MAINTAINING STANDARDS IN UK HIGH-STAKES EXAMINATIONS USING ITEM RESPONSE THEORY (IRT) METHODS OF TEST EQUATING

Christopher Wheadon

ABSTRACT

The dual processes of setting and maintaining of standards in high-stakes public examinations in the UK have long used the same methodology: a combination of judgement and statistics. While judgement is necessary to set standards, if it is accepted that standards are a social construct, it has been found to be a blunt instrument with built-in biases. As such it is arguably not suitable for maintaining standards once they have been set. Standard statistical cohort-referencing and predictive models of maintaining standards cannot be used in isolation because they are subjective models that do not allow for improvement over time. This paper considers whether Item Response Theory (IRT) methods of test equating are suitable for maintaining standards over time in UK high-stakes examinations. It concludes that these methods are readily applicable for assessments that use short response test formats but that a full research programme is required to investigate whether IRT methods are suitable with longer response test formats.

INTRODUCTION

At the end of the GCE Accounting awarding meeting in July 2007, the Chair of Examiners for Accounting reported that while the committee had followed the statistical recommendations on the standards that had been put to them to consider, they

> "would have welcomed the opportunity to revise standards in the candidates' favour".
> Chair of Examiners Report, GCE Accounting, June 2007

Unlike the GCE English committee in 1990 who decided to cut the pass rate at Grade A from 5.7% to 0.7% in one fell swoop (Baird, 2007) this committee were not given that opportunity. Their task was twofold. They were required to use their judgement to examine specific candidates' performance on the test content to determine the extent to which they may have been affected by any changes in the difficulty of an examination from the previous year; but in making any compensation for that perceived change, they were to take into account the statistical information which informed them about various characteristics of the cohort taking that examination. This model of setting and maintaining standards has been dubbed rather derogatorily, but perhaps aptly, the contest model (Newton, 2005), as it requires participants to bring together sources of evidence that occasionally conflict. This process, however, is the essence of any standard setting procedure: it requires participants to bring to bear information about both test content and test takers using a combination of judgement and statistics. What is unique to the model used in the UK is that a single methodology has developed for setting standards and maintaining standards. In this US there are two very different traditions for standard setting (see Cizek & Bunch, 2007 for an review of methodologies) and maintaining standards, which is largely done using test equating (see Kolen & Brennan, 2004 for an

overview of methodologies). The Accounting committee, having set the initial height of the bar in 2002, have been given the opportunity to spend five years wrangling over whether it was indeed set at the right height, when that height may not even exist in any objective sense! This paper argues that once standards have been set, IRT methods of test equating could be used to maintain them.

MAINTAINING STANDARDS OVER TIME

If it is accepted that standards do not exist in any objective sense as they represent the expectations of society and reflect the uses to which those standards may be put (Baird, Cresswell, & Newton, 2000) then no purely statistical technique could ever be substituted for an initial standard setting process. Once a standard has been set, however, the task of maintaining that standard over short periods of time should be fairly straightforward. In practice the weight of evidence currently rests firmly with statistical methods (Stringer, 2008a). These methods assume that last year's outcome provides a statistical starting point for standard setting this year, *ceteris Paribas* (Cresswell, 1992). There are two fundamental problems with a purely statistical approach, however. Firstly, no examination work is ever inspected. The system could go ahead blithely awarding grades based on predictions without candidates ever turning up for examinations. Secondly, the choice of variables that compose any statistical method of maintaining pass rates is a subjective decision. Under the present system, for example, no statistical controls are made for different rates of progress by gender.

As a result over five hundred committees of eight senior examiners (on average) are convened to make judgements of students' performance on GCSE and GCE examinations in England every year (Baird & Dhillon, 2005). Work is inspected, and the statistically predicted boundaries ratified in almost all cases (Stringer, 2008a). Occasionally the judgment suggests that alternative statistical indictors be sought, but this useful purpose has to be set against the evidence that giving the candidates the benefit of the doubt has, even within the tight statistical constraints that are imposed, introduced an inflationary bias into the system (Stringer, 2008b). There are ways of minimising the threat from this bias which ensure that the statistics are ratified in all but the most compelling cases, but the subjective nature of the predictive models and the uncertainty associated with the predictors that are used remain of concern (Pinot de Moira, 2008).

There is an alternative: in the US the setting and maintaining of standards for high-stakes university admissions tests (in particular the SAT®, the PSAT/NMSQT®, and the Advanced Placement Program® (AP®)) have long since developed separate methodologies. Standard setting is undertaken using a combination of judgement and statistics usually through Angoff or IRT bookmarking procedures (Cizek & Bunch, 2007) whereas the maintenance of standards is undertaken using IRT test equating (Kolen & Brennan, 2004). IRT test equating has the advantage over the statistical methods of maintaining standards used in the UK as it considers the actual performance of candidates on a subset of the total items they have taken. It has been used to set standards between tiers at GCSE (Wheadon & Beguin, 2007) but has not yet been explored for longitudinal equating. The equating done for GCSE acts as complementary evidence: the re-engineering of the assessment system that has begun in order to realise the benefits inherent in e-assessment could take a more radical approach: poring over scripts after assessments have been marked has no place in an on-demand world.

