	4.00	3
Straightforward	4.00	3
Focused	3.67	3
Complicated	3.00	3
Confusing	1.67	3

Table A4: Test constructors' confidence in the tests constructed using the prototype in comparison with tests constructed in the usual way (mean rating all test constructors: 1 = less confident, 5 = more confident)

Test constructed using prototype	Mean rating	Total N
contains content area (topic) coverage that is comparable with past papers?	4.67	3
is similar to past question papers?	4.00	3
contains appropriate proportions of content area (topic) coverage?	4.00	3
contains appropriate assessment objective coverage?	4.00	3
is able to discriminate amongst candidates of varying ability?	4.00	3
is comparable in demand with past question papers?	3.67	3
is not predictable in terms of questions and topics covered?	3.67	3
is capable of being answered in the time allowed?	3.67	3
would allow an average candidate to gain around 60% of the marks?	3.67	3

Table A5: Comments from test constructors on their test construction experience and the tests constructed using the prototype

Comment	Source (test constructor)
"*Much quicker way of producing a paper which is balanced and appropriate. As new questions are easy to produce by changing a simple number would like a very simple edit facility (also to be able to cut up questions). On the whole it soon became very easy to use. Need a much bigger bank then the process would be even quicker as initial searching would reveal exactly what was searched for.*"	GCSE Mathematics
"*Enjoyed the experience and can see how it is beneficial. Need ability to view hard copy of paper before completion of paper construction.*"	GCE Economics

Table A6: Test constructors' ratings of functionalities within the prototype item bank/test construction interface (percentage all test constructors)

		Percentage responses (%)			Total N
		Not useful	Useful	Very useful	
Availability of item performance statistics in general		0.0	66.7	33.3	3
Searching for items based on					
	- facility	33.3	0.0	66.7	3
	- mean difficulty and range	33.3	33.3	33.3	3
	- assessment objective	0.0	33.3	66.7	3
	- content area	0.0	0.0	100.0	3
Searching for items based on a combination of statistical data and qualitative information		0.0	33.3	66.7	3
Ability to create and store multiple tests		33.3	0.0	33.3	2

Table A7: Test constructors' satisfaction ratings with aspects of the prototype item bank/test construction interface (mean satisfaction rating: 1 = very dissatisfied, 5 = very satisfied, "No opinion" option was provided, but not used)

	Mean rating	Total N
Sizing of items, tests and statistical information	4.00	3
On-screen navigation	3.67	3
Screen layout	3.67	3
Quality of graphics	3.67	3
Accessibility of statistical information about item performance	3.67	3
Formatting of test, mark scheme and assessment grid	3.33	3

Table A8: Comments from test constructors on the functionalities of the prototype item bank/test construction interface

Comment	Source (test constructor)
"*Facility to store 'possible' questions on a separate screen or to compare questions on the same topic. Need some integration of/consideration of the 'Targets and Tariffs' document.*"	GCSE Mathematics

Table A9: Mean rank order of potential benefits of an item bank/test construction interface as perceived by the test constructors (rating scale of 1 = no benefit, 11 = great benefit converted to mean rank order)

	Mean rank order	Total N
Construction of tests with known and recorded discrimination properties	1	3
If on-line, improved question paper security as no longer sending disks or papers by post	2	3
Construction of tests with known and recorded difficulty/facility	3	3
Construction of tests, mark schemes and assessment grid pre-formatted to AQA's house style	= 4	3
Constructing tests based on statistical information about question performance	= 4	3
Separating test construction from item/question writing	6	3
Availability of a question paper setting schedule through the item bank/test construction interface	7	3
Not having to draft and check a separate assessment grid	8	3
Not having to draft and check a separate mark scheme	9	3
Availability of documents from Ofqual or AQA relevant to the component through the item bank/test construction interface	10	2
Availability of copyright documentation through the item bank/test construction interface	11	1

Table A10: Frequency of activity during usual test construction experience (percentage all test constructors)

How often do you	Low	Medium	High	Not used	Not available	Total N
use the specification during drafting of question paper(s)?	33.3	66.7	0.0	0.0	0.0	3
use past question papers and mark schemes during drafting of question papers?	33.3	66.7	0.0	0.0	0.0	3
use assessment grids from past question papers?	66.7	0.0	0.0	33.3	0.0	3
use statistical information about item performance?	33.3	33.3	33.3	0.0	0.0	3
use CMI+ Item Analyses?	0.0	33.3	0.0	0.0	66.7	3
write individual items/questions?	0.0	0.0	66.7	0.0	33.3	3
construct tests from an item bank?	0.0	0.0	33.3	0.0	66.7	3

