Centre for Education Research and Policy

ITEM BANKS AND ON-DEMAND TESTS

Claire Whitehouse

Introduction

The aim of this chapter is to provide an overview of the fundamental principles of item banking. Item banking is a means to producing better questions (or items) and question papers (or tests) through the use of up to date statistical information about how tests and items perform when sat by candidates. The use of this alternative to the conventional process of question paper setting can also improve the efficiency of the production of examination materials; provide for the controlled re-use of questions (or items); and, lead to a reduction in the time between candidates sitting an examination (or test) and receiving their results. Item banking can be used to support new models of assessment that begin to move away from offering one or two examinations per year on fixed dates. The destination of such a journey is on-demand testing which offers the ultimate in flexibility and speed. This chapter discusses the differences between the conventional test-led approach to question paper setting and the item-led approach that item banking is able to facilitate.

The ability of item banking to lead to improved efficiencies in question paper setting is strongly dependent on computing power. Sophisticated electronic technologies are needed to store large numbers of items, information about each item and to apply statistical procedures to analyse the results of tests. This chapter assumes that such technologies are available, but does not describe them.

Tests versus items: two approaches to setting question papers

The conventional approach to setting question papers or tests involves a number of steps in a process that may last up to two years from the initial concept for the test to the point at which candidates sit that same test: see Figure 1(a). In this test-led approach the senior examiner charged with writing the test takes the first step in the process by deciding which topics from the subject to include and how they should be assessed. Using the outcomes of these decisions the senior examiner drafts and revises an entire test. A committee of subject experts evaluates the draft of the test, which may then undergo further revisions. Once a consensus is reached on such issues as accuracy, accessibility and relevance the draft goes through a number of proof stages before a question paper is printed on paper and ready for despatch to schools. Quality assurance is applied to the entire test at each stage. Each task in this process is scheduled to take place at a certain point in the calendar year to ensure that the question paper is ready for the day stated in the examination timetable. After the test has been sat by the candidates and the paper-based scripts marked, grade boundaries are set during the awarding process. Candidates receive their results many weeks after sitting their examinations. For those candidates with results from high stakes examinations, such as GCSEs and GCEs in the UK, and their teachers, there follows a comparatively short period of time in which they must make life-changing decisions based on these results.

The test-led approach to question paper setting is linear with a start and an end. There is a feedback loop in the middle which is based on human judgement of how the test may perform rather than statistical information about how items or tests have actually performed. If candidates' responses are analysed using statistical procedures, the results are of limited use as they are rarely able to influence the contents and format of the next test, or even the one

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

after that. Test writers do not receive timely information about how individual items performed which would help them to improve their item and test writing skills. Items and tests tend to be used once and then formally released to the public domain as past papers for students to practise with.

The test-led approach is suited to an environment in which examinations are offered on paper once or twice a year. There are pressures in the form of regulation, government policies, improving technology and the changing needs of candidates and teachers that are forcing awarding bodies to try alternative approaches to producing tests. One of these alternative approaches is item banking and the item-led approach to test construction, which is described diagrammatically in Figure 1(b).

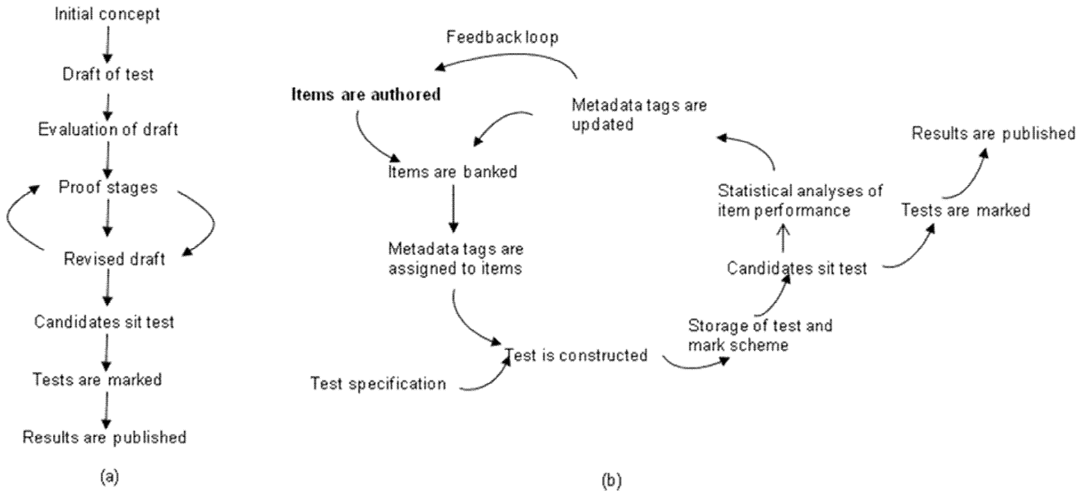

Figure 1: Comparison of (a) the test-led approach to question paper setting and (b) the item-led approach to test construction