IRT MODELS

Item Response Theory would seem perfectly suited to the task of maintaining standards as, unlike Classical Test Theory, which is test oriented, it models performance at an item level in order to separate the characteristics of the population taking that test from the characteristics of the items in that test (Lord, 1980). As such it frees the measurement of ability from dependence on a fixed set of items, and frees the measurement of item difficulty from dependence on a fixed population. Given the right conditions, therefore, and the acceptance of some strong statistical assumptions (see later) that do not hold precisely in real testing situations, IRT can be used to compensate for the variation in candidate performance that is due to the variation in demand of a test (Kolen & Brennan, 2004).

One of the most widely used IRT models is the three-parameter logistic model (Birnbaum, 1968; Lord, 1980). Under this model the probability that persons of ability equal to the ability of person *i* correctly answer item *j* is defined as:

$$p_{ij} = p_{ij}\left(\vartheta_i; a_j, b_j, c_j\right) = c_j + \left(1 - c_j\right)\frac{\exp\left[Da_j\left(\vartheta_i - b_j\right)\right]}{1 + \exp\left[Da_j\left(\vartheta_i - b_j\right)\right]} \qquad 1.0$$

Where ϑ_i is the ability parameter for person *i*, c_j is the pseudo-chance parameter for item b_j is the difficulty parameter for item *j*, and a_j is the discrimination parameter for item *j* . The item characteristic curve (ICC) relates the probability of correctly answering an item to candidate ability. There are simplifications to this model: the two-parameter model does not explicitly accommodate examinee guessing while the Rasch model (Rasch, 1960) requires all items to be equally discriminating. Only the Rasch model is an objective model in which raw scores are sufficient statistics.

CHOOSING BETWEEN IRT MODELS

For dichotomous items the Rasch model has some perceived theoretical advantages in that it maintains a monotonic relationship between person/item measures and total test/item scores. Allowing the discrimination of two items to differ means that measurement scales free from person ability cannot be constructed (Wright, 1997). The perceived disadvantage of the lack of a guessing parameter is conceptual as guessing is a characteristic of a person as much as an item, but it can also be addressed by a test construction and validation process which removes items liable to guessing.

Most assessments delivered at GCSE and GCE, however, require polytomous responses: these range from long essay questions in GCE English Literature to well defined multi-part questions in GCSE Mathematics. Within the Rasch family of models the Partial Credit Model (Wright & Masters, 1982) extends the dichotomous model so that partial credit can be given to ordered responses to a single stimulus. Within the IRT family of models Samejima's Graded Response Model (1969) extends the 2-parameter logistic model to polytomous ordered categories while Muraki's Generalised Partial Credit model (1992) allows the slope parameters of the Partial Credit Model to vary. OPLM (Verhelst, Glas, & Verstralen, 1995) imputes rather than estimates discrete discriminations for better fit.

The assumption of all of these polytomous models is that the category responses are ordered: a higher category implies more of the construct being measured. In fact Bode (2004) expresses the kind of polytomous item that would be suitable as one which requires completion of one step before progression to the next. As He (2008) notes, this is not the case with the polytomous items that are typically used in high-stakes examinations in the UK where there may be different ways of achieving a particular score on a polytomous item. Figure 1 gives an example of this from a GCSE Mathematics paper. The notation 'ft' in the mark scheme allows candidates who have got a previous part of the question wrong to answer subsequent parts correctly. Fitting a Rasch model to this test shows that despite the different ways of achieving scores on this item it is behaving well, displaying ordered categories so that a higher score implies more of the latent trait (table 1).

Figure 1: A Typical Multi-Part GCSE Mathematics Question and Associated Mark-Scheme

Question 1
Paul is paid £5 per hour.
Shelley is paid 50% more per hour than Paul. One day Shelley earned £48.75
How long did she work during that day? Give your answer in hours and minutes.
(4 marks)