Table A11: Test constructors' level of understanding of aspects of item banking after the evaluation (percentage all test constructors)

	Low	Medium	High	Not applicable	Total N
Statistical information about item performance	0.0	0.0	100.0	0.0	3
Writing of individual items, separate from a question paper	0.0	0.0	100.0	0.0	3
Test construction by compiling individual items	0.0	0.0	100.0	0.0	3
Item exposure rates	33.3	33.3	33.3	0.0	3

APPENDIX B

Results of survey by questionnaire from participants based in subject administration for Stage 2: Blind comparison

Table B1: Ratings for characteristics of question papers given to the NCT and a past question paper (mean rating all subject administrators: 1 = not at all, 5 = completely)

In my opinion the test …	NCTs		Past question papers	
	Mean rating	Total N	Mean rating	Total N
would allow an average candidate to gain around 60% of the marks	5.00	1	4.50	2
can be answered satisfactorily in the time allowed	4.33	3	4.33	3
contains content area (topic) coverage that is comparable with past papers or current versions	4.00	2	5.00	2
is similar to past question papers or current versions	3.33	3	4.00	3
is comparable in demand with past question papers or current versions	3.33	3	4.00	3
is not predictable in terms of questions and topics covered	3.33	3	2.33	3
is able to discriminate amongst candidates of varying ability	3.00	2	3.00	2
is free from bias	3.00	3	3.00	3
presents questions in the best order	3.00	1	4.00	1
provides a suitably demanding challenge for higher-achieving candidates	3.00	2	3.00	2
contains a balanced coverage of content areas	2.00	3	3.00	3
assesses the full range of skills and abilities as defined by the assessment objectives	2.00	2	3.00	2
contains an appropriate gradient of difficulty	2.00	2	3.50	2

Table B2: Ratings for the appropriateness of assessment objective weightings and content area weightings given to the NCT and a past question paper (mean rating all subject administrators: 1 = disagree, 5 = agree)

Using the information within the grids, in my opinion the test …	NCTs		Past question papers	
	Mean rating	Total N	Mean rating	Total N
contains appropriate content area (topic) weightings	3.00	3	3.33	3
contains appropriate assessment objective weightings	2.67	3	3.33	3

Table B3: Identification of NCTs by participants who are based in subject administration

Correct identification of NCT (% participants)		
Before viewing statistical information	After viewing statistical information	Total N
100%	100%	3

Table B4: Additional documents participants who are based in subject administration used or would have found useful to consult during the evaluation of the two question papers

Adult Numeracy	GCE Economics ECN1/1	GCSE Mathematics 43003F Part A
"*Specification.* *Adult Numeracy Core Curriculum booklet.*"	"*Test 2 seems to have a better spread of content areas and question types. I didn't use any additional documents. If I had Economics subject expertise, I would have used the specification and the assessment grid. You really need to put these questions to a subject expert/examiner, and supply the reference material too.*"	"*Specimen papers for 43003F*"

Table B5: Comments from participants who are based in subject administration on aspects of the two question papers they considered to be inappropriate or otherwise noteworthy

	NCTs	Past question papers
Adult Numeracy	"*1. Please see point 1 above.* *2. Questions repeated therefore narrowing the balance of subject content even further.* *3. What is or isn't the correct order?*"	"*1. Not all the subject content is tested and perhaps this should be looked into when versions are commissioned.* *2. What is or isn't the correct order?*"
GCE Economics ECN1/1	"*I've checked this test against the Test Construction Grid used by question paper setters. Test 1 should have been 6 items on 10.2 (not 7) 3 items on 10.6 (not 2). Facilities range from 0.280 to 0.792.*"	"*Checked as for Test 1. facilities range from 0.496 to 0.915*"
GCSE Mathematics 43003F Part A	"*With the possible exception of Q6, there is little opportunity in the test to demonstrate the ability to apply skills, to communicate in mathematics and to reason through a problem.*"	

Table B6: Ratings of additional characteristics considered when evaluating a question paper (mean rating all subject administrators: 1 = unimportant, 5 = important)

