Maintaining standards in on-demand testing using item response theory

Qingping He
February 2010
Product code: Ofqual/10/4724

Contents

Summary

This report builds on findings from a research project on on-demand testing commissioned by Ofqual and carried out by the Assessment and Qualifications Alliance (AQA) and looks at the implications of on-demand testing for maintaining standards of examinations in general qualifications offered by UK awarding organisations. The study has investigated various technical issues associated with the implementation of on-demand testing, including item pre-testing and analysis, item bank development and the grading of components and qualifications. A model involving the use of item response theory (IRT) for maintaining standards at both component and qualification levels has been proposed for implementing on-demand testing in general qualifications.

Introduction

Although on-demand testing is being increasingly used in many areas of assessment, it has not been adopted in high stakes examinations like GCSEs and GCEs in England. (For a definition of on-demand testing, see Wheadon *et al*, 2009; Boyle 2010.) In order to develop a regulatory framework for on-demand testing, Ofqual commissioned a research project to AQA to look into the current status of on-demand testing and issues associated with its application in general qualifications in the UK (Boyle, 2010). Findings from the project have been included in a recently published report (available at http://www.ofqual.gov.uk/2525.aspx). The research has identified a number of drivers that could promote the use of on-demanding, but it also identified a number of barriers to its adoption in high stakes examinations. The nature of on-demand testing – including the use of multiple versions or forms of the same test in different test sessions throughout the year and the variability in the ability distribution and the number of candidates at different times – creates many difficulties for awarding organisations. One of the major issues is that some of the methods used for maintaining standards in conventional testing are no longer available (see discussions below), and a new definition of the meaning of maintaining standards in on-demand testing situations and the development of new methods that suit on-demand testing would be required.

Building on findings from the research conducted by AQA, this study looks at the implications of on-demand testing for maintaining standards for general qualifications offered in the UK, investigates the various technical issues associated with its implementation and explores new approaches to maintaining standards over time in on-demand testing.

This report sets out an approach to maintaining standards using IRT in on-demand testing. It is a detailed technical monograph that enunciates a range of issues that a

person charged with maintaining standards in an awarding organisation would have to address. The purpose of this paper is to make plain the issues that need to be addressed and to propose a method for tackling them. It does not imply that this is the only approach that one could take to maintain standards in an on-demand world, nor that this approach is mandated by the regulator. However, it is proposed that this is a coherent and complete approach to maintaining standards, which could be adapted by people to suit their own needs and circumstances.

Approaches to maintaining standards in conventional testing

Conventionally, examinations of general qualifications are taken during a small number of testing windows in a year (mainly in January and June), and continuity of standards over time at both component/unit and qualification levels is normally achieved through the use of expert judgement and statistical information (see Baird, 2007; Robinson, 2007). Cresswell (2000) suggested that the definition of standards in a particular examination would require considering two aspects:

- what should be assessed
- the levels of attainment that are comparable to those represented by each grade in other examinations.

He argued that the aspect of public examination standards that concerns what should be assessed changes over time as a result of change in the curriculum and other related areas in education resulting in change in examination standards. With regard to comparability of grade standards between examinations, Cresswell argued that it is difficult to determine the attainment required for the same grade to be comparable with that from different examinations, particularly examinations on different subjects. Baird (2007) has provided a comprehensive evaluation of the various definitions of comparability of examination standards and their interpretation and operational implications.

Robinson (2007) has provided a detailed description of the procedures that are commonly used by awarding organisations to set and maintain standards, which generally involves the use of both expert judgement and statistical information (see also discussions by Cresswell, 2000; Baird, 2007). The use of statistical information can be broadly classified into two categories: cohort-referencing and candidates' prior achievement.

In the case of cohort-referencing, entries to the current examination series are assumed to be similar to those in previous series in terms of ability distribution, and

therefore the proportion of candidates in each grade for the current series will be similar to that in previous series at both component/unit and qualifications levels. This is used to set component/unit and overall qualification grade boundaries. Cohort-referencing is suitable for situations where candidate entries for a qualification are stable over time.

In the case of using candidates' prior achievement, the relationships between candidates' grades and prior achievement established for previous series is assumed to be applicable to the current series and these relationships are used to set grade boundaries. It should be stressed that both cohort-referencing and prior achievement are used for the whole cohort or population. It should also be noticed that the current practice is to ensure comparability of standards over time at the overall qualification level, and that component/unit boundaries taken just before the awarding are frequently changed to achieve this. Component standards may therefore not be maintained in such situations.

In on-demand testing, candidates in different test sessions will take different forms or versions of the same test that will contain different questions. Although questions in different forms are assumed to measure the same knowledge and skills, the difficulty level may be different. Because the number of candidates in each test session will only comprise a small proportion of the population and the prior achievement information for some candidates will not be available, the two statistical approaches (cohort-referencing and prior achievement) in maintaining standards in conventional testing outlined above will not be valid in on-demand testing situations.

Meaning of maintaining standards in on-demand testing and approaches to comparability

Use of IRT for educational measurement

IRT assumes that, given a test and the examinee sample, the overall performance of an examinee is determined by his/her ability and the characteristics of the test items (Wright and Stone, 1979; Lord, 1980; Hambleton *et al*, 1991; Sijtsma and Junker, 2006). An item response model can therefore be developed to predict the probability that an examinee will answer a particular item correctly based on his/her ability and the characteristics of the item. Once an item response model has been established for a test, values for both the ability scores for examinees and item parameters and other test statistics can be estimated from candidates' responses. The better a model fits the test data, the more reliable are the estimated results. An important feature of IRT is that when test data meets model assumptions, item parameter estimates can be treated as invariant over different samples drawn from the population.

One of the most important applications of IRT modelling is to analyse the performance of test takers and test items for diagnostic purposes. For example, Hartig and Höhler (2008) have used multidimensional item response theory (MIRT) to model performance in complex domains, simultaneously taking into account multiple basic abilities. Wu and Adams (2006) have examined students' responses to mathematics problem-solving tasks and applied a general MIRT model at the response category level. They used IRT modelling to identify cognitive processes and extracted additional information. They have demonstrated that MIRT models can be powerful tools for extracting information from responses from a limited number of items by looking at the within-item multidimensionality. Through the analysis, they were able to understand how students interacted with the items and how the items were linked together and therefore to construct problem-solving profiles for students.

Another important application of IRT modelling in psychological and educational testing is equating test forms using either common items built into the forms or common persons taking the different forms (see for example, Wright and Stone, 1979; Hambleton *et al*, 1991; Kolen, 1999; Kolen and Brennan, 2004; Yu and Popp, 2005; also see later discussions). Test equating with IRT places items in different forms on the same measurement scale and therefore makes it possible for examinees taking different forms to be compared in terms of the underlying ability trait (Wright and Stone, 1979; Hambleton *et al*, 1991). Test equating has many practical applications, including the construction of calibrated item banks, in which all items are of known difficulty, and the delivery of IRT-based testing (Wright and Stone, 1979; Rudner, 1998; Ward and Murray-Ward, 1998; Umar, 1999). For example, the monitoring of students' learning progress has received increasing attention from policy makers and education practitioners in recent years, and a prerequisite of this is that tests taken at different times must be compared in terms of difficulty, which can be achieved through test equating (see, for example, Chin *et al*, 2006). A calibrated item bank can be used in a wide range of applications. For example, computer adaptive testing (CAT), which targets items at the individual ability level of the examinees, has been increasingly used in many areas of assessment and a large calibrated item bank is essential for its successful implementation (see, for example, Masters and Evans, 1986; Masters and Keeves, 1999; van der Linden, 1999; Weiss and Schleisman, 1999; Wheadon and He, 2006; Hol *et al*, 2007). It is frequently required to construct tests targeted at specific ability groups of examinees and this can be achieved by using items from a calibrated bank.

With regard to examinations standards in on-demand testing for general qualifications, the report by Wheadon *et al* (2009) suggested that the following principles should be adopted:

i. *Decisions to move each syllabus to on-demand testing should be supported by a clear educational case. This case should have a sound theoretical basis and be supported by the teaching profession.*
ii. *On-demand testing should be underpinned by Item Response Theory methods of test-equating.*
iii. *Policies on item to test ratio, item re-use, pre-test procedures and evidence of coherence of scales should all be available.*
iv. *Where items are re-used, item parameters should be monitored for unexpected changes over time or between versions that may indicate security breaches, drift, over-use or changes in testing conditions such as reduced time available for question completion.*
v. *Systems should be in place to monitor and help explain changes in aggregate qualification outcomes over time.*
vi. *The reliability of tests should be such that there is little to gain from repeated re-sitting.*

In view of the nature of on-demand testing, the wide application of IRT in educational measurement, particularly in test equating, and the availability of complex IRT analysis software, coupled with increasing IRT expertise in awarding organisations, the use of IRT could prove to be one of the viable approaches to maintaining standards over time or between test sessions in an on-demand environment.