Philosophically, using item banking to set tests is very different to the test-led approach. In the item-led approach subject experts author, review and revise items, not tests. The item becomes the unit of work. An item contains a number of elements, the demand being only one of these. Other elements that may be present are: a rubric; a stem; background information (photographs, video, audio, diagrams, graphs, tables, descriptions, maps *etc.*); maximum mark, correct response; and, explanations as to why the correct response is correct and incorrect responses are incorrect.

A basic electronic item bank receives deposits of items from the subject experts at any time. The bank stores the items in an accessible manner whilst value is added to them by attaching descriptions of their qualitative and quantitative characteristics. Together these characteristics are known as metadata which are tagged to an item. Items that are suitable for withdrawal from the bank for use in a particular test are identified by searching the bank using the metadata tags. The basic function of item and metadata storage in a virtual environment can help to reduce the time between the creation of a test and when candidates sit it. It also allows changes to a test close to the examination day or even the replacement of an entire test to be done rapidly and with comparative ease.

Qualitative metadata include information about the elements that make up the item, authoring details, assessment objective that the item addresses and content area or topic covered by the item. Tags for these are set during item authoring, before the items are selected for use in a test. Quantitative metadata are the result of statistical analyses carried out on the performance of items in live tests. The results of the statistical analyses include values for

- summary statistics (mean marks and standard deviations) for items and tests
- how easy or difficult items are (item facility index and item difficulty). An item with a high value for facility index is easy in comparison with other items in the same test. A high value for item difficulty, however, indicates an item that is hard for candidates to score marks on in comparison with other items in the item bank. In this case the item bank stores items for use in one test.
- how well items differentiate between candidates of high and low ability (discrimination index). A high discrimination value indicates that the item is able to differentiate well between candidates of differing abilities.

These numerical characteristics are a measure of the interaction between candidates and items under test conditions.

Before the performance of items can be measured a test must be constructed for candidates to sit. Similarly to test-led question paper setting, tests constructed from an item bank use a specification or template that contains a high level of detail. This sets out the format of the test including rubrics, guidance for candidates and the start and end points for sections within the test. It also states the percentages of each assessment objective that are required within a test and whether there are certain topics that must appear in the test. Unlike test-led question paper setting, a test construction programme or algorithm that can be used to search the metadata that are relevant to the test specification and selects suitable items for use in a test. The algorithm is similar to a cookery recipe in that it is a set of step-by-step instructions for building a test that will meet the criteria laid out in the specification. The test construction algorithm may be fully automated with the human input being limited to starting the algorithm running and a visual check of the resulting test. Alternatively, it may be semi-automated. In this case a test constructor uses the algorithm to carry out searches of the item bank through the metadata tags, select items and place them in to a test. Usually the algorithm selects more than one suitable item. The test constructor chooses the best item from the selection based on a set of heuristics that have not yet been incorporated into the algorithm. Provided the item bank is stocked with a reasonable quantity of suitable items it may take only a day to construct a few tests. Mark schemes and assessment grids (sometimes known as item maps) are compiled automatically at the same time as tests because all the information needed for these two documents is stored with the items. Once constructed, the tests and their mark schemes are stored in an electronic repository until needed. At this point the tests are either delivered electronically for candidates to sit on-screen or printed for a traditional pencil-and-paper test.

As with any test, once it has been sat it is marked, graded and the results are sent to the candidates and published. A distinct advantage that item-led test construction has over the test-led approach is that information about the performance of tests and items is fed back into the item bank with very little delay. Metadata tags for the quantitative characteristics are updated to take account of the most recent test results. The feedback of the results of the statistical analyses is extended to item authors, usually in a formal environment. In this way authors of items are able to use the most recent performance data to amend draft items and write new items that are better suited to the assessment. This does not mean that items become easier as time progresses. Rather, items can be written to target specific abilities within the range of abilities present in a group of candidates sitting a test. Thus a test is better able to differentiate amongst candidates because its most important constituents, the items, are designed to perform their task more accurately.