	Mean rating	**Total N**	**Not considered**
Whether the mark scheme shows the correct response for each question	5.00	2	1
Stimulus materials used in questions	4.50	2	0
The contexts of questions	4.33	3	0
Positioning of a question for potentially supplying correct answer to another question	4.33	3	0
Whether the mark scheme contains a range of responses that are mark worthy for a free text response question	4.33	3	0
Mark allocations for questions are visible on the question paper	4.00	3	0
Number of questions containing graphs	4.00	2	1
Number of questions requiring the use of tables	3.33	3	0

Table B7: Other characteristics of question papers considered by participants who are subject administrators

GCE Economics ECN1/1	**GCSE Mathematics 43003F Part A**
"*Topic covered within each content question. Eg. Section 10.2 should have 6 questions but these six need to be on different topics. For automatic selection of items to form QPs, the subject content needs to be more comprehensively categorised (eg within 10.2 we need 10.2.1, 10.2.2 etc)*"	"*An overall sense of balance across the paper considering length of questions, demand, context/no context*"

Table B8: Ratings given to potential benefits of an item bank/test construction interface (mean rating all subject administrators: 1 = unimportant, 5 = very important)

	Mean rating	**Total N**
Separating test construction from item/question writing	4.67	3
Availability of documents from Ofqual or AQA relevant to the component through the item bank/test construction interface	4.67	3
Construction of tests with known and recorded difficulty/facility	4.33	3
Construction of tests with known and recorded discrimination properties	4.33	3
Automatic construction of a grid showing the coverage of content areas	4.33	3
Automatic production of mark scheme with the question paper	4.33	3
If on-line, improved question paper security as no longer sending disks or papers by post	4.00	3
Automatic production of an assessment objective grid with the question paper	3.67	3
Constructing tests based on statistical information about question performance	3.33	3
Construction of tests, mark schemes and assessment objective grid pre-formatted to AQA’s house style	2.80	5
Availability of a question paper setting schedule through the item bank/test construction interface	2.80	5
Availability of copyright documentation through the item bank/test construction interface	2.40	5

APPENDIX C

Results of survey by questionnaire from participants who are senior examiners in a cognate subject for Stage 2: Blind comparison

Table C1: Ratings for characteristics of question papers given to the NCT and a past question paper (mean rating all cognate subject examiners: 1 = not at all, 5 = completely)

	NCTs		Past question papers	
In my opinion the test …	Mean rating	Total N	Mean rating	Total N
can be answered satisfactorily in the time allowed	4.80	5	4.80	5
contains content area (topic) coverage that is comparable with past papers or current versions	4.60	5	4.80	5
is similar to past question papers or current versions	4.20	5	4.00	5
is comparable in demand with past question papers or current versions	4.20	5	4.80	5
is free from bias	4.20	5	4.00	4
is able to discriminate amongst candidates of varying ability	3.80	5	3.60	5
presents questions in the best order	3.80	5	4.00	5
provides a suitably demanding challenge for higher-achieving candidates	3.80	5	3.80	5
contains an appropriate gradient of difficulty	3.80	5	4.00	5
contains a balanced coverage of content areas	3.60	5	3.60	5
assesses the full range of skills and abilities as defined by the assessment objectives	3.40	5	3.60	5
would allow an average candidate to gain around 60% of the marks	3.20	5	3.20	5
is not predictable in terms of questions and topics covered	2.60	5	3.00	5

Table C2: Ratings for the appropriateness of assessment objective weightings and content area weightings given to the NCT and a past question paper (mean rating all cognate subject examiners: 1 = disagree, 5 = agree)

	NCTs		Past question papers	
Using the information within the grids, in my opinion the test …	Mean rating	Total N	Mean rating	Total N
contains appropriate assessment objective weightings	3.60	5	3.20	5
contains appropriate content area (topic) weightings	3.00	5	4.00	5

Table C3: Identification of NCTs by participants who are senior examiners in a cognate subject

Correct identification of NCT (% participants)		
Before viewing statistical information	After viewing statistical information	Total N
80%	80%	5

Table C4: Additional documents participants who are senior examiners in a cognate subject used or would have found useful to consult during the evaluation of the two question papers

Adult Numeracy	**GCE Economics ECN1/1**	**GCSE Mathematics 43003F Part A**
"*Specification* *Specimen question paper* *Specimen mark scheme* *Adult Numeracy Core Curriculum* *Entry Level 2008 specification*" "*None*"	"*1 additional past paper & assessment grids for past papers*" "*A further past paper - basing all judgements on a single exemplar of such a paper*"	"*Very difficult to decide - I used targets and tariffs but this does not set grade in stone!*"