The on-demand testing architecture

For on-demand testing to be effective, tests must be designed by selecting items from a large calibrated item bank that contains valid and reliable items. Umar (1999) discussed the advantages of using item banks for the development and operation of assessment systems. These advantages include economy, flexibility, consistency, and enhanced security. Figure 1 depicts an on-demand testing architecture proposed by Wheadon *et al* (2009).

It is envisaged that an on-demand testing system would contain the following key components:

- An item bank containing calibrated items (i.e. items with known characteristics). The item bank should have the following sub-components:
  - an item creation component for item writers to write and review items

- a test design component for designing specific tests meeting specified objectives required by test centres – this component may need to possess IRT-related functions
- a marking component for marking candidates' responses in the case that responses cannot be marked by the test delivery system instantly
- an analysis component for scoring, analysing items and grading if required – this component may need to possess IRT-related functions.

- A registration system for centres to register their students for test sessions. This system may need to be able to access the item bank.
- A test delivery system for delivering tests on-screen to centres. This system may also need to possess IRT-related functions and functions for instant marking, scoring and grading. This system will need to be able to access the item bank.

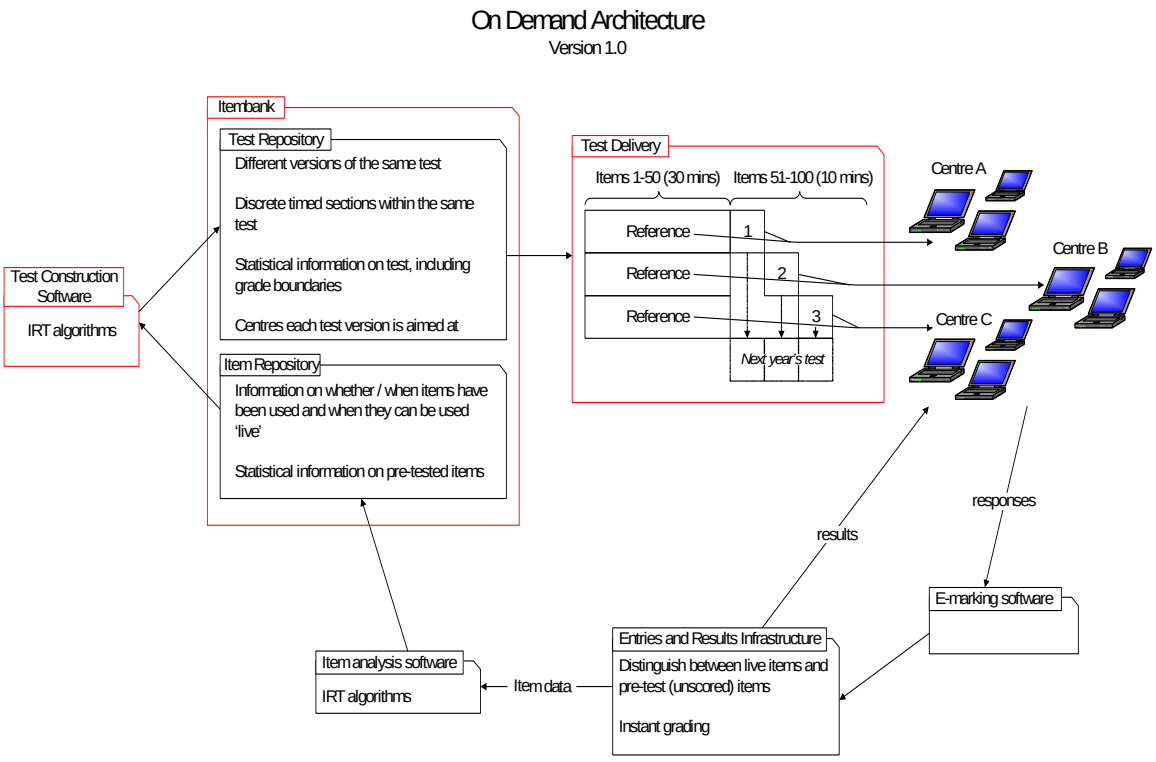

Figure 1 An on-demand testing architecture (adapted from Wheadon *et al*, 2009).

Meaning of maintaining standards and implications for comparability

As indicated earlier, one of the possibilities to address the difficulty in maintaining standards in on-demand testing is the use of IRT in linking and equating different test forms. The following procedures will need to be taken to achieve this:

- pre-testing items
- analysing items and equating test forms using an IRT model to place all items on the same common scale so that they are directly comparable in terms of the constructs they measure and the difficulty level they represent
- establishing a calibrated item bank for a component/unit using the pre-tested items
- establishing (component) grade boundaries using the IRT measurement scale
- constructing tests for specific test sessions according to pre-specified criteria
- administering tests to candidates
- scoring candidates' responses using IRT procedures and grading candidates using pre-established grade boundary values; standards are automatically maintained this way.

The implications of using the above procedures are as follows:

- The meaning of maintaining standards in on-demand testing are fundamentally different from that in conventional testing. While in conventional testing, maintaining standards is associated with ensuring the continuity of outcomes over time for similar populations (at component or qualification level), maintaining standards in on-demand testing using IRT requires the continuity of the underlying latent ability of the candidates being measured.
- While the focus of maintaining standards over time in conventional testing is placed on the overall qualification (even for unitised qualifications, standards over time at component/unit level may not be maintained), on-demand testing maintains standards at both component/unit and qualification levels (if all components in a qualifications are assessed using on-demand testing).

Item banking using IRT

General procedures for developing an IRT-based item bank

As indicated earlier, the establishment of an on-demand testing system will involve the development of a calibrated item bank in which item characteristics are known and are on the same measurement scale. Tshering (2006, also see Eggen, 2007) has identified seven steps for developing a calibrated item bank that have been used as the basis for developing the following procedures for IRT-based item banking:

- identification of the goals and objectives of the item bank to be developed
- the deployment of appropriate personnel to develop high quality items and performing item-content matching to reflect the knowledge and skills to be assessed; Items should be properly classified using tags such as assessment objectives, topic areas, knowledge and skills, etc
- creating test forms that meet the assessment objectives and content topic coverage to assess the intended knowledge and skills
- pre-testing items through the use of different test forms across a wide range of abilities using representative samples
- calibration of items from different test forms and equating the forms to align item parameters on a common scale by using suitable IRT models and the evaluation of model fit statistics – items violating model assumptions (large misfit) should be excluded from the item bank
- the development of an item bank database
- storing information associated with the calibrated items selected for inclusion into the database
- replenishment of the item bank with new calibrated items
- training of users of the item bank so that they have a good knowledge of the package and other related computer packages used in item banking.

Some of these procedures will be discussed in detail in the following sections.

Item types and implications for on-demand testing

There are generally four major types of items used in general qualifications in the UK:

- *Objective items*: Items of this type can be marked objectively. These include multiple choice questions, multiple selection questions, matching questions and others. Computers can be used to mark this type of questions effectively.

- *Structured response questions*: These are short answer questions that require candidates to response freely. Depending on the nature of the answers, these questions can be marked either by computers or by human markers.
- *Essays*: Candidates are required to write long passages to demonstrate their understanding, application and evaluation skills. Questions of this type are generally marked by humans although extensive research has been undertaken to study automatic essay marking.
- *Portfolios*: These are a collection of tasks accumulated over a long period of time and are generally assessed by humans.

Question type and format have important implications for implementing an on-demand testing system. For example, objective questions are relatively easy to write, and tests consisting of objective questions can be marked by computers. These questions are particularly suitable for computer adaptive testing and instant feedback or grading. It is not possible to provide instant feedback or grading for tests that are composed of questions requiring human marking.