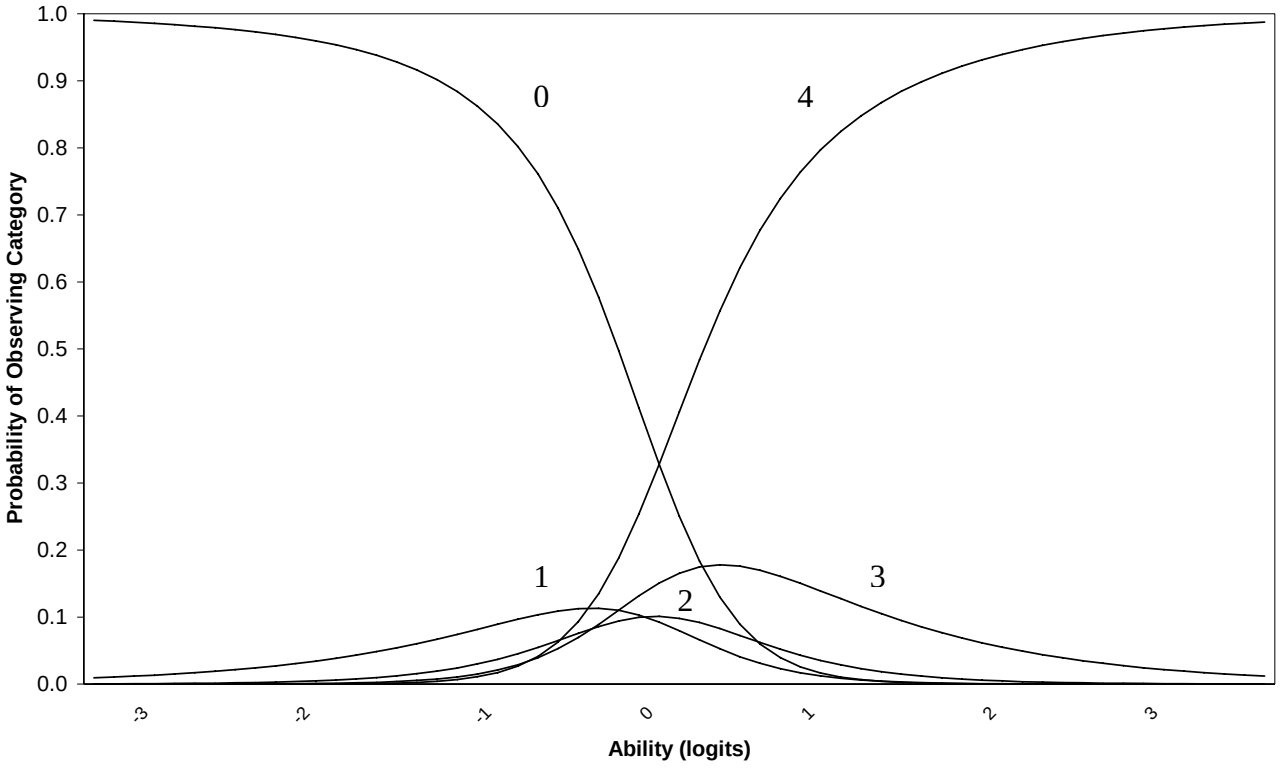
1(b)	Shelley is paid £7.50 an hour	B1	
	£48.75 / *their* £7.50	M1	their £7.50 must be £5 or more Build up must be completely correct method.
	= 6.5	A1 ft	ft their division to 1dp or better
	= 6 hours 30 minutes	B1 ft	ft their decimal time correctly converted to minutes. Allow rounding to nearest minute. Must not be exact number of hours. 6 hours 50 minutes or 6 hours 5 minutes no working SC2

Table 1: Item Statistics for Question 1

Score	Respondents	Respondents (%)	Average Measure	Outfit (MnSq)	Item Threshold	Item x total score correlation
0	8467	70	0.01	1.1	-	-.49
1	932	8	0.67	0.5	1.27	.12
2	585	5	0.81	1.1	-0.09	.13
3	667	6	1.08	1.0	-0.40	.20
4	1438	12	1.22	2.7	-0.78	.37

Figure 2 shows that while the item displays ordered categories, it displays disordered thresholds. The location of the category probability curves along the ability axis show increasing ability with increasing score. The cross-over points (thresholds) between the likelihood of a certain response, however, are disordered. The threshold between a score of 0 and 1 occurs at 1.27 logits while that between 1 and 2 occurs at -0.09 logits. Disordered thresholds imply that these score points correspond to a narrow interval on the latent variable (Linacre, 2004), so the item will discriminate more finely than if it were marked out of 0 or 1. This is a positive quality, and tribute to the detailed preparation of questions and mark schemes by examiners. Whether such detail is sustainable, given the driving force behind e-assessment and the restrictions on response format entailed, however, is debatable. Given current mark schemes the Rasch model displays a key advantage here in that it is robust to the small sample sizes that are apparent across multiple categories. The Graded Response Model, the Generalised Partial Credit model and OPLM fail to converge for this test. For this reason it would seem sensible to use Rasch models of test-equating.