Table C5: Comments from participants who are senior examiners in a cognate subject on aspects of the two question papers they considered to be inappropriate or otherwise noteworthy

	NCTs	**Past question papers**
Adult Numeracy	"*A.O.3 under represented but better on this paper. Subject areas 5, 9, 12, 17, 19, - Printout did not give % for Part A 23 - 24 but appears to be very similar to test 1.*"	"*A.O. 3 would appear to be under represented. Subject areas 5, 9, 12, 17, 19, 23, 24, 25, 26, 27, 29, 30, 31, 32, 33 not covered.* *15/34 Rather a lot*"
GCE Economics ECN1/1	"*I do not agree with the classification of Q5. I consider it be covering syllabus section 10.2 and AO1.*" "*I'm not always convinced by the classification of skills - for example is Q5 really an analytical Q? Seems more like K & U. More Qs asked on section 10.2 here than on 08 paper. Section 10.4 looks short-changed.*"	"*I do not agree with the classification of Q15. I consider it to be covering syllabus section 10.2.*" "*Less emphasis on 10.2*"
GCSE Mathematics 43003F Part A	"*Q7 is higher grade than Q8 so switch*"	"*Q3 is usually the starter question. Q7, 8, 9 all tough!!*" "*Really not sure!!* *The difficulty is that the paper as a whole has to meet criteria. As this is one section both are out of tolerance.*"

Table C6: Ratings of additional characteristics considered when evaluating a question paper (mean rating all cognate subject examiners: 1 = unimportant, 5 = important)

	Mean rating	**Total N**	**Not considered**
The contexts of questions	4.8	5	0
Whether the mark scheme shows the correct response for each question	4.6	5	0
Mark allocations for questions are visible on the question paper	4.4	5	0
Stimulus materials used in questions	4.2	5	0
Positioning of a question for potentially supplying correct answer to another question	4.2	5	0
Whether the mark scheme contains a range of responses that are mark worthy for a free text response question	4.0	5	0
Number of questions requiring the use of tables	3.6	5	0
Number of questions containing graphs	3.2	5	0

Table C7: Other characteristics of question papers considered by participants who are senior examiners in a cognate subject

Adult Numeracy	**GCE Economics ECN1/1**	**GCSE Mathematics 43003F Part A**
"*Can the question be answered in the space provided. In both test 1 and test 2 candidates were asked to develop a graph but in both cases the development would have to be hindered by the position of the key.*"	"*Mark scheme makes clear how the different AO's will be rewardable i.e. nature of the answer required to gain marks under each AO.*"	"*Balance of topics. Novelty of questions, overall feel for level of difficulty, coverage of spec (over time) (using a tracking document). Our assessment grid meets criteria.*" "*The timing issues. The degree to which the context is accessible by students.*"

Table C8: Ratings given to potential benefits of an item bank/test construction interface (mean rating all cognate subject examiners: 1 = unimportant, 5 = very important)

	Mean rating	Total N
Automatic production of an assessment objective grid with the question paper	3.80	5
Construction of tests with known and recorded difficulty/facility	3.80	5
Construction of tests with known and recorded discrimination properties	3.80	5
Automatic construction of a grid showing the coverage of content areas	3.60	5
Automatic production of mark scheme with the question paper	3.40	5
Constructing tests based on statistical information about question performance	3.40	5
If on-line, improved question paper security as no longer sending disks or papers by post	3.00	5
Separating test construction from item/question writing	3.00	5
Availability of documents from Ofqual or AQA relevant to the component through the item bank/test construction interface	3.00	5
Construction of tests, mark schemes and assessment objective grid pre-formatted to AQA's house style	2.80	5
Availability of a question paper setting schedule through the item bank/test construction interface	2.80	5
Availability of copyright documentation through the item bank/test construction interface	2.40	5

APPENDIX D

Summary of statistical information for NCTs and past question papers used in the blind comparison