UIRT models

There are two types of IRT models, unidimensional IRT (UIRT) models, for items measuring a single ability in common, and multidimensional IRT (MIRT) models for items measuring multiple abilities (see Reckase, 1985, 1997; Ackerman, 1992, 1994; 1996; Adams and Wilson, 1996; Adams *et al*, 1997; Embretson, 1996, 1997; Embretson and Reise, 2000; Reckase and Martineau, 2004; Sijtsma and Junker, 2006; Wu and Adams, 2006). Frequently used UIRT models include the 1PL (the Rasch model), 2PL and 3PL models for analysing dichotomous items, and the partial credit model (PCM) and the rating scale model (RSM) for analysing polytomous items (Rasch, 1960; Andrich 1978; Wright and Stone, 1979; Lord, 1980; Masters, 1982; Wright and Masters, 1982). The 3PL model for dichotomous items can be expressed as (Lord, 1980):

$$P(\theta) = c + (1-c)\frac{\exp(Da(\theta - b))}{1+\exp(Da(\theta - b))} \quad (1)$$

where:

θ = person ability in logits

D = 1.7

$P(\theta)$ = probability for a person with an ability θ to answer the item correctly

a = the item discrimination parameter

b = the item difficulty parameter

c = the item guessing parameter.

Equation (1) is also called item response function (IRF or item characteristic curve or ICC) and indicates that the probability of an examinee answering an item correctly increases with an increase in his/her ability or a decrease in item difficulty. When c = 0, it becomes the two-parameter logistic model. When a = 1 and c = 0, Equation (1) reduces to the Rasch model. For the Rasch model, when the item difficulty is close to the person ability, the test taker will have a 50% chance of answering the item correctly.

The partial credit model for polytomous items developed by Masters can be expressed as (Masters, 1982, 1984, 1999; Wright and Masters, 1982; Masters and Evans, 1986):

$$P(\theta, x) = \frac{\exp \sum_{k=1}^{x} (\theta - \delta_k)}{1 + \sum_{x=1}^{m} \exp[\sum_{k=1}^{x} (\theta - \delta_k)]} \qquad (2)$$

where:

$P(\theta, x)$ = the probability of a person with ability θ scoring x on a polytomous item with a maximum score m

δ_k = the k^{th} threshold location of the item on the latent trait continuum which is also referred to as the item step difficulty

m = maximum score of the item.

Adams *et al* (1997; also see Wu *et al*, 2007) have developed an IRT model called the unidimensional random coefficients multinomial logit model (URCMLM) for both dichotomous and polytomous items:

$$P(X_j = 1 | \theta) = \frac{\exp(b_j \theta + \overrightarrow{a'_j} \overrightarrow{\xi})}{\sum_{k=1}^{K} \exp(b_k \theta + \overrightarrow{a'_k} \overrightarrow{\xi})} \qquad (3)$$

where:

$P(X_j = 1 | \theta)$ = the probability of a person with ability θ whose response is in category *j* (with a category response value of $X_j = 1$)

K = the total number of categories of the item

b_j = the response score representing the performance level of an observed response in category *j*

$\vec{\xi}$ = the item parameter vector (in the form of a matrix with the number of elements equal to the number of item parameters)

$\vec{a}$ = the design vector (matrix) reflecting the nature of the model (for example dichotomous or polytomous, 1PL or PCM)

$\vec{a'}$ = the transpose of $\vec{a}$.

The URCMLM includes a number of IRT models such as the Rasch model, the partial credit model, the rating scale model and a few other models.

Maraki (1992; 1997) expanded the PCM to the generalised PCM (GPCM) to include item discrimination parameters in addition to difficulty parameters or category thresholds. Other commonly used UIRT models included Samejma's (1969) graded response model (GRM) that models the cumulative category response function and Bock's (1972) nominal model (also see Kolen and Brennan, 2004).

Two important assumptions are required under these unidimensional IRT models: unidimensionality and local independence. Unidimensionality requires that one ability or a single latent variable is being measured by the test. Local independence requires that test takers' responses to any question in the test are statistically independent when the ability influencing their performance on the whole test is held constant. In situations where items having the same stimulus are analysed as individual items, the assumption of local independence could be violated. Such items can be grouped to form a testlet and treated as one polytomous item and analysed using polytomours IRT models (Thissen *et al*, 1989; Yen, 1993).

MIRT models

When a test is designed to measure more than one latent variable, which is frequently the case given that a test needs to meet certain validity criteria such as required content or curriculum coverage, multidimensional IRT models can be used (see, for example, Reckase, 1985, 1997; Embretson and Reise, 2000; Ackerman,

1994, 1996; Reckase *et al*, 2005; DeMars, 2006; Yao and Schwarz, 2006). Given the features of public examinations in the UK, MIRT models would seem to be particularly suitable. The widely used compensatory multidimensional 3PL model can be viewed as an extension to the unidimensional 3PL model and can be expressed as:

$$P(\vec{\theta}) = c + (1-c)\frac{\exp(\vec{a'}\vec{\theta} - b)}{1+\exp(\vec{a'}\vec{\theta} - b)} \quad (4)$$

In a compensatory MIRT model, a test taker's low ability in one dimension can be compensated by high ability in other dimensions when answering questions. Although the definitions of the variables in Equation (4) are similar to those for Equation (1), both the item discrimination parameter $\vec{a}$ and the latent trait $\vec{\theta}$ are vectors represented using matrices (both have the same number of elements which is the number of ability dimensions). $\vec{a'}$ is the transpose of $\vec{a}$.

The URCMLM discussed previously has also been extended to the multidimensional random coefficients multinomial logit model (MRCMLM) to accommodate the multidimensionality of items in a test (Wu *et al*, 2007):

$$P(X_j = 1 | \vec{\theta}) = \frac{\exp(\vec{b_j'}\vec{\theta} + \vec{a_j'}\vec{\xi})}{\sum_{k=1}^{K}\exp(\vec{b_k'}\vec{\theta} + \vec{a_k'}\vec{\xi})} \quad (5)$$

where $\vec{\theta}$ is the latent trait vector, $\vec{b_j}$ is the category score vector for category j, $\vec{b_j'}$ is the transpose of $\vec{b_j}$, and $\vec{a_j'}$ is the transpose of $\vec{a_j}$.

Wang (1995) and Adams *et al* (1997) have introduced the concepts of between-item multidimensionality and within-item multidimensionality in MIRT modelling to assist in the discussion of different types of multidimensional models and tests (see also Wu *et al*, 2007). In the case of between-item multidimensionality, each item in a test measures just one distinct ability dimension and items measuring the same latent dimension are grouped to form a subscale (for example, a mathematics test may contain some items measuring geometry ability and some items measuring algebra ability). In this case, the actual MIRT model reduces to a combination of unidimensional IRT models for individual item groups. In the case of within-item multidimensionality, an item may measure more than one latent dimension. Between-

item multidimensionality is easier to deal with mathematically and computationally than within-item multidimensionality.

Item calibration, test information, and evaluation of model assumptions and model fit

Once a test has been taken by test takers, responses can be analysed using IRT software to produce estimates for the item parameters identified by the specified IRT model and the person ability measures. IRT places both item and person measures on the same measurement scale. It should be noted that the item response function is a function of both person ability trait and item parameters. For most IRT models, the IRF is invariant when certain linear transformations are applied to both person ability trait and item parameters, and this creates indeterminacy when establishing the IRT scale (i.e. the origins of person trait and item parameters cannot be determined independently). This model indeterminacy in IRT scaling has an important impact on IRT equating (see later discussions). One possible way to deal with this for the Rasch model is to set the average difficulties of the items in the test to be zero to determine values for item parameters. The procedures implemented in item calibration by most IRT analysis software systems are the joint maximum likelihood estimation (JMLE), marginal maximum likelihood estimation (MMLE) or the conditional maximum likelihood estimation (CMLE) (Embretson and Reise, 2000; Linacre, 2008; Simon, 2008).

An important concept in IRT modelling is the item information function (IIF) and test information function (TIF). TIF $I(\theta)$ for a UIRT model is defined as the sum of item information function over all items in the test:

$$I(\theta) = \sum_{i=1}^{N} I_i(\theta) \tag{6}$$

where N is the number of items in the test, and $I_i(\theta)$ is the item information function of item i. The standard error (standard deviation) σ_θ of a person ability measure is inversely proportional to the test information:

$$\sigma_\theta = \sqrt{\frac{1}{I(\theta)}} \tag{7}$$

Since the test information is a function of person ability, the standard error of person ability is also a function of ability.