The item-led approach to test construction from item banks that incorporate statistical information on the items enables the construction of tests with known difficulty and provides the assurance that, within stated confidence limits, a test is comparable to previous and future tests. When new items with no measurements of difficulty are used in a test, estimates based on past experience may be used, but their inaccurate nature should be flagged. The tagging of items with statistical information about their past performance in one or more tests also allows grade boundaries to be set in advance of the test being sat by candidates. The one proviso here is that statistical information is attached to all of the items in a test. On construction of a test the statistical information for each item is added together to produce information about the test as a whole. This test information includes the most appropriate grade boundaries, taking into account the likely ability range of the candidates who will sit the test. Awarding bodies are able to use the pre-set boundaries to return results to candidates almost immediately after they have sat a test, assuming that the items in the test lend themselves to automatic marking. With grade boundaries that are pre-set the post-test awarding process becomes redundant. It is possible with this model to move to on-demand testing which provides tests when candidates want them by creating the tests automatically from an item bank.

Item Response Theory: the power behind the statistical analyses

A basic electronic item bank is a storage facility for items and their metadata. It is the statistical analyses that enhance a basic item bank to the point where it is a useful tool for managing items and constructing tests of known difficulty and pre-set grade boundaries. Statistical analyses are applied to the responses that the candidates make to the items in a test. The results of these analyses are tagged to items as the quantitative metadata. The analyses are often based on Item Response Theory (IRT), also known as latent trait theory. IRT encompasses a number of mathematical models of which probably the most well know is the Rasch model. IRT was developed originally to handle items with a maximum mark of one (dichotomous items), making it suitable for multiple-choice type items. More recently, researchers have extended its application to polytomous items, that is, those with maximum marks ranging from 2 to 8.

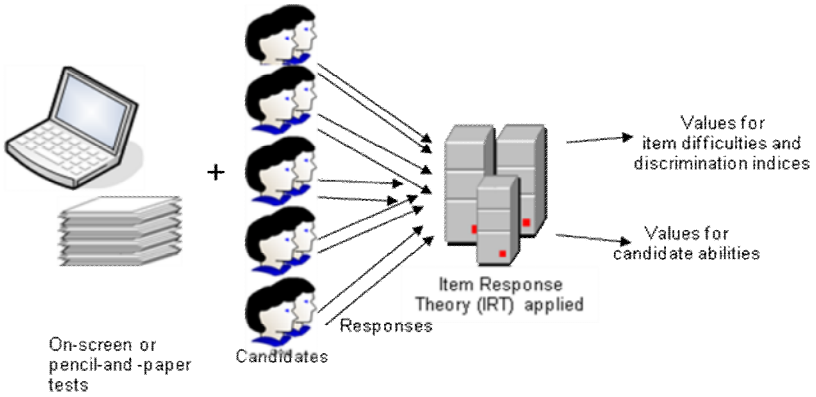
+
Values for item difficulties and discrimination indices
Values for candidate abilities
Item Response Theory (IRT) applied
Responses
On-screen or pencil-and -paper tests
Candidates

Figure 2: The application of IRT to candidates' responses

IRT models assume that the probability of a correct response to an item may be described by a mathematical equation that combines terms for item difficulty, item discrimination and the ability of the candidate. The ability of a candidate is determined by their total score on the test. By plugging candidates' scores for each item, which are calculated from their responses to items, into the equation the values for the difficulties of items are separated out from the abilities of candidates: see Figure 2 for a pictorial representation of this process. The separation of the two sets of values means that the item difficulties are no longer dependent on the ability of those candidates who responded to that set of items in a particular test sat at a particular time. In IRT terms this is known as placing the items on the same scale or the calibration of items. This is IRT's main advantage over Classical Test Theory which uses facility index to describe the performance of items. The facility index of an item is only comparable to the facility indices of items in the same test sat by a specific cohort of candidates.

Once calculated using IRT an item's difficulty should be the same each time the item is used in a test, no matter that the candidates are different or that the surrounding items have changed. In some circumstances, changing motivation of candidates, possible time pressures or changing the position of an item in a test, the value for item difficulty drifts away from the original value. The identification of items that show drift and consideration of the reasons for this drift is part of the management of the item bank.