Table D1: Statistical information for tests for GCE Economics ECN1/1

Weightings of assessment objectives	**NCT**	**Past question paper**
AO1: demonstrate knowledge and understanding of the specified subject content	53.33%	53.33%
AO2: apply knowledge and critical understanding to economic problems & issues	33.33%	33.33%
AO3: analyse economic problems and issues	13.33%	13.33%
AO4: evaluate economic arguments and evidence	0.00%	0.00%
Weightings of content areas		
10.1 The Economic Problem	13.33%	13.33%
10.2 The Allocation of Resources in Competitive Markets	43.33%	36.67%
10.3 Monopoly	6.67%	6.67%
10.4 Production and Efficiency	6.67%	6.67%
10.5 Market Failure	13.33%	13.33%
10.6 Government Intervention in the Market	16.67%	23.33%
Classical test theory		
Mean facility	0.59	0.68
Mean discrimination	0.46	0.39
Approximate mean facility index by grade boundary for entire test expressed as percentage of maximum mark		
A	84.00	86.67
B	63.33	71.33
C	51.33	60.00
D	45.33	53.33
E	36.67	44.67

Table D2: Statistical information by item for tests for GCE Economics ECN1/1

	NCT					Past question paper				
Position of item in test	**AO**	**Content area**	**Facility**	**Equated difficulty**	**Discrimination**	**AO**	**Content area**	**Facility**	**Difficulty**	**Discrimination**
1	AO1	10.1	0.623	0.329	0.436	AO1	10.1	0.758	-0.373	0.398
2	AO1	10.1	0.484	0.972	0.471	AO1	10.4	0.646	0.219	0.399
3	AO2	10.6	0.640	0.212	0.439	AO2	10.2	0.655	0.179	0.373
4	AO1	10.6	0.792	-0.619	0.468	AO1	10.2	0.601	0.436	0.446
5	AO3	10.5	0.623	0.327	0.467	AO2	10.6	0.496	0.916	0.431
6	AO2	10.5	0.779	-0.503	0.368	AO3	10.5	0.755	-0.361	0.320
7	AO1	10.4	0.679	0.052	0.484	AO2	10.5	0.779	-0.503	0.368
8	AO3	10.2 & 10.6	0.638	0.261	0.465	AO1	10.1	0.726	-0.190	0.402
9	AO2	10.3	0.744	-0.351	0.495	AO1	10.2	0.814	-0.741	0.393
10	AO2	10.2	0.529	0.758	0.444	AO1	10.2	0.509	0.857	0.421
11	AO1	10.2	0.527	0.756	0.569	AO1	10.2	0.641	0.244	0.405
12	AO2	10.2	0.468	1.038	0.493	AO3	10.2 & 10.6	0.638	0.261	0.465
13	AO1	10.2	0.481	0.978	0.462	AO2	10.6	0.649	0.206	0.377
14	AO1	10.2	0.509	0.857	0.421	AO1	10.6	0.915	-1.693	0.296
15	AO1	10.2	0.280	1.962	0.387	AO2	10.3	0.578	0.544	0.387

Table D3: Statistical information for tests for Adult Numeracy

	Part A		Part B	
Weightings of assessment objectives	**NCT**	**Past question paper**	**NCT**	**Past question paper**
AO1: read and understand information given by numbers, symbols, diagrams and charts	57.69%	57.69%	39.51%	37.65%
AO2: generate results to a given level of accuracy using given methods, measures etc	23.08%	23.08%	58.02%	59.88%
AO3:present and explain results which meet the intended purpose using appropriate etc	19.23%	19.23%	2.47%	2.47%
Weightings of content areas				
1: Use whole numbers, fractions and decimals to measure and mark observations	15.38%	7.69%	0%	3.70%
1.1: Count, read, write, order and compare numbers up to 1000	0%	7.69%	3.70%	0%
1.2: Add or subtract using three-digit numbers	23.08%	23.08%	0%	0%
1.3: Recall addition and subtraction facts to 20	0%	0%	2.47%	2.47%
1.4:Multiply two-digit whole numbers by single-digit whole numbers	0%	0%	0%	0%
1.5: Divide two-digit whole numbers by single-digit whole numbers and interpret remainders	0%	0%	2.47%	4.32%
1.6: Recall multiplication facts, e.g. multiples of 2, 3, 4, 5, 10	0%	0%	2.47%	2.47%
1.7: Approximate by rounding numbers less than 1000 to the nearest 10 or 100	0%	0%	7.41%	3.70%
1.8: Estimate answers to calculations	0%	0%	0%	0%
1.9:Use and interpret +, -, x, / and = in practical situations for solving problems	23.08%	23.08%	1.85%	9.26%
1.10: Read, write and understand common fractions, e.g. 3/4, 2/3, 1/10	0%	0%	2.47%	1.23%
1.11: Recognise and use equivalent forms, e.g. 5/10 = 1/2	0%	0%	0%	0%
1.12: Read, write and understand decimals up to 2 decimal places in practical contexts	0%	0%	13.89%	12.04%
1.13: Estimate, calculate and compare money	0%	0%	12.96%	12.96%
1.14: Read, measure and record time	0%	0%	18.52%	18.52%
1.15: Read, estimate, measure and compare length, capacity, weight and temperature	0%	0%	8.33%	8.33%
1.16: Choose and use appropriate units and measuring instruments	0%	0%	0%	0%
2: Use space and shape to record information	0%	0%	7.41%	7.41%
2.1: Sort 2-D and 3-D shapes to solve practical problems using properties	0%	0%	0%	0%
3: Use numerical information from lists, tables, diagrams and simple charts to help understanding	0%	12.82%	8.95%	7.72%
3.1: Extract numerical information from lists, tables, diagrams and simple charts	12.82%	12.82%	0%	0%