Figure 2: Category Probability Curves for Question 1



EQUATING TESTS

In order to equate the scores from two different tests the parameters from the different tests need to be placed on the same scale. The ϑ scale is often defined as having a mean of 0 and a standard deviation of 1. If the calibration is undertaken on two groups that are not equivalent, their abilities would be scaled to have a mean of 0 and a standard deviation of 1, even though the abilities of the groups are different. If the Rasch model holds, a linear equation can be used to convert Rasch parameter estimates to the same scale. Where A and B are constants in the linear equation and ϑ_{Ji} and ϑ_{Ii} are values of ϑ for individual *i* on Scale *J* and Scale *I*, the item parameter values on the two scales are related as follows:

$$b_{Jj} = Ab_{Ij} + B\,, \qquad (2.0)$$

In terms of groups of items or persons, it follows that:

$$A = \frac{\sigma(b_J)}{\sigma(b_I)} \qquad (2.1)$$

$$B = \mu(b_J) - A\mu(b_I) \qquad (2.2)$$

If a plot of the b-parameters from the common items on each separate scale is shown to be close to the identity line then the constant A can be assumed to be equal to 1, leaving a shift constant B, the mean difference between the b-parameters on the common items, to be applied to bring the items onto a common scale (Linacre, 2006; Wright, 1979). Once these equations have been solved for the common items, the ability scales of the two tests are related as follows:

$$\vartheta_{Ji} = A\vartheta_{Ii} + B \qquad (2.3)$$

The equating therefore relies on the tests sharing items: without common items (or randomly equivalent groups) the scaling constants remain unknown (Kolen & Brennan, 2004).

The problem with using a shift constant with polytomous items is that the b-parameters equated are the threshold values between categories (Masters, 1984) which, with lower sample sizes than if the item was dichotomous, are more liable to misfit and tend not to be stable (Linacre, 2006). One alternative is to use characteristic curve methods of equating which transform the test characteristic curves, given as:

$$\tau(\vartheta_i) = \sum_j p_{ij}(\vartheta_i) \qquad (2.4)$$

These methods use a linear transformation to minimise the difference between the test characteristic curves for common items. This transformation is then applied to bring the tests onto the same scale (Haebara, 1980; Stocking & Lord, 1983). This method takes advantage of the fact that item characteristic curves can be estimated accurately even when item parameters are not estimated precisely, but it is mathematically complex.

Another alternative to separate estimation and transformation of one scale into another is to calibrate all items and persons from both tests together. This is known as concurrent estimation. As the estimation has been done in a single pass the parameters will be expressed on a single scale (Kolen & Brennan, 2004). The method of estimating item parameters is critical. Maximum Likelihood Estimation (MLE) requires sample distributions to be specified in advance when often the distributions are a finding, not an assumption. MLE may result in bias if the examinee groups taking the different test forms differ in ability (Hanson & Beguin, 1999). Concurrent calibration using Conditional Maximum Likelihood Estimation does not result in bias when the ability of the groups to be equated differs, but it lacks computational precision and only allows a few, well defined patterns of missing data (Linacre, 2008). Little is known about the theoretical properties of Joint Maximum Likelihood Estimation (Hanson & Beguin, 1999) but it does not require specification of the population / sample ability distribution and item difficulty distribution.

Finally, He (2008) has developed an extension to the Rasch model for polytomous items. This model, should it prove robust, is mathematically simple to implement and simple to interpret. This does not only have logistical advantages. Models that are simple to interpret are more likely to engage examiners in the whole process of test construction, use and evaluation in a constructive fashion. This process could therefore become a continuous feedback cycle improving the quality of the tests and engaging examiners with standard setting much earlier in the test production process. He's model must, however, be carefully evaluated in a number of different contexts.

Once item parameters have been placed on the same scale, the test scores on one test form can be mapped to scores on the other test form through either true-score equating or cumulative frequency distribution equating. True-score equating should be able to provide more detail than is available from the observed test scores should the model hold. Empirical tests have shown that this is the case: IRT models represent a smoothed score distribution which reduce fluctuations in the results of equating and improve the invariance of equating over populations (Beguin, 2000).

TEST EQUATING DESIGNS

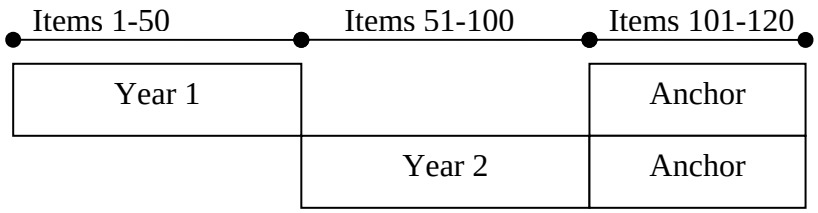
Internal Anchors

It has been shown therefore, that it is relatively straightforward to equate different tests as long as some proportion of the items from the tests to be equated are taken by a sample of the entire cohort. As a rule of thumb, around one-fifth of the items should be taken by several hundred candidates, although specific requirements may require more or less of either (Kolen & Brennan, 2004). There are various designs which achieve this, but they all rely on the premise that the tests to be equated should be built to exactly the same specification and measure the same construct, and where common items are employed, these should ideally represent a miniature of the entire test (Kolen & Brennan, 2004).