As indicated earlier, the use of IRT models for analysing tests requires that test data meet model assumptions. Items that violate those assumptions should be excluded from the item bank. This is important because the stability of item parameters will to a certain extent be affected by the degree that the item meets the model assumptions (such as the unidimensionality and local independence for UIRT models). Analysis should be conducted to identify the reasons why certain items do not meet model assumptions, and this would improve the development of high quality items. The use of an IRT model also assumes that the model is correct. As indicated by Barnard (1999), an evaluation of how well the model chosen fits the data is essential in IRT modelling to ensure the usability of test results. Embretson and Reise (2000) outlined procedures for assessing the fit of IRT models.

Basic principles for equating test forms and properties of equating

As discussed earlier, in many situations, multiple forms or versions of the same test are used for reasons like security. Different forms are administered to different groups of test takers and the results from all test takers are compared. Although test developers try to make those alternative forms as similar as possible in terms of content coverage, knowledge and skills to be assessed, and the difficulty level of the test, they are not strict parallel forms. The comparability between test scores from different forms must be established so that candidates can be compared fairly. In classical test theory (CTT), this is normally achieved through the process of test equating which establishes a mathematical relationship between the test scores so that they can be used interchangeably regardless which form of the test a test taker has taken. Procedures involved in test equating generally include:

- developing test forms
- developing test data collection designs
- administering tests and collecting responses
- establishing the relationship between scores from different forms using statistical procedures.

Assuming that there are two forms X and Y of the same test taken by two groups of test takers, the goal of equating is to establish a mathematical function so that test score can be transformed from one form (say form Y, the equating form or new form) to the other (say form X, the base form or reference form). Mathematically, this can be expressed as (Kolen and Brennan, 2004; van der Linden, 2005a):

$$Y_X = F_{CTT}(Y) \quad (8)$$

where Y is the score that a test taker obtained on form Y and Y_x is the score on form X if he/she took form X. The actual form of the transformation function F_{CTT} is dependent on the assumptions made about the ability distributions of the test takers between the two groups (Livingston, 2004). It should be noted that F_{CTT} is established for the two groups of test takers and used to transform test scores for individual test takers. Livingston (2004) compared CTT-based test equating with IRT-based test equating in terms of the advantages and weakness of the two methods. IRT equating is more flexible in terms of equating designs for linking test forms, but it is complex, both conceptually and procedurally (see later discussions).

Kolen and Brennan (2004) (also see Lord, 1980; Petersen *et al*, 1989; Yu and Popp, 2005) outlined some of the desirable properties of equating relationships. These include the following:

- *Same specification*: The two test forms are built to the same content and statistical specifications, so that they measure the same underlying trait or construct.
- *Equity*: For examinees of identical performance level on the underlying trait or true score, the conditional frequency distribution of scores on form Y (including observed score means and standard deviations), after transformation, must be the same as the conditional frequency distribution of scores on form X.
- *Group invariance*: The equating relationship must be the same regardless of the group of examinees from which it was derived.
- *Symmetry*: The equating transformation must be symmetric or invertible, that is, the mapping of scores from form X to form Y must be the same as the mapping of scores from form Y to form X. This requires that the function used to transform scores on form X to the form Y scale be the inverse of the function used to transform scores on form Y to the form X scale.

Test equating design

Bèguin (2000), Kolen and Brennan (2004) and Livingston (2004) discussed some of the most widely used equating designs and their advantages and disadvantages in CTT and IRT equating. These include the following:

- *The single group design:* This the simplest equating design, which involves the same test takers taking both the new form and the reference form. It is assumed that in this case the equating relationship between test scores derived from the

group of test takers applies to the target population that are going to use the new form. The equating relationship is directly and explicitly established for test scores on the two forms.

- *The counterbalanced design*: This design is to remove the effect of ordering the test forms and involves dividing the test-takers into two groups and 'counterbalancing' the order in which the groups take the two forms. One group takes the new form first and the reference form second; the other group takes the reference form first and the new form second. The results from the two groups on the two occasions are combined to establish the equating relationship.

- *The equivalent groups design*: It can be difficult in situations for the same examinees to take the two forms of the test and the equivalent-groups design can be used. This design involves separating examinees into two groups that are equal in terms of the knowledge and skills to be measured by the test. The two equivalent groups then take the two forms separately. This design also assumes that equating relationship derived from the two groups can be applied to the target population.

- *The internal anchor design*: In this design, the new form contains items (anchor items or common items) from the reference form and is given to the examinees. Performance on the anchor items is used to establish the equating relationship between the new form and the reference form. Scores on anchor items are counted into the total test scores on the new form.

- *The external anchor design*: The anchor items form an anchor which is separate from the new form itself (scores on anchor will not be counted towards the total score on the new form). Performance on the anchor items is used to establish the equating relationship between the new form and the reference form.

- *Pre-equating non-equivalent groups design*: The reference form of the test is administered with subsets of new items to different groups of test takers. The new items are then analysed and used to compose new forms of the test.

In the case of IRT equating, the designs outlined above can also be conceptualised as two broad designs: common item design and common person design. In the case of common item design, two forms containing a subset of common items of the test are given to two groups of examinees which have similar ability distributions, and the performances on the common items are compared to establish the equating relationship which is used to transform test scores between the two forms. In the case of common person design, the two forms do not contain common items and are given to the same group of examinees, the scores on the two tests are directly comparable and are used to establish the relationship between items parameters in

the two forms. Figure 2 provides a graphic representation of the two equating designs.

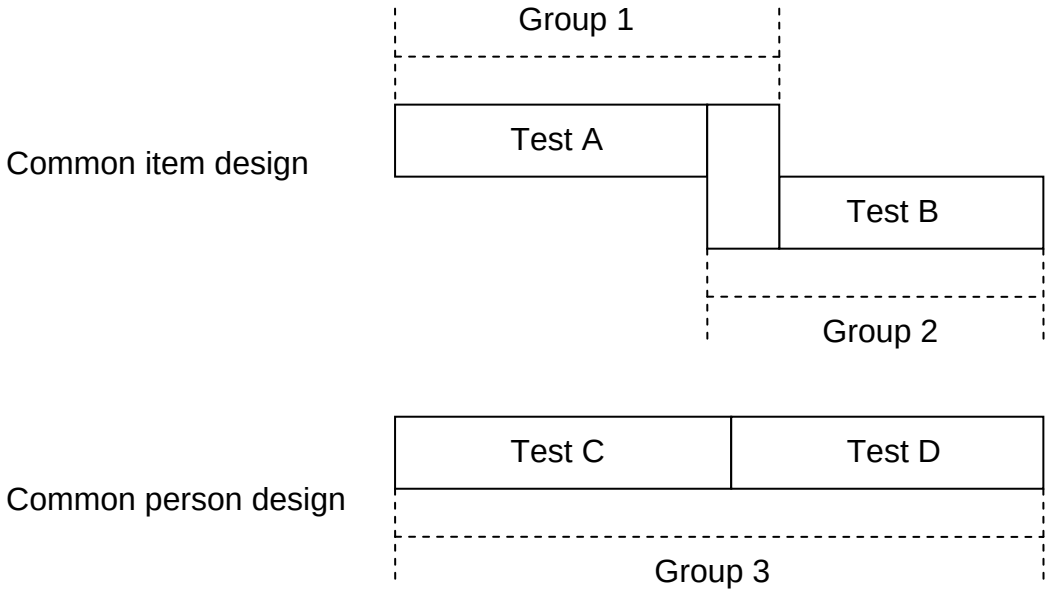

Figure 2 Common item and common person designs in IRT equating.

Effective on-demand testing requires a large number of calibrated items. Pre-testing items therefore present huge tasks for awarding organisations. Wheadon *et al* (2009) have proposed an equating design model involving embedding uncalibrated items in live testing to calibrate items (see Figure 3). The uncalibrated items will not count towards candidates' final scores if instant feedback or grading is needed. These new items can be calibrated based on the performance of candidates on these items and their performance on the calibrated items. It is, however, also possible to count these items towards candidates' final scores once they have been calibrated. The newly calibrated items can then be used to form a live test, and the whole process continues. As suggested by Wheadon *et al* (2009), choice of a test equating design requires the various claims of stakeholders to be balanced and evaluated. These researchers recognised that although pre-testing items in live tests provides the highest level of quality assurance, it could be seen as unfair to the candidates and to make testing less transparent to teachers and schools as different candidates may take a different set of uncalibrated items with varying difficulties and under different content topics and required skills.