As with any theory, IRT is based on some assumptions about what happens whilst candidates sit tests. Of these assumptions two stand out as being important for the design of tests. The first of these assumptions is that the items in a test measure only one ability or skill. This is the assumption of unidimensionality, where a dimension is another word for ability. To start with this was viewed as meaning, quite literally, one ability only, for example, the ability to carry out simple mathematical calculations such as addition and subtraction. Further research revealed that tests nearly always measure the intended ability plus other abilities, such as test taking ability, mathematical ability in a test of chemistry and quality of written communication in almost any test. Thus, in reality a test measures a composite of abilities that are relevant to a subject. An item bank that contains items with appropriate tagging assists greatly in the construction of comparable tests that measure a composite of abilities. However, there is still a risk that the presence of items testing a specific ability within a composite may advantage or disadvantage one or more sub-groups of candidates. Therefore, the quantitative metadata attached to items need to be monitored for such effects.

The second important assumption is that the responses given by candidates to any one item in a test are not reliant on the responses that they give to any other item(s) in the same test. For example, one item or the range of possible responses to it, should not give away the correct response to another item. This is termed the local independence of the characteristics of items. Local independence is necessary to ensure that an item is worthy of the marks stated in the mark scheme. The preferred strategy to ensure local independence is to write items that are unconnected to each other. Sometimes there is no educational benefit to this strategy and items are connected, perhaps by being set on a common theme or using the same stimulus material or context. These items and a candidate's responses to them can be grouped together and treated as one polytomous item with one response. This aggregation of responses reduces the reliability of the results of the statistical analyses and needs to be weighed against any benefit accrued from connecting items.

If these two assumptions are not met then the use of an IRT model for analysis purposes is undermined as the values of difficulty and discrimination become less reliable. To prevent this, test constructors need to ensure that the items selected for a test measure the composite ability that the test is intended to measure and that all the items are independent of each other.

Maintaining standards in a world of items

Even when tests are constructed to be as similar as possible, whether that is done by using human judgement based on experience or the results of statistical analyses of the results of past tests, there will be some differences in the statistical characteristics of both items and tests.

These differences may be attributed to one test being more difficult or easy than the other or to differences between the groups of candidates sitting the tests. The aim of the standard setting process is to ensure that from test to test candidates of equal ability are rewarded to the same extent by accounting for such differences.

In an assessment environment where the test-led approach to question paper setting dominates standards tend to be maintained between tests by setting grade boundaries after candidates have sat the examination. The standard setting process involves human judgement, statistical forecasting or a mixture of both. Where national high stakes tests, such as GCSEs and GCEs, are concerned, the setting of grade boundaries is a resource intensive process that takes place over a relatively short duration to ensure that deadlines for the reporting of results are met.

The scores on different tests are not directly comparable. In the item-led approach to test construction, the use of IRT enables the comparison of tests by a statistical operation known as equating. Test equating adjusts the scores from the different tests so that the performance of two groups of candidates sitting two tests measuring the same ability can be compared. This is facilitated through a set of items that are common to both tests. The two tests may be sat on the same day in different locations, one test per location, or sat at different times separated by gaps ranging from hours to months.

For test equating purposes, the rule of thumb is that there should be at least 20% of items on each of the two equivalent tests that are common. Sample sizes of a minimum of a few hundred candidates are also recommended. There are a number of designs for equating tests that incorporate the use of common items. One of these designs uses an anchor test for the common items; see Figure 3.

Figure 3 shows test 1 containing 40 items being sat by candidates in group 1 and test 2, containing a different set of 40 items being sat by a different set of candidates in group 2. Both tests also contain an additional 20 items in the anchor test, so all candidates respond to 60 items.. So long as the candidates in groups 1 and 2 attempt the items in the anchor test and the difficulties of the items in the anchor test are already known, the two tests can be equated and grade boundaries set on both tests that take account of the abilities of the candidates and the difficulties of the items.