Perhaps the most commonly used design involves candidates taking some previously used questions that are not scored (an internal anchor), along with some new questions (figure 3). The anchor test links results on the new items to the standard previously established on the anchor test. This design is known as the non-equivalent anchor-test design (NEAT), and is the approach used by the US high stakes university admissions tests, the SAT®, the

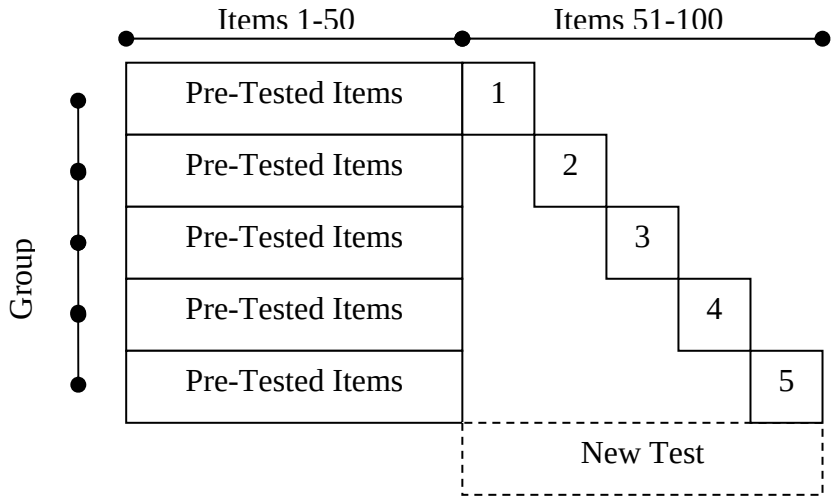
PSAT/NMSQT®, and the AP®. The advantage of this design is that all the questions are taken under live conditions so levels of motivation and preparation do not affect item calibrations. The disadvantages are that items designated for re-use in an anchor cannot be released to the public, any that are leaked compromise the standard being set, and levels of motivation of candidates may be affected by items that do not contribute to their scores.

Figure 3: Non-Equivalent Group Anchor-Test Design (NEAT)



A variation on this design involves candidates pre-testing unscored items during their live tests so the standard on these items is known before they are subsequently used and scored (figure 4). This design, known as the pre-equating non-equivalent groups design, is used for the SweSat, the test used for university entrance in Sweden (Emons, 1998). To minimise the security risk involved in exposing new items before they are used live, the items are split across a number of candidate groups so each new item is only exposed to a proportion of the total cohort. This design combines the advantages of maintaining standards with the enhanced quality control that comes from pre-testing new items.

Figure 4: Pre-Equating Non-Equivalent Groups Design



External anchors

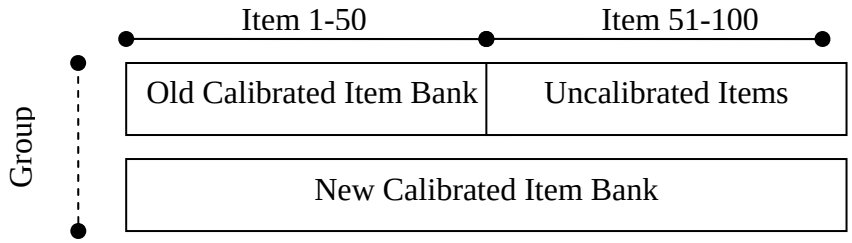
Rather than using an internal anchor, an external anchor could be used based on either of the above designs, and administered before the test. This approach gets over the problem of using items that are not scored in a live test, but it assumes that the response behaviour of candidates is similar in the pre-test to the live test, which is problematic. A variation of this approach is used for the Key Stage Tests in the UK which equates results on a pre-test one year to a pre-test the next year (Hayes, 2008). It is assumed therefore that the candidates' motivation and preparation will be similar from pre-test to pre-test. There is no reason to believe, however, that the test equating function from pre-test to pre-test will be the same as the test equating function from live test to live test; and item security remains a key concern. This design would require a huge amount of time and resources to be spent on pre-testing sessions, and is therefore not operationally suitable to the myriad of qualifications offered at GCSE and GCE.

The design adopted in the Netherlands for public examinations is to use learners who are not participating in the examinations to take some of the old and some of the new items after the administration of the live tests. This is known as the post-equating non-equivalent groups design. In this way the security of all items is preserved as the equating takes place post-hoc, and all items used in the live situation contribute to candidates' scores (Alberts, 2001). With the drive for faster results in the UK and competition between awarding bodies to release results as quickly as possible, this design is untenable.