Table D3: Statistical information for tests for Adult Numeracy *contd*

	Part A NCT	 Past question paper	Part B NCT	 Past question paper
3.2: Make numerical comparisons from bar charts and pictograms	12.82%	12.82%	0%	0%
3.3: Organise and represent information in different ways so that it makes sense to others	12.82%	0%	0%	0%
4: Make observations and record numerical information using a tally	0%	0%	0%	0%
4.1: Calculate using whole numbers and decimals to solve problems in context	0%	0%	0%	0%
4.2: Check calculations	0%	0%	0%	0%
5: Use given materials and methods	0%	0%	0%	0%
6: Present and explain results	0%	0%	5.25%	4.01%
6.1: Use whole numbers, common fractions and decimals to present results	0%	0%	0%	0%
6.2: Use common measures and units of measure to define quantities	0%	0%	0%	0%
6.3: Use tables, charts and diagrams to present results, e.g. for amounts	0%	0%	0%	0%
6.4: Use given methods to check results	0%	0%	0%	0%
6.5: Use given methods to present results	0%	0%	0%	0%
6.6: Use appropriate methods and forms to describe outcomes	0%	0%	1.85%	1.85%
Classical test theory indices				
Mean facility	0.91	0.89	0.81	0.81
Mean discrimination	0.40	0.47	0.42	0.37
Approximate mean facility index by grade boundary for entire test expressed as percentage of maximum mark				
Pass	92.31	96.92	86.67	88.67

Table D4: Statistical information by item for tests for Adult Numeracy

	NCT					Past question paper				
Position of item in test	AO	Content area	Facility	Equated difficulty	Discrim-ination	AO	Content area	Facility	Difficulty	Discrim-ination
Part A 1	AO1	1.9	0.926	-0.791	0.270	AO1 & AO2	1	0.783	0.550	0.679
2	AO1	1.2	0.873	-0.100	0.500	AO1	1 & 1.1	0.925	-0.732	0.293
3	AO1	3.1, 3.2, 3.3	0.893	-0.378	0.607	AO1 & AO2	1.9	0.855	0.119	0.533
4	AO1	1	0.925	-0.732	0.293	AO1 & AO2	1.9	0.905	-0.415	0.418
5	AO1	1.9	0.855	0.119	0.533	AO1 & AO2	1.9	0.940	-0.951	0.358
6	AO1	1.9	0.960	-1.507	0.212	AO1 & AO3	3.1, 3.2 & 3.3	0.942	-1.209	0.547
Part B 1	AO1	1.14	0.940	-1.039	0.487	AO2	1	0.970	-1.708	0.130
2	AO2	1.13	0.725	1.060	0.523	AO2	1.15 & 1.16	0.850	0.164	0.385
[34] 3	AO2	1.7	0.876	-0.157	0.341	AO1 & AO2	1.9	0.890	-0.235	0.324
4	AO1 & AO2	1.9 & 1.12	0.644	1.484	0.542	AO1 & AO2	1.5 & 1.9	0.850	0.164	0.313
5	AO2	1.15 & 1.16	0.850	0.164	0.385	AO1 & AO2	1.13	0.735	0.999	0.462
6	AO1	1.10, 3.1 & 6.1	0.945	-1.202	0.455	AO1 & AO2	1.12 & 1.13	0.770	0.775	0.376
7	AO2	1.1	0.985	-2.551	0.010	AO1 & AO2	1.12 & 1.13	0.690	1.264	0.552
8	AO1 & AO2	1.12 & 1.13	0.934	-0.988	0.240	AO2	1.3, 1.5 & 1.6	0.910	-0.482	0.311
9	AO1, AO2 & AO3	2.1	0.671	1.233	0.378	AO1 & AO2	1.9 & 1.12	0.765	0.808	0.480
10	AO2	1.15 & 3.1	0.720	1.113	0.403	AO1 & AO2	1.9 & 1.12	0.780	0.708	0.549
11	AO2	1.14	0.828	0.300	0.434	AO1 & AO2	1.12, 1.15, 3.1 & 6.1	0.787	0.552	0.567
12	AO1	1.10, 3.1 & 6.1	0.710	1.149	0.402	AO2	1.7	0.945	-1.049	0.135
13	AO2	1.12 & 1.13	0.570	1.893	0.528	AO1	1.10, 3.1 & 6.1	0.710	1.149	0.402
14	AO1 & AO2	1.12, 1.15, 3.1 & 6.1	0.849	-0.585	0.705	AO2	1.14	0.885	-0.180	0.198
15	AO2	1.3, 1.5 & 1.6	0.910	-0.482	0.311	AO1 & AO2	1.14	0.930	-0.918	0.283
16	AO2	1.7	0.855	0.057	0.483	AO1	1.14	0.875	-0.074	0.257
17	AO1 & AO2	1.12: & 1.13	0.713	1.151	0.464	AO1	1.14	0.880	-0.126	0.269