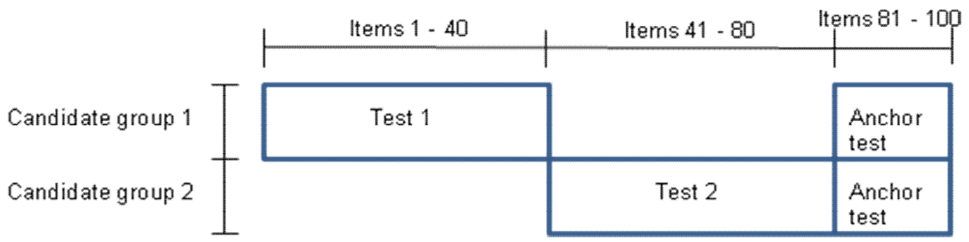

Figure 3: An equating design that uses an anchor test

The test equating design in Figure 3 allows grade boundaries to be set after candidates have sat both tests. It is not the best design for maintaining item security, especially if large cohorts of candidates sit the tests. Once the candidates have seen the items, they cannot be used again until the candidates in those cohorts have completed their programmes of study and will not re-sit the tests. There are other equating designs that maximise the security of items, especially new items whose performance in tests needs to be calibrated and placed on the same scale as items that have been used previously. This is pre-testing of new items.

Pre-testing of items to measure their performance raises two important issues. The first of these is whether to pre-test in live tests. This is known as live seeding and is cost effective as it piggy-backs on the administrative arrangements made for operational tests. It also has the advantage that items are calibrated under live test conditions. Pre-testing in non-live test conditions is expensive, time-consuming and less reliable. The students who take part are usually not representative of live test candidates and are less motivated.

The second issue is whether to include candidates' scores on the pre-test items in their total scores for the test. These items are new and lack statistical information about their performance, hence the need to pre-test. When a pre-test item counts towards the total score for a test and is not flawed, it is considered to part of a fair test. A pre-test item is problematic when it does not count towards the final score and is either flawed or its targeted ability was misjudged. Only statistical analyses are able to confirm this. Candidates may spend too much time responding to such a pre-test item, preventing them from doing as well as they are able to on the actual test items. There is no one solution to this problem as the following examples illustrate, other than that the chosen strategy should be disclosed to candidates.

The UK Driving Standards Agency calibrates new items by adding a few in a separate section at the end of the driving theory test. Candidates are offered the choice of whether or not to respond toe these items which do not contribute to their final score. In contrast, the providers of the GRE[1] incorporate new items within the test and request in their literature and on their website that candidates respond to all items and do not try to guess which items are the new ones that will not in actual fact count towards their final scores.

Item banking: what is it good for?

Item banking offers solutions to some of the challenges presented when providing high stakes assessments in a market place that is coming to terms with consumer choice. These challenges are outlined below.

1. The UK government's personalisation agenda in education. As learning and assessment programmes are tailored more to the individual learner, it is possible that assessment will become decoupled from the fixed examination series with which we are familiar. Provided an item bank is well stocked, the item-led approach to test construction can reduce greatly the amount of time needed to create a test so that tests can be offered with greater frequency.
2. The provision of greater choice and improved services to customers (learners, teachers, schools and colleges) through increasing the availability of tests at times that are better suited to learners and test centres. One way of achieving this is to deliver more tests on-screen to prevent a gap developing between students' everyday learning experiences and what happens during an examination. A second way is to decrease the time between sitting a test and receiving the result.
3. The use information technology to meet the regulators' drive for efficiency and flexibility in assessment such that results are timely and accurate to facilitate decision-making for learners, employers and tertiary education providers. Another by-product could be the streamlining of the awarding process by moving standard setting from the end of the question paper setting process to the beginning. The maintenance of standards becomes a process of continual monitoring of the performance of items and tests.

[1] The Graduate Record Examinations, a set of tests for entrance to postgraduate courses in the USA.

4. Increase the reliability of tests through the use of timely item performance data.

Item banking may also pave the way for on-demand tests. Ideally, on-demand tests are made available by test providers (awarding bodies) through test centres (schools, colleges, workplaces, authorised test centres) to individual learners

- at a time of the learner's choosing
- in a place of the learner's choosing.

These choices may be made by the learners themselves or, in conjunction with their teachers, parents or school or college management.

On-demand testing requires tests to be available to candidates when they are ready to sit them. Tests can be constructed from an item bank as soon as a candidate's registration for a test is received in preparation for that candidate to sit the test immediately. With pre-set boundaries, candidates are able to receive the results of their test soon after they have finished it.