Item Banking

Item banking can be conceptualised as an IRT test-equating design (figure 5). If Form X1 is an item pool which has been calibrated on a single scale, then an item pool with uncalibrated items can be referred to as Form X2. The test equating procedures described earlier can then be used to transform the item parameters from Form X2 onto the theta scale for Form X1 (Kolen & Brennan, 2004). This design is only really feasible through onscreen delivery as the rate at which the item bank can be built up with paper testing is too slow. An algorithm that delivers a random uncalibrated item from an item bank at specified anchor positions within a test to each candidate would allow a large number of items to be quickly calibrated while minimising their exposure and thus the security risk.

Figure 5: Common-Item Equating to a Calibrated Item Bank



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

Brennan (2004) recommend using an anchor design before moving to a full item-banking design so that potential problems can be teased out before this next step is taken.

Evaluating the Test Equating Designs

For most high-stakes assessments offered by AQA, the NEAT and pre-testing designs seem most suitable. Table 2 summarises the differences between them. NEAT has the advantage on security: items will be only be re-used as part of the anchor where they won't be scored, so there is little incentive to steal them; under the pre-testing design items designated for the new test are exposed so future candidates could gain an advantage through item theft.

Table 2: Advantages of NEAT against the Pre-testing Design

	Same test for all candidates	Item theft cannot affect individual scores	Builds in pre-testing	Allows release of all scored items	Increase to test length	Candidates take unscored questions
NEAT	Yes	Yes	No	Yes	20%	Yes
Pre-testing	No	No	Yes	Yes	<20%	Yes

On public accountability the pre-testing design has the advantage, with a rigorous pre-testing design to assure the quality of the assessments delivered. It may be argued that pre-testing is a luxury the UK assessment system has done without for a long time, but if the pre-testing design offers an economical solution for pre-testing then the quality of assessments delivered could only benefit. It should be noted that the SAT®, the PSAT/NMSQT®, and the AP® are all pre-tested separately from the test-equating procedure: in the US pre-testing is not considered a luxury.

On logistics grounds the NEAT has the advantage that candidates in the same test session sit the same test. This is similar to current methods of test delivery in the UK. E-assessment, however, promises a fresh approach to how tests are delivered and offers the chance to reappraise systems. It would be straightforward to implement the delivery of different forms of the same test through an e-assessment system. In a dual economy of paper and e-assessment, results from the e-assessment forms could be used to drive the paper based standards once take-up of the e-assessment has reached critical numbers. Delivering different tests in the same session should not be an issue for stakeholders. The notion of a single sampling of the curriculum being delivered to an entire cohort has already been eroded through modular GCSEs, for example, which offer three testing sessions a year, and are being retaken by up to 40 *per cent* of candidates.

Both of these designs require tests that are longer than the existing tests: the NEAT would require an anchor that is a miniature replication of the entire test. This anchor must consist of at least 20 *per cent* of the total items in the test. The pre-testing design would not require such a large anchor, as each new section is anchored against the entire live test. In terms of demands on the testing timetable therefore, which are currently quite severe, the pre-testing design has the advantage.

Perhaps the critical factor against which to weigh up these designs, however, is their requirement for candidates to answer questions that do not count towards their total test score. Under the NEAT design, all candidates see the same items in the same order; under the pre-testing design candidates would take different items which may be of different difficulty. The iniquity of one candidate sweating over a particularly difficult question while their neighbour breezes past an easy question is compounded by the fact that those questions aren't scored.

Under the NEAT design the anchor items are not scored to minimise the incentive to steal them. Under the pre-equating design the new items could be scored using IRT algorithms, but this would defeat the purpose of pre-testing them to assure their quality. On balance it would seem the rights of the candidate are better served by the pre-testing design unless the NEAT items are separately pre-tested. It is not impossible to conceive a shift in opinion to the point at which it becomes no longer acceptable to deliver items to candidates without a sound knowledge of the properties of those items. This would outweigh the disadvantage that candidates' motivation could be differentially affected by the test form they are presented with.

There is little to choose between these designs, and the best position, though complex logistically, may be a hybrid. The items to be pre-tested could be delivered in specifically accredited test centres while non-accredited centres could be given a reference test anchor instead. An alternative would be to pre-test questions in one geographical location and use them live in another. Care would have to be taken to minimise any sample bias that limiting the numbers of centres in the pre-testing may introduce. Such complex operations could only be achieved through an architecture specifically designed to manage this process (figure 6).

VIOLATIONS OF THE IRT MODEL

Use of unscored items is not the reason that IRT methods of test construction and equating are not prevalent in the UK while they firmly underpin standards in the US and Australia. What brought experiments in IRT in the UK to a sudden halt in the late 1980s was a high profile paper which expressed the view that the assumptions of IRT would always be violated in practical testing situations in the UK, and that the assessments would have to be watered down to meet these requirements (Goldstein & Wood, 1989).