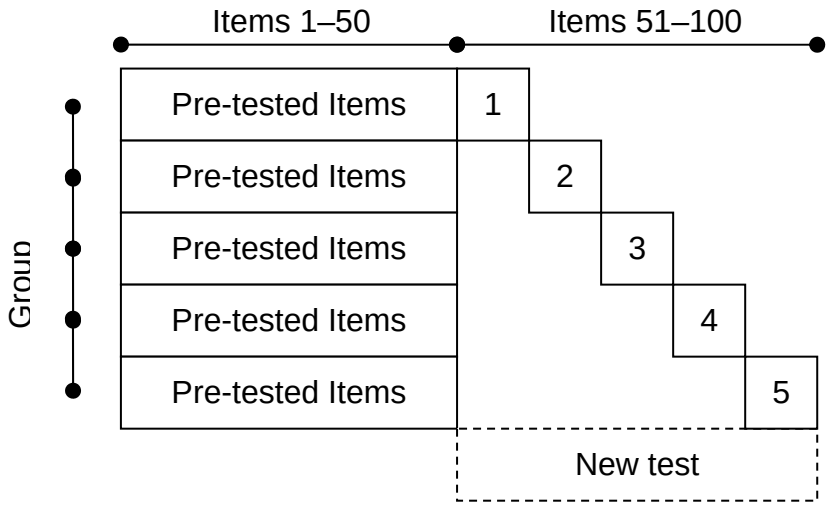

Figure 3 An equating design using live tests to pre-test items (adapted from Wheadon *et al*, 2009).

When a qualification (like most GCSE subjects) is taken by a large number of test centres, it is possible to use some centres to pre-test items using mock tests. These pre-tested items are then used to construct tests for live testing in centres that are not involved in pre-testing. Figure 4 illustrate this pre-testing design (common person equating design in this case). An advantage of this design is that a large number of items can be calibrated and equated relatively quickly. A weakness of this design is that there might be a pre-test effect on the performance of test takers that will affect item parameter estimates and equating relationships. Research has been undertaken to take account of pre-test effect when equating tests using this type of pre-testing design (Maughan *et al*, 2009; Pyle and Shamsan, 2009).

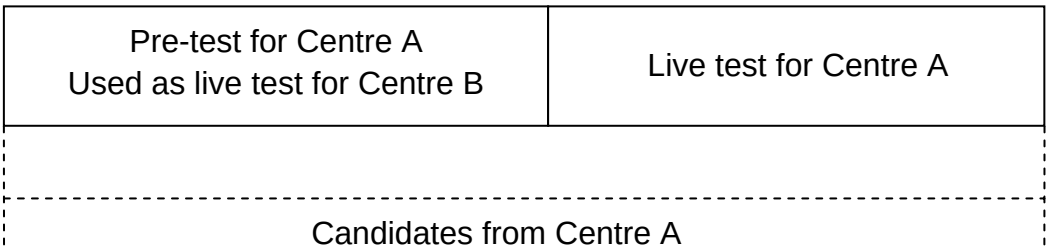

Figure 4 An equating design using mock tests to pre-test items in test centres which are used as live tests for other centres in the future.

Test equating using UIRT models

The difference in test equating between CTT and IRT is that CTT uses directly raw score transformation between test forms while IRT is based on an abstraction of the underlying person trait being measured by the test. The use of common items or common persons for equating using IRT involves similar procedures that are to place item parameters from different test forms on the same common scale. Two methods are commonly used in item calibration and test equating using IRT, the concurrent calibration method and the separate calibration method (see Kim and Cohen, 1998; Bèguin, 2000; Hanson and Bèguin, 2002; Kolen and Brennan, 2004; Chin *et al*, 2006). In the concurrent calibration method, test forms are combined and all items are calibrated onto the same common scale simultaneously based on performances on all items (the common items provide a link between other items).

In the separate calibration method, items in each test form are calibrated separately, and the item measures of the common items or the person ability measures of the common persons are compared between the forms to establish a common scale for the individual calibrations which would involve deriving a mathematical equation to transform item parameter values for non-common items from different forms onto the common scale. As mentioned before, because the two forms are assumed to measure the same underlying latent trait, item parameter values derived from the two separate calibrations should be linearly related to maintain the invariance of the IRFs. Therefore, to place the two scales on a common scale, one simply needs to transform item parameter values from one scale to the other through a linear transformation. Assume that form X is taken as the reference form, form Y is to be equated, the following transformation will convert person and item parameters from scale Y onto scale X:

$$\begin{aligned} \theta_X &= F_\theta(A_1, A_2, ..., A_K, \theta_Y), \\ \eta_{X,k} &= F_\eta(A_1, A_2, ..., A_K, \eta_{Y,k}), \qquad k = 1,2,...,K \end{aligned} \tag{9}$$

where:

θ_Y = ability measure for a test taker on the Y scale

θ_X = ability measure on the X scale after transformation

$\eta_{Y,k}$ = measure of the k^{th} item parameter on the Y scale

$\eta_{X,k}$ = measure of the k^{th} item parameter on the X scale after transformation

K = number of parameters in the IRT model

$A_1, A_2, ..., A_K$ = transformation parameters

F_θ, F_η = transformation functions.

The form of the transformation functions is determined by the IRT model. Values of the transforming parameters can be estimated based on the performances of the test takers on the common items in the case of equating using common items, and a number of methods have been developed for this purpose (Hambleton *et al*, 1991; Kolen and Brennan, 2004; Simon, 2008). The difference in item parameters on the common items between the two groups resulted from the difference in the ability distributions between the groups.

In the case of 2PL and 3PL models for tests containing dichotomous common items, there are two transforming parameters A_1 and A_2, and four of the most widely used methods to estimate them are the mean/sigma, mean/mean, Haebara, and the Stocking and Lord approaches (Kolen and Brennan, 2004; Simon, 2008):

- the mean/sigma method involves the use of item difficulty parameters
- the mean/mean method involves the use of the item discrimination parameters
- the Haebara method involves minimising the cumulative squared difference between the ICCs for each item at different ability locations
- the Stocking and Lord method involves minimising the squared sum of the difference of the test characteristic curves (TCCs) of the common items on the two forms at different ability locations.

Kolen and Brennan (2004) and Simon (2008) discussed the strengths and weaknesses of the different approaches. It was found that the Haebara method produces more accurate and stable results. Once item parameters are placed on the same common scale, IRT true score equating and observed score equating methods can be used to equate raw scores on different forms (Kolen and Brennan, 2004; van der Linden, 2005a). Kolen and Brennan (2004) discussed various procedures for equating tests containing polytomous items using UIRT models like the GRM and the GPCM (also see Song, 2009).

Results from a number of studies have suggested that both concurrent calibration and separate calibration methods produce similar results (cf. Hanson and Bèguin, 2002; Kolen and Brennan, 2004; Chin *et al*, 2006).

Issues with the stability of item parameters and test equating

The accuracy of the comparability of standards of the different forms of the same test over time or between test sessions depends to a certain degree on the accuracy and stability of item parameters, which are affected by a range of factors.

- *Sample characteristics and size*: To obtain accurate estimates for item parameters, a reasonable sample size will be required for item calibration. The more item parameters in a model, the more likely that the data would fit the model better, but a larger sample size is also required. Ideally, the sample used to calibrate the items should be representative of the target population. Kolen and Brennan (2004) discussed sample size requirements for some IRT models.

- *Common items*: Kolen and Brennan (2004) suggested that as a rule of thumb, over one fifth of items should overlap between two test forms to produce reliable equating results. As indicated by Wheadon *et al* (2009), common items should ideally represent a miniature of the entire test (also see Kolen and Brennan, 2004).

- *Violation of model assumptions*: As indicated earlier, when model assumptions are violated, item parameter values can be sample dependent and therefore not stable. Local independence for UIRT models could be easily violated for dichotomous items sharing the same stimulus, and in this case, those items should be grouped and analysed using polytomous IRT models.

- *Model fit*: The application of IRT to analyse test data assumes that the model reflects the functioning of the test correctly. The degree that the model fits the test data can be evaluated using model fit statistics which are normally calculated and exported by the software used for item calibration.

- *Curriculum change and population change*: When test data meets IRT model assumptions, item parameter estimates will be sample independent. When there is a curriculum change, it is likely that the relationships between the various aspects of the curriculum will also have changed. This will result in old items and new items measuring different latent variables and therefore instability in item parameters.