Table D4: Statistical information by item for tests for Adult Numeracy *contd*

	NCT					**Past question paper**				
Position of item in test	**AO**	**Content area**	**Facility**	**Equated difficulty**	**Discrimination**	**AO**	**Content area**	**Facility**	**Difficulty**	**Discrimination**
18	AO1 & AO2	1.15 & 3.1	0.840	0.252	0.327	AO2	1.13	0.725	1.060	0.523
19	AO1	1.14	0.870	-0.028	0.507	AO1, AO2 & AO3	2.1	0.755	0.504	0.330
20	AO1 & AO2	1.12 & 1.13	0.770	0.775	0.376	AO1 & AO2	1.15 & 3.1	0.840	0.327	0.327
21	AO2	1.12 & 1.13	0.708	1.119	0.598	AO2	1.15 & 3.1	0.670	0.408	0.408
22						AO2	1.12 & 1.13	0.570	0.528	0.528

Table D5: Statistical information for tests for GCSE Mathematics 43003F Part A

Weightings of assessment objectives	NCT	Past question paper
AO1 Using and Applying Mathematics	0%	0%
AO2-NA-1: Using and Applying Number and Algebra	23.44%	14.58%
AO2-NA-2: Numbers and the Number System	25.00%	30.21%
AO2-NA-3: Calculations	29.69%	48.96%
AO2-NA-4: Solving Numerical Problems	21.88%	6.25%
AO2 - NA-5: Equations, Formulae and Identities	0%	0%
AO2-NA-6: Sequences, Functions and Graphs	0%	0%
AO3-SSM-1:Using and Applying Shape, Space and Measures	0%	0%
AO3-SSM-2: Geometrical Reasoning	0%	0%
AO3-SSM-3: Transformations and Coordinates	0%	0%
AO3-SSM-4: Measures and Construction	0%	0%
AO4-HAD-1: Using and applying handling data	0%	0%
AO4-HD-2: Specifying the Problem and Planning	0%	0%
AO4-HD-3: Collecting Data	0%	0%
AO4-HD-4: Processing and Representing Data	0%	0%
AO4-HD-5: Interpreting and Discussing Results	0%	0%
Weightings of content areas		
AO1: Using and Applying Mathematics	0%	0%
AO2-1: Problem Solving	6.25%	6.25%
AO2-1: Communicating	17.19%	5.21%
AO2-1: Reasoning	0%	2.08%
AO2-2: Integers	6.25%	8.33%
AO2-2: Powers and Roots	3.13%	6.77%
AO2-2: Fractions	0%	5.47%
AO2-2: Decimals	0%	0%
AO2-2: Percentages	9.38%	5.21%
AO2-2: Ratio	3.13%	0%
AO2-3: Number Operations and the Relationships Between Them	17.19%	16.93%
AO2-3: Mental Methods	3.13%	12.50%
AO2-3: Written Methods	9.38%	9.38%
AO2-3: Calculator Methods	3.13%	10.16%
AO2-4: Solving Numerical Problems	21.88%	11.72%
AO2-5: Use of Symbols	0%	0%
AO2-5: Quadratic Functions	0%	0%
AO2-5: Index Notation	0%	0%
AO2-5: Inequalities	0%	0%
AO2-5: Equations	0%	0%
AO2-5: Linear Equations	0%	0%
AO2-5: Formulae	0%	0%
AO2-5: Simultaneous Linear Equations	0%	0%
AO2-5: Quadratic Equations	0%	0%
AO2-5: Simultaneous Linear and Quadratic Equations	0%	0%
AO2-5: Numerical Methods	0%	0%