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

A study of any test will reveal different dimensions. Figure 7, for example, shows a section of the GCSE Mathematics assessment that a Principal Components Analysis of Residuals consistently identifies as testing a separate construct from the rest of the examination. The

wider question is whether this item, purportedly testing knowledge of the number system, is really testing Mathematics at all; but what is in the syllabus must be tested. Bejar (1983), however, provided a key clarification of the requirement for unidimensionality; that it is not necessary for a single latent trait to account for the performance of all the items in a test as long as a coherent scale can be constructed. IRT methods of test equating have elaborated on this premise, finding that where different dimensions have been found to exist, they appear to share the same equating function, as the same linear composite of latent traits underlies the item responses on both tests. The overwhelming consensus is that IRT methods of test-equating are robust to violations of the assumption of unidimensionality within homogenous populations (Harris, 1993). Dimensionality, however, remains an empirical issue to be monitored; and less work has been done on the interactions between population sub-groups and violations of unidimensionality.

Figure 7: A Different Dimension from GCSE Mathematics

Here are five digits: 4 1 6 9 3

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

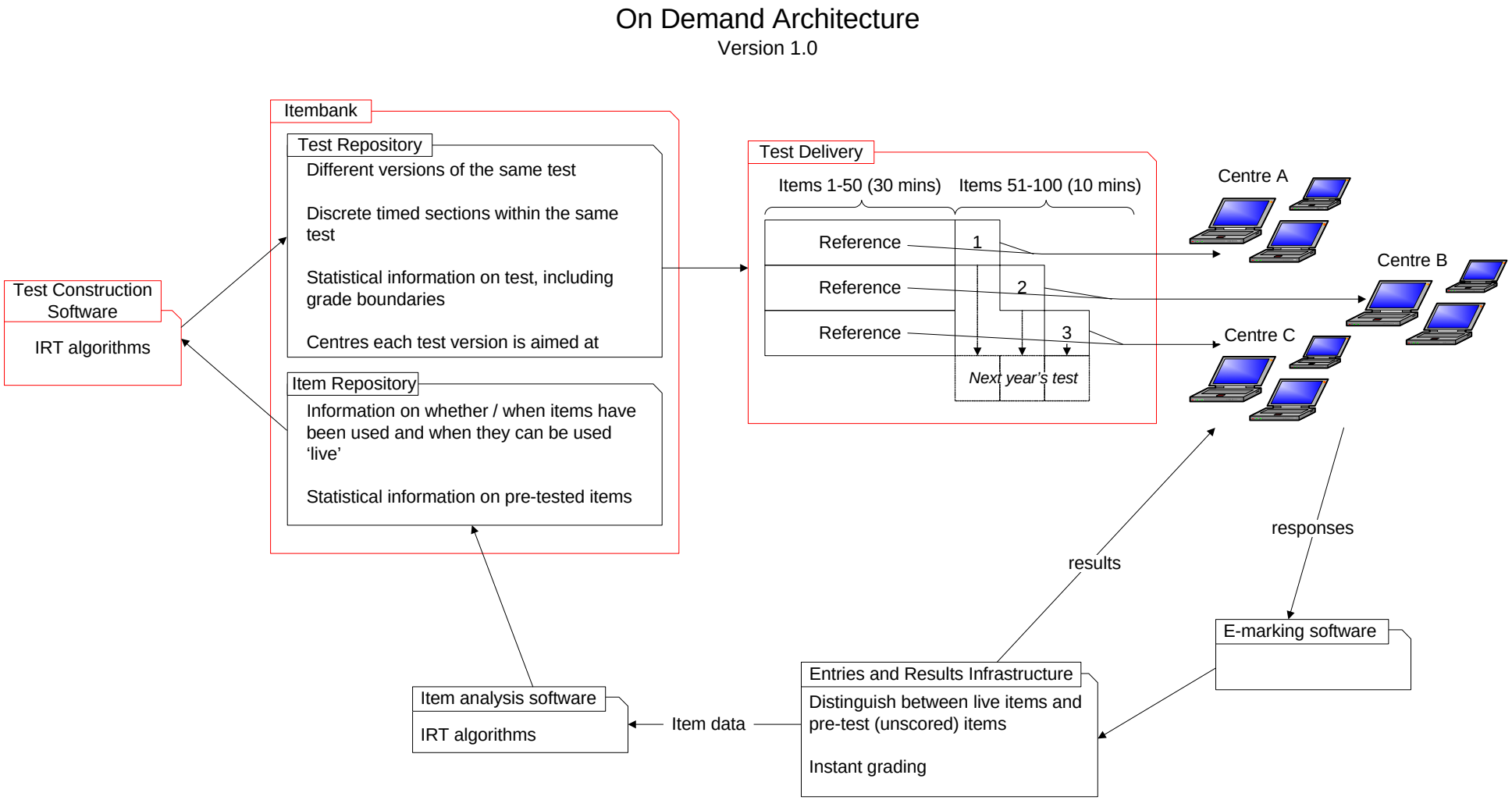
A second assumption that may be violated is that of local independence of item parameters. This requires that candidates' responses to any question are statistically independent when the ability influencing their performance on the whole test is held constant. Figure 8 shows a question that clearly violates this assumption. Answers to the first question will lead to different chances of success on the second, all other factors being equal. This design is typical of UK assessments which tend to group questions around a context such as a passage or a diagram. The solution is simple: responses that are not conditionally independent should be aggregated.