- *Calibration methods*: As discussed previously, different calibration methods produce item or person estimates with slightly different precision. This needs to be taken into account when calibrating items. Consistent calibration methods should be used over the lifetime of the item bank.

- *Equating methods*: Different equating methods are likely to produce slightly different equating results. Whenever possible, the equating error should be evaluated to choose the best equating methods.

- *Item parameter drift*: Other factors such as security breaches can also result in instability of item parameters. Overuse of particular items increases the risk of the items becoming easier and therefore parameter drift. If items are reused, it is necessary to monitor item parameter drift. Item parameter drift has an impact on equating results.

Test equating using MIRT models

In MIRT, model indeterminacy also exists (i.e. IRFs can maintain invariant under some linear transformations of traits and item parameters). If items from two forms of the test are analysed concurrently, both person trait measures and item measures are placed on the same scale. In the case of separate calibration, as indicated by Simon (2008) and Yao and Boughton (2009), the scale of form Y can be linearly transformed onto the scale of the reference form X using procedures similar to those used for UMIT models. Because multiple sets of parameter estimates exist, this transformation will normally involve the use of matrices. Again common items and common persons are the most widely used linking methods. Simon (2008) provided a comprehensive review of MIRT linking and equating approaches. Yau and Boughton (2009) have developed procedures for MIRT linking with both dichotomous and polytomous items. Reckase (2009) presented different methods for MIRT linking and equating.

Computer software for item calibration and test equating

Below is a list of IRT analysis software commonly used to undertake item analysis:

- BILOG/BILOG MG (www.ssicentral.com/irt/example1.html)
- BMIRT (www.ctb.com)
- ConQuest (www.assess.com)
- FACETS (www.winsteps.com/facets.htm)
- MULTILOG (www.ssicentral.com/irt/example1.html)
- OPLM (norman.verhelst@citogroep.nl)
- PARSCALE (www.ssicentral.com/irt/example1.html)
- RUMM, RUMM202, RUMM2030 (www.rummlab.com.au)
- TESTFACT (www.scienceplus.nl/testfact)
- WINSTEPS (www.winsteps.com/winsteps.htm).

Test construction and delivery

Once a calibrated item bank has been developed, designing tests using items in the bank is relatively easy. One of the advantages of using a calibrated item bank to design tests is that the designed tests can meet the requirements of pre-specified statistical and psychometric performances such as required reliability and measurement precision. The tests designed should also ensure that the requirements of important constraints such as the required number of items measuring the intended knowledge and skills and the variety of item types in the tests are met. van der Linden (2005b) has provided a comprehensive introduction to designing tests with IRT. Effective test design will be an important requirement for on-demand testing to be effective.

General considerations in test design using an item bank

When designing a test, the following aspects should be considered for the test to meet the assessment objectives:

- *Content balance*: This is to ensure that items in the test cover the required content topic areas so that the intended knowledge and skills are assessed.
- *Test length*: Appropriate length of the test must be kept to ensure that the test will produce reliable results.
- *Test difficulty level*: Targeting items at the ability level of the intended candidates will produce more reliable results. In the case of using IRT-based item bank to construct test, the test information function (TIF) which can be generated dynamically when a test is being designed can be used to suit this purpose (see discussion below).

To facilitate quick test design for on-demand testing, it is needed to have a test design tool that works with the item bank database (i.e. the test design component of the item bank). This tool should be able to perform the following functions:

- *Item searching criteria*: This tool should be able to accept a range of search criteria (such as topic area, skills to assess, and other item attributes) that the test designer wishes to perform.
- *Item searching function*: The tool should be able to identify items that meet the search criteria.
- *Display of searched items*: The tool should be able to display the searched items for the test designer to inspect.

- *Display of test information function and other properties*: The tool needs to be able to display the test information function and other properties of the test being designed.

- *Test preview and revision*: The tool should allow the designer to preview the test under construction and to revise it if necessary.

Parallel forms and equivalent forms

As indicated earlier, an important application of IRT in the analysis of items and tests is the use of test information function. In general, when two tests have similar test information functions in addition to other constraints like content and skill coverage, the two tests can be viewed as parallel forms (Boekkooi-Timminga, 1990, Hambleton and Jones, 1993; van der Linden and Adema, 1997; Luecht, 1998; van der Linden, 2005b; Lin, 2008; Sun *et al*, 2008; Reckase, 2009).

Automatic test design

With a calibrated item bank, it is possible to design tests according to specified requirements automatically. Automatic test design generally involves the following procedures (see van der Linden and Boekkooi-Timminga, 1989; van der Linden and Adema, 1997; Luecht, 1998; Lin, 2008; Cor *et al*, 2009; Finkelman *et al*, 2009):

- specifying the requirements for the test to be designed

- developing the algorithms to meet the requirements that frequently involves the development of an objective function under pre-specified constraints that is to be maximised or minimised; for example, when the ability distribution of the target population is known, a test can be designed to produce maximum information for the population in addition to other constraints like non-statistical characteristics (such as test length, content topics, item format etc.)

- implementing the algorithms using a computing programming language.

Test delivery

Once a test has been designed, it can be delivered to centres through the test delivery system for administration. The test delivery system will be able to display questions to candidates and record their responses. Depending on the nature of the test being delivered, the system may also be able to mark candidates' responses and estimate their abilities. Candidates' responses or scores will need to be transferred back to the item bank for making and/or analysis.

Pure adaptive testing

The basic principle of computer adaptive testing is that the items to be administered to an examinee during testing are selected based on the responses which were given

to previous administered items using certain pre-specified rules (Weiss and Schleisman, 1999). One of the main advantages of adaptive testing is that different examinees take a different set of questions which provide maximum measurement precision, yet they still can be compared in terms of the latent trait being measured.

Weiss and Schleisman (1999) outlined the basic requirement for adaptive testing:

- a pre-calibrated item bank containing sufficient number of items exists
- a procedure must be developed for selecting the first item to administer to an examinee
- a procedure must be developed for selecting items during the testing process
- a procedure must be developed to terminate the testing process.

One of the widely used procedures for selecting next items is the use of maximum information in which the next item is selected to make the test information have maximum value. Eggen (2007) has provided a detailed description of the psychometric aspects and practical considerations of conducting CAT. Both UIRT and MIRT models involving both dichotomous and polytomous items have been used to conduct computer adaptive testing (see, for example, Masters and Evens, 1986; Eggen, 2007; Scalise and Wilson, 2007; Reckase, 2009)

Weiss and Schleisman (1999) also discussed issues with adaptive testing. These include context effects in which the prior items in a test influence the answer to succeeding items, unbalanced content in which there is an emphasis on particular content area for different examinees. The choice of the starting item requires attention as examinees having different abilities should be provided items with different difficulties. In CAT, examinees generally cannot review items already answered and therefore cannot change answers, which could affect the performance or ability estimate. It is important to consider these issues when developing procedures for administering the first item and item selection during the testing process.

It should also be noted that candidate's response to a question during testing must be accurately marked so that his/her ability can be estimated dynamically, which is required for selecting the next question. This requirement of instant marking implies that items for adaptive testing must be marked by the adaptive test delivery system. In view of the current technology available to the UK awarding organisations and practice, adaptive testing would only be possible for examinations consisting of objective questions such as multiple choice questions, multiple selection questions, matching questions, and short free response questions that can be marked by computers.

Test scoring and component level grading

In on-demand testing using IRT scales, raw test scores cannot be compared directly, but the IRT ability measures can. Once grade boundaries have been made on the IRT scale, they will be applicable to test forms designed using calibrated items for all test sessions.

Scoring unidimensional tests

In the case that test forms measure a single latent trait, IRT-based scores or trait measures can be used to represent the overall performance of the candidates on the test. These are equivalent to total test scores in conventional testing. Candidates' ability measures can be either estimated by the test delivery system (for example, when items in the test can be marked by the delivery system instantly) or by the analysis system after responses have been marked by human markers. A frequently used measurement unit in IRT is logits (see Wright and Stone, 1979, for a definition).

Scoring multidimensional tests – component composite trait

When a test measures multidimensional latent traits, to facilitate interpretation of test results and grading candidates, the overall trait $\overline{\theta}_c$ on the test can be calculated as a weighted mean of the measures on individual trait dimensions:

$$\overline{\theta}_c = \sum_{i=1}^{N_t} w_i \theta_i \qquad (10)$$

where θ_i is the trait estimate on dimension i and w_i is the assigned weight, and N_t is the number of trait dimensions of the test. Equation (10) is particularly useful when items measuring the same trait in the test can be grouped (e.g. items measuring the same content topics or the same knowledge or skills) and analysed using between-item multidimensionality MIRT models. Weights should remain constant over time (between test sessions) once a decision on weighting has been made.