Table D5: Statistical information for tests for GCSE Mathematics 43003F Part A *contd*

Weightings of assessment objectives	**NCT**	**Past question paper**
AO2-6: Sequences	0%	0%
AO2-6: Graphs of Linear Functions	0%	0%
AO2-6: Gradients	0%	0%
AO2-6: Interpreting Graphical Information	0%	0%
AO2-6: Quadratic Functions	0%	0%
AO-6: Other Functions	0%	0%
AO2-6: Transformations of Functions	0%	0%
AO2-6: Loci	0%	0%
AO3-2: Angles	0%	0%
AO3-2: Properties of Triangles and Other Rectilinear Shapes	0%	0%
AO3-2: Properties of Circles	0%	0%
AO3-2: 3-D Shapes	0%	0%
AO3-3: Specifying Transformations	0%	0%
AO3-3: Properties of Transformations	0%	0%
AO3-3: Coordinates	0%	0%
AO3-3: Vectors	0%	0%
AO3-4: Measures	0%	0%
AO3-4: Constructions	0%	0%
AO3-4: Mensuration	0%	0%
AO4-2: Specifying the Problem and Planning	0%	0%
AO4-3: Collecting Data	0%	0%
AO4-4: Processing and Representing Data	0%	0%
AO4-5: Interpreting and Discussing Results	0%	0%
AO2-5: Direct and inverse proportion	0%	0%
Classical test theory indices		
Mean facility	0.55	0.45
Mean discrimination	0.55	0.57
Approximate mean facility index by grade boundary for entire test expressed as percentage of maximum mark		
C	79.38	69.06
D	62.81	52.19
E	48.13	37.81
F	40.31	27.81
G	30.94	20.63

Table D6: Statistical information by item for tests for GCSE Mathematics 43003F Part A

	NCT					**Past question paper**				
Position of item in test	**AO**	**Content area**	**Facility**	**Equated difficulty**	**Discrim-ination**	**AO**	**Content area**	**Facility**	**Difficulty**	**Discrim-ination**
1	AO2-NA-2	AO2-2	0.820	-1.375	0.336	AO2-NA-2 & AO2-NA-3	AO2-2, AO2-3, AO2-3 & AO2-4	0.741	-0.879	0.646
2	AO2-NA-1 & AO2-NA-4	AO2-1 & AO2-4	0.890	-2.466	0.527	AO2-NA-1, AO2-NA-2 & AO2-NA-3	AO2-1, AO2-2 & AO2-3	0.443	1.152	0.739
3	AO2-NA-1 & AO2-NA-4	AO2-1 & AO2-4	0.739	-0.874	0.528	AO2-NA-2	AO2-2	0.820	-1.375	0.336
4	AO2-NA-1 & AO2-NA-3	AO2-1 & AO2-3	0.562	0.184	0.485	AO2-NA-3	AO2-3	0.361	1.570	0.599
5	AO2-NA-2 & AO2-NA-3	AO2-2, AO2-3, AO2-3 & AO2-3	0.599	-0.091	0.626	AO2-NA-2 & AO2-NA-3	AO2-2 & AO2-3	0.368	1.510	0.521
6	AO2-NA-1 & AO2-NA-3	AO2-1 & AO2-3	0.339	1.171	0.677	AO2-NA-1 & AO2-NA-2	AO2-1, AO2-2 & AO2-2	0.498	0.845	0.562
7	AO2-NA-2 & AO2-NA-3	AO2-2 & AO2-3	0.196	1.881	0.489	AO2-NA-3	AO2-3	0.360	1.540	0.563
8	AO2-NA-2, AO2-NA-3 & AO2-NA-4	AO2-2, AO2-3 & AO2-4	0.285	1.210	0.746	AO2-NA-1 & AO2-NA-4	AO2-1 & AO2-4	0.396	1.320	0.613
9						AO2-NA-3	AO2-3 & AO2-3	0.074	3.830	0.420