There are different levels of on-demand testing. The bubble chart in Figure 4 places tests relative to each other in terms of two factors. The first factor is the number of places in which the tests are administered, with high stakes assessments (GCEs and GCSEs) being sat currently in schools and colleges. In a fully networked virtual world candidates will take tests anywhere, through any interface. The second factor is the number of opportunities to take a test. Tests for GCEs and GCSEs take place in one, two or three examination series held at fixed points in the academic year. Examination series are test windows: periods of time during which candidates take tests under secure conditions. For some tests, for example, the multiple-choice tests (MCTs) offered as part of AQA's GCSE Sciences, there is some leeway within a test window for a number of test sessions to be scheduled to ease the logistical burden on test centres. High stakes assessments in the UK are usually made available in a test window with one test session. In an ideal on-demand environment candidates will take tests at any time that is suited to their needs and the requirements of the test centre. The academic year may become a historical curiosity.

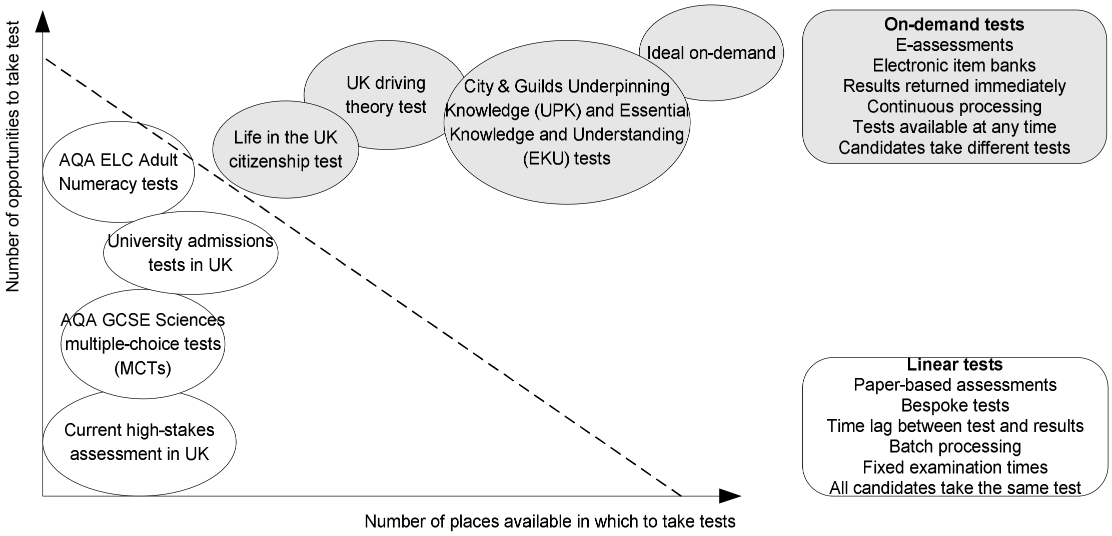
Number of opportunities to take test
Number of places available in which to take tests
Ideal on-demand
UK driving theory test
City & Guilds Underpinning Knowledge (UPK) and Essential Knowledge and Understanding (EKU) tests
Life in the UK citizenship test
AQA ELC Adult Numeracy tests
University admissions tests in UK
AQA GCSE Sciences multiple-choice tests (MCTs)
Current high-stakes assessment in UK
On-demand tests
E-assessments
Electronic item banks
Results returned immediately
Continuous processing
Tests available at any time
Candidates take different tests
Linear tests
Paper-based assessments
Bespoke tests
Time lag between test and results
Batch processing
Fixed examination times
All candidates take the same test

Figure 4: Levels of on-demand testing and the characteristics of on-demand and linear testing[2]

[2] The sizes of the bubbles in this chart are not intended to convey information about the relative numbers of tests.

The progress towards on-demand testing will be gradual, with the management of item banks and item authoring being established first. Constructing tests manually from item banks by using metadata is likely to be implemented next. Semi-automated and automated test construction with algorithms will take place as the amount of statistical information stored in the item bank increases. Low stakes formative assessments lend themselves to automated test construction and have the advantage of not requiring the same large numbers of items as banks that supply high stakes tests do, nor the same levels of security.

Management of item banks

Item banks need to be managed. There are three primary functions of item bank management. These are (1) populating the item bank; (2) monitoring the performance of items and tests; and, (3) maintaining item and test security.