Figure 8: Conditionally Dependent Questions from GCSE Geography

(i) Name an English port (other than London) that has sea links to Europe.
(1 mark)

(ii) For the port in (i) above name:
the main road link leading to/from the port
a port in Europe to which it is linked
(2 marks)

Figure 6: An On-Demand Architecture



Aggregation of responses introduces a third consideration, which has been less examined. At what level of aggregation does the Rasch model cease to be useful? Long responses, for example, are marked on a number of criteria which are then implicitly or explicitly aggregated. Is the Rasch model appropriate in such cases? It is a little studied area, largely because assessments in the US tend to be multiple choice or short-answer. Research in this area is clearly required to determine the scope of any IRT approach.

POPULATION INVARIANCE

Apart from the assumptions of the IRT model, there are five requirements that need to be satisfied in evaluating whether test equating is appropriate:

- Equal Construct: Tests that measure different constructs cannot be equated.
- Equal Reliability: Tests that measure the same construct but that differ in reliability cannot be equated.
- Symmetry: The equating function for equating the scores of Y to those of X should be the inverse of the equating function for equating the scores of X to those of Y.
- Equity: It should not matter to examinees which test they take.
- **Population Invariance: The choice of subpopulation used to estimate the equating function between the scores of tests X and Y should not matter. That is, the equating function used to link the scores of X and Y should be population invariant.**

  (Dorans & Holland, 2000)

Of these, population invariance is one of the most difficult to satisfy. If the equating functions used to link the scores on two tests are not invariant across different subpopulations of candidates, the two tests cannot be equated (Dorans, Jinghua, & Shelby, 2008). This is a complex requirement, for, as Brennan (2008) notes, it is entirely possible for population invariance to be satisfied for each demographic categorisation separately, male / female or white / non-white, but not for the crossed categorisation white males / white females / non-white males / non-white females. Exploration of these categories becomes increasingly problematic as subdivisions reduce the sample sizes.

Peterson (2008) argues that all equatings are first and foremost population dependent, and that the characteristics measured by tests built to exactly the same specifications may differ between different sub-groups. Developing her argument, she states that the equating results will be particularly compromised if the selection variable for constructing the subgroups is related to the construct being measured. This situation would not necessarily be problematic, however, unless the composition of the population is liable to change. Cook and Petersen (1987), for example, found that when relatively parallel forms of a biology test were equated using groups of students at different times of the year, they got different equating results. They concluded the difference was due to an interaction between the recency of their course interacting with test content. Research is required in the UK to identify the key population characteristics that could cause relative item difficulty to change so appropriate equating samples can be selected.

LINKAGE PLANS

If it is assumed that all equating is population dependent to a degree, a linking design that minimises the difference between equatings must be considered. Kolen and Brennan (2004) suggest four rules by which these designs can be evaluated:

Rule 1: Avoid equating strains by minimising the number of links that affect the comparison of scores on forms at successive times.
Rule 2: Use links to the same time of the year as often as possible.
Rule 3: Minimise the number of links connecting each form back to the initial form.
Rule 4: Avoid linking back to the same form too often.

The design must also take into account the practical consideration of retake candidates: for modular GCSEs, for example, the evidence suggests that there will be a high level of retakes within any one year.

If the invariance of item parameters is an issue between testing sessions, then the linkage plan in figure 9 would be an option. This follows rule 2, but violates the other rules. A comparison between the autumn and spring tests in year 4 for example would have to be traced back through eight links which represent a severe strain on the equating. It also violates rule 3, as by year 4 each test form has four links separating it from the original test form. It does mean, however, that items will not be reused within a year, which is an important consideration for retake candidates. Whether the violations of these rules is considered an issue depends on the priorities of comparison. Violation of rule 3 risks changes in the absolute standard over time, while violation of rule 1 risks comparability of test forms within any one year. Arguably, the highest priority for an awarding body should be the absolute maintenance of standards over time so the number of links separating test J from the legacy standard would be an issue. Double links from successive series back to the legacy standard could manage this risk (figure 10), but care would have to be taken not to present retake candidates with questions they have already taken. Current statistical procedures do not have a link back to the legacy standard; this extra link would represent a safeguard against incremental drift.

Figure 9: A Single Link Plan