Standard error of measurement

IRT software normally exports standard errors for trait measures. In the case of UIRT analysis, the measurement error in the overall trait measure for a candidate is the software exported standard error of measurement. In the case of MIRT, the standard errors associated with dimensional trait measures can be used to calculate the measurement error for the overall trait estimate.

Setting initial trait grade boundaries and grading

When the first time the test is used for on-demand testing, grade boundaries must be set. That is the relationship between grades (A*, A, B, C …) and the corresponding latent trait measures must be established. Once they have been set, they need to remain constant over time because the items are aligned on the same measurement scale. This is different from conventional testing in which grade boundaries (cut scores) have to be set for each exam series.

Once a test or its various forms have been taken by a sufficient number of test takers, a relationship between their IRT-based grades and prior achievement could be established. This relationship could be compared with that established before for conventional testing for the same test.

Issues with comparability at qualification level

Establishing score metric and forming qualification composite

If all components/units in a qualification use IRT-based scoring metric, then the overall qualification level trait θ_Q can be defined as the weighted mean of component traits:

$$\theta_Q = \sum_{i=1}^{N_Q} w_{c,i}\bar{\theta}_{c,i} \qquad (11)$$

where $\bar{\theta}_{c,i}$ is the trait estimate on component i calculated using Equation (10), $w_{c,i}$ is the weight assigned to component i and N_Q is the number of components in the qualification. Equation (11) is equivalent to the overall qualification score in conventional testing. Component weights should remain constant over time (test sessions) once a decision on weighting has been made.

If some of the components in the qualification are assessed using conventional testing methods with the raw score metric, it is necessary to convert the IRT scales for components assessed using on-demand testing and the raw score scales for components assessed using conventional testing to a common scale like the uniform mark scale (UMS) used in unitised GCSEs and GCEs.

Qualification grading and qualification level comparability

Whether it is the IRT-based scoring metric or UMS-like scoring metric, qualification level grading should be similar to that used for unitised GCSEs or GCEs. Further, if all components in the qualification are assessed using on-demand testing, then comparability in standards at qualification level over time should be automatically maintained. This has to be recognised. It is also suggested that any qualifications that have components which are assessed using on-demand testing should try to maintain standards at component level. This is important for all components to move to on-demand testing in the future.

Further discussion

Choice of IRT models

The choice of an IRT model to a large degree depends on the nature of the assessment. While tests measuring a single latent trait can be analysed using UIRT models, tests that are of multidimensional nature will require MIRT models in order to capture all their features. The interrelationships between questions in the test will also impact on the choice of IRT models. For example, tests consisting of between-item multidimensionality can be modelled using MIRT models differently from test consisting of within-item multidimensionality. Choice of model will have important impact on item calibration and test equating. Models containing more item parameters (such as MIRT models) can describe a test more accurately or fit the test data better, but they will also require large samples to pre-test items in order to obtain stable parameter estimates. A balance between model complexity and practical constraints has to be reached. The degree to which test data meets the assumptions of the selected model must be evaluated.

Pre-testing items and equating design

It has been shown that pre-testing items and equating tests are crucial for establishing a large calibrated item bank and on-demand testing in general. The choice of a pre-equating design needs to take into account a range of factors, including security. The two pre-equating designs described in this report need to be investigated further to confirm their suitability for actual implementation for specific assessments. Other equating designs can also be explored (see Bèguin, 2000; Kolen and Brennan, 2004; Wheadon *et al*, 2009). Care needs to be taken to ensure that there is no security breach when pre-testing items.

Form of on-demand testing

The format of on-demand testing such as the adoption of multiple testing windows or pure adaptive testing depends to a certain extent on the size of the calibrated item

bank (or the speed to pre-test items) and the format of the questions in the test. As indicated earlier, objective questions are relatively easy to write, and tests consisting of objective questions can be marked by computers. These questions are particularly suitable for computer adaptive testing. For tests that are composed of questions requiring human marking, smaller infrequent test windows would be necessary.

Cooperation between awarding organisations

The development of an on-demand testing system that is based on the use of IRT is technically demanding. It requires expertise in the areas of item writing, test development, psychometric and statistical analysis and modelling, computing, and other relevant skills. As identified by Wheadon *et al* (2009), resources in these areas in UK awarding organisations are relatively scarce and further enhancement is needed. A possible approach to implementing on-demand testing in general qualifications is for awarding organisations to work together, drawing expertise from different areas to develop a shared understanding of the pros and cons of approaches to maintaining standards in an on-demand environment that can be used by individual awarding organisations.

References

Ackerman, T. (1992) A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of educational measurement*, **29**: 67-91.

Ackerman, T. (1994) Creating a test information profile for a two-dimensional latent space. *Applied psychological measurement*, **18**: 257-276.

Ackerman, T. (1996) Graphical representation of multidimensional item response theory analyses. *Applied psychological measurement*, 20: 311-330.

Adams, R. and Wilson, M. (1996) Formulating the Rasch model as a mixed coefficients multinomial logit. In *Objective measurement III: Theory into practice* (eds G. Engelhard and M. Wilson). Ablex, Norwood, NJ, USA.

Adams, R., Wilson, M. and Wang, W. (1997) The multidimensional random coefficients multinomial logit model. *Applied psychological measurement*, **21**: 1-23.

Andrich, D. (1978) A binomial latent trait model for the study of Likert-style attitude questionnaires. *British journal of mathematical and statistical psychology*, **31**: 84-98.

Baird, J. (2007) Alternative conceptions of comparability. In *Techniques for monitoring the comparability of examination standards* (eds P. Newton, J. Baird, H. Goldstein, H. Patrick and P. Tymms), 124-156. QCA, London, England.

Barnard, J. (1999) Item analysis in test construction. In *Advances in measurement in educational research and assessment* (eds G. Masters and J. Keeves), 195-206. Pergamon, New York, USA.

Bèguin, A. (2000) Robustness of equating high-stakes tests. PhD thesis, University of Twente, the Netherlands.

Bock, R. (1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Phychometrika*, **37**: 29-51.

Boekkooi-Timminga, E. (1990) The construction of parallel tests from IRT-based item banks. *Journal of educational and behavioral statistics*, **15**: 129-145.

Boyle, A. (2010) *Regulatory research into on-demand testing*. Ofqual, Coventry.

Chin, T., Kim, W. and Nearing, M. (2006) Five statistical factors that influence IRT vertical scaling. Paper presented in the Annual Meeting of National Council on Measurement in Education at San Francisco, USA.

Cor, K., Alves, C. and Gierl, M. (2009) Three applications of automated test assembly within a user-friendly modeling environment. *Practical assessment, research and evaluation,* **14***, No 14.*

Cresswell, M. (2000) The role of public examinations in defining and monitoring standards. In *Educational standards* (eds H. Goldstein and A. Heath), Oxford University Press Inc, New York, USA.

DeMars, C. (2006) Applications of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of educational measurement*, **43**: 145-168*.*

Eggen, T. (2007) Choices in CAT models in the context of educational testing. In *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing* (ed J. Weiss). www.psych.umn.edu/psylabs/CATCentral/.

Embretson, S. and Reise, S. (2000) *Item response theory for psychologists*. Lawrence Erlbaum Associates, New Jersey, USA.

Embretson, S. (1996) The new rules of measurement. *Psychological assessment*, **8**: 341-349.

Embretson, S. (1997) Multicomponent response models. In *Handbook of modern item response theory* (eds W. van der Linden and R. Hambleton) Springer-Verlag, New York, USA.

Finkelman, M., Kim, W. and Roussos, A. (2009) Automated test assembly for cognitive diagnosis models using a genetic algorithm. *Journal of educational measurement*, 46: 273-292.

Hambleton, R. and Jones, R. (1993) An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational measurement: Issues and practice*, **12**: 38-47.

Hambleton, R., Swaminathan, H. and Rogers, H. (1991) *Fundamentals of item response theory*. Sage Publications, London, England.

Hanson, B. and Bèguin, A. (2002) Obtaining a common scale for item response theory item parameters using separate versus cocurrent estimation in common-item equating design. *Applied psychological measurement*, **26**: 3-24.