How many calibrated items should an item bank contain? The answer is dependent on a number of factors. These factors include (1) the number of items in an operational test; (2) the frequency of tests; and, (3) whether the tests are high or low stakes. For existing tests that are switching from the test-led approach to the item-led approach to test construction items may be authored straight into an item bank. Item bank managers commission item authors to write items that are suitable for the qualification. Throughout the lifetime of the qualification the manager monitors the performance of items and commissions further items to fill any gaps in the bank. These may be gaps in the coverage of topics or assessment objectives, but they can also be gaps in the ability range that tests are targeted at.

Strategies for populating an item bank with calibrated items should be part of the development of a new qualification. This is the case particularly with on-demand tests. Sufficient items need to be written to ensure that the item bank is large enough to offer the number of tests needed.

It is also possible to re-use items from a legacy qualification in a new qualification. If the items from the legacy qualification already have difficulty values, then they can be used to continue the standards from legacy to new. An anchor test equating design would be optimal for this situation (see Figure 3).

Item re-use is dependent on the performance of items in live tests. Therefore, it is important that item performance is monitored. In particular an item bank manager should look for items that need to be retired, items that are being used too much in tests and items that are not being used sufficiently. Candidates and teachers may become familiar with items that appear frequently in tests causing the difficulties and discriminations of these items to drift. Items that are rarely used, but are otherwise satisfactory, can lead to apparent gaps in coverage which item authors may be asked to fill unnecessarily.

Item security is linked to item exposure which needs to be controlled in on-demand testing. When tests are delivered on-screen, candidates need to rely on their memories to reproduce items, rather than referring to the hard copy of a pencil-and-paper test. So, malicious exposure by individuals is greatly reduced, although such item exposure through the concerted efforts of a group of individuals is still a possibility. Equating designs can aid in controlling item exposure.

A consequence of this emphasis on item security is that past tests cannot be released into the public domain as this would compromise the integrity of future tests if the items remained active. If the items were retired prior to regular releases of past tests, the turnover of items within the bank would be unsustainable. Therefore, on-demand tests require different strategies for communicating with learners and teachers. A limited number of practice tests illustrating each type of item, strategies for responding and their correct responses replace the regular publication of past tests. Learners and their teachers are provided with clear descriptions of the purpose and structure of tests so that there are no surprises during the tests. The metadata can also provide information about candidates' performance by topic area, rather than by item, to highlight areas of strength and weakness in preparation for the next round of testing.

There is a fourth management function, which is not directly related to items. This is the development of qualifications that are suitable for administration from an item bank. Such qualifications are likely to be delivered entirely on-screen. The current model of test specification used in high-stakes assessment in the UK provides weightings for the assessment objectives to be covered in a test and a pool of content and skills from which question paper setters choose when drafting a question paper that meets the weightings for the assessment objectives laid down in the specification. The assessment objectives tend to be broad descriptions: knowledge, understanding, evaluation. Content is rotated through the series of examinations, with some topics being neglected because they are difficult to assess or do not quite fit with the weightings of the assessment objectives.

There are also subject-specific requirements. For example, in GCSE Mathematics specific mathematical operations are allocated a number of marks. GCSE Sciences MCTs provide a second example as there must be a certain number of 'how science works' questions that cut across all assessment objectives. The more constraints that an automated or semi-automated test construction algorithm has to meet, the larger and more varied the item bank needs to be and the more complex test construction becomes. For even semi-automated test construction, on-demand tests require assessment objectives and content topics that are clear and precise. However, this precision has the potential to lead to tests that assess smaller and smaller chunks of a qualification without creating the cognitive links between the chunks. Care needs to be taken to ensure that guidance is offered to teachers and learners on how to synthesise their chunks of learning. This is especially so if assessment of learning is by synoptic tests.

Summary

The item-led approach to test construction using item banking has the following benefits. These benefits can be realised for most assessment models, starting with tests that are offered once or twice a year, all the way to on-demand tests.

1. Tests of known difficulty can be constructed which means tests are comparable before they are sat by candidates. Comparable tests that are easily assembled are most suited to on-demand testing.
2. Grade boundaries can be pre-set when tests are of known difficulty. This removes the need for an awarding process, allowing candidates to receive the results of their tests immediately.
3. The continuous cycle of item banking, test construction and feedback of item performance data smooths out the peaks and troughs of the work load of item authors and item bank managers in comparison with that experienced during test-led question paper setting.
4. With items and their associated metadata stored in one location, the management of the item authoring and test construction processes can be done with real time information. The disadvantage of storing everything in one virtual location is the need for strong security protocols.