Hartig, J. and Höhler, J. (2008) Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Journal of psychology*, **216**: 89-101.

Hol, A., Vorst, H. and Mellenbergh, G. (2007) Computerized adaptive testing for polytomous motivation items: administration model effects and a comparison with short forms. *Applied psychological measurement*, **31**: 412-429.

Kim, S. and Cohen, A. (1998) A comparison of linking and concurrent calibration under item response theory. *Applied psychological measurement*, **22**: 131-143.

Kolen, M. (1999) Equating of tests. In *Advances in measurement in educational research and assessment* (eds G. Masters and J. Keeves), 164-175, Elsevier Science, the Netherlands.

Kolen, M. and Brennan, M. (2004) *Test equating, scaling, and linking: methods and practices*, Springer, Berlin.

Lin, C. (2008) Comparisons between classical test theory and item response theory. In *Automated assembly of parallel test forms, Journal of technology, learning, and assessment*, **6**, Number 8, April.

Linacre, J. (2008) *A user's guide to WINSTEPS/MINISTEPS Rasch-model computer programs*. Chicago, USA.

Livingston, S. (2004) Equating test scores (without IRT). *Educational testing service*, New Jersey, USA.

Lord, F. (1980) *Applications of item response theory to practical testing problems*. New Jersey: Lawrence Erlbaum, New Jersey, USA.

Luecht, R. (1998) Computer-assisted test assembly using optimization heuristics. *Applied psychological measurement*, **22**: 224-236.

Masters, G. and Keeves, J. (1999) *Advances in measurement in educational research and assessment*. Elsevier Science, the Netherlands.

Masters, G. and Evans, J. (1986) Banking non-dichotomously scored items. *Applied psychological measurement*, **10**: 355-367.

Masters, G. (1982) A Rasch model for partial credit scoring. *Psychometrika*, **47**: 149-174.

Masters, G. (1999) Partial credit model. In *Advances in measurement in educational research and assessment* (eds G. Masters and J. Keeves), 98-109. Elsevier Science, the Netherlands.

Maughan, S., Styles, B., Lin, Y. and Kirkup, C. (2009) *Partial estimates of reliability: parallel form reliability in the key stage 2 science tests*. Ofqual: Coventry.

Muraki, E. (1992) A generalized partial credit model: Application of an EM algorithm. *Applied psychological measurement*, **16**: 159-176.

Muraki, E. (1997) A generalized partial credit model. In *Handbook of modern item response theory* (eds W. van der Linden and R. Hambleton), 153-164. Springer-Verlag, New York, USA.

Petersen, N., Kolen, M. and Hoover, H. (1989) Scaling, norming, and equating. In *Educational measurement* (ed. R. Linn), 221-262. Macmillan, New York, USA.

Pyle, K. and Shamsan, Y. (2009) Investigation of the factors affecting the pre-test effect in national curriculum science assessment development in England. Paper presented at the 2009 AEA-Europe Annual Conference, Malta.

Rasch, G. (1960) *Probabilitistic models for some intelligence and attainment tests*. Denmark Paedagogiske Institute, Copenhagen, Denmark.

Reckase, M. and Martineau, J. (2004) The vertical scaling of science achievement tests. Paper commissioned by the Committee on Test Design for K-12 Science Achievement Center for Education National Research Council www7.nationalacademies.org/bota/Vertical%20Scaling.pdf

Reckase, M. (1985) The difficulty of test items that measure more than one ability. *Applied psychological measurement,* **9**: 401-412.

Reckase, M. (1997) A linear logistic multidimensional model for dichotomous item response data. In *Handbook of modern item response theory* (eds W. van der Linden and R. Hambleton), 271-286. Springer, New York, USA.

Reckase, M. (2009) *Multidimensional item response theory*. Springer-Verlag, New York, USA.

Reckase, M., Ackerman, T. and Carlson, J. (2005) Building a unidimensional test using multidimensional items. *Journal of educational measurement*, **25**: 193-203.

Robinson, C. (2007) Awarding examination grades: current progresses and their evolution. In *Techniques for monitoring the comparability of examination standards* (eds P. Newton, J. Baird, H. Goldstein, H. Patrick and P. Tymms), 97-123. QCA, London, England.

Rudner, L. (1998) Item banking. *Practical assessment, research and evaluation*, **6**: 4. Available online: http://pareonline.net.

Samejima, F. (1969) Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, **34**: 100-114.

Scalise, K. and Wilson, M. (2007) Bundle models for computerized adaptive testing in e-learning assessment. In D. J. Weiss (ed) *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved 08/03/10 from www.psych.umn.edu/psylabs/CATCentral/

Sijtsma, K. and Junker, B. (2006) Item response theory: past performance, present developments, and future expectations. *Behaviormetrika*, **33**: 75-102.

Simon, M. (2008) Comparison of concurrent and separate multidimensional IRT linking of item parameters. PhD Thesis, University of Minnesota, USA.

Song, T. (2009) Investigating different item response models in equating the examination for the certificate of proficiency in English (ECPE) University of Michigan. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, **7**: 85-98.

Sun, K., Chen, Y., Tsai, S. and Cheng, C. (2008) Creating IRT-based parallel test forms using the genetic algorithm method. *Applied measurement in education*, **21**: 141-161.

Thissen, D., Steinberg, L. and Mooney, J. (1989) Trace lines for testlets: a use of multiple-category response models. *Journal of educational measurement*, **26**: 247-260.

Tshering, G. (2006) IRT in item banking, study of DIF Items and test construction. Mater Thesis, the University of Twente, the Netherlands.

Umar, J. (1999) Item banking. In *Advances in measurement in educational research and assessment* (eds G. Masters and J. Keeves), 207-219. Elsevier Science, the Netherlands.

van der Linden, W. and Adema, J.J. (1997) Simultaneous assembly of multiple test forms. *Journal of educational measurement*, **35**: 185-198.

van der Linden, W. and Boekkooi-Timminga, E. (1989) A maximin model for IRT-based test design with practical constraints. *Psychometrika*, **54**: 237-247.

van der Linden, W. (1999) Computerized educational testing. In *Advances in measurement in educational research and assessment* (eds G. Masters and J. Keeves), 138-150. Elsevier Science, the Netherlands.

van der Linden, W. (2005a) *Evaluating equating error in observed-score equating*. Law School Admission Council, Pennsylvania, USA.

van der Linden, W. (2005b) *Linear models for optimal test design*. Springer, New York.

Wang, W. (1995) Implementation and application of multidimensional random coefficients multinomial logit. Unpublished doctoral dissertation. University of California, Berkeley.

Ward, A. and Murray-Ward, M. (1994) Guidelines for the development of item banks. An NCME instructional module. *Educational Measurement: Issues and Practice* 13: 34-39.

Weiss, D. and Schleisman, J. (1999) Adaptive testing. In *Advances in measurement in educational research and assessment* (eds G. Masters and J. Keeves), 129-137. Elsevier Science, the Netherlands.

Wheadon, C. and He, Q. (2006) An investigation of the response time for maths items in a computer adaptive test. *Proceedings of the 10th International Computer Assisted Assessment Conference*, Loughborough University, UK, 455-466.

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

Wright, B.D. and Masters, G (1982) *Rating scale analysis. Rasch measurement.* Chicago, IL: MESA Press, USA.

Wright, B.D. and Stone, M.H. (1979) *Best best design. Rasch measurement.* Chicago, IL: MESA Press, USA.

Wu, M. and Adams, R. (2006) Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics education research journal*, **18**: 93-113.

Wu, M., Adams, R. J., and Wilson, M. R. (2007) ConQuest: Multi-aspect test software (computer software). Melbourne: Australian Council for Educational Research.

Yao, L. and Boughton, K. (2009) Multidimensional linking for tests with mixed item types. *Journal of educational measurement*, **46**: 177-197

Yao, L. and Schwarz, R. (2006) A multidimensional partial credit model with associated item and test statistics: an application to mixed-format tests. *Applied psychological measurement,* **30**: 469-492.

Yen, W. (1993) Scaling performance assessments: strategies for managing local item dependence. *Journal of educational measurement*, **30**: 187-213.

Yu, C. and Popp, S. (2005) Test equating by common items and common subjects: concepts and applications. *Practical assessment research and evaluation*, **10**: 4. http://pareonline.net/getvn.asp?v=10&n=4.

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

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

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

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

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346

www.ofqual.gov.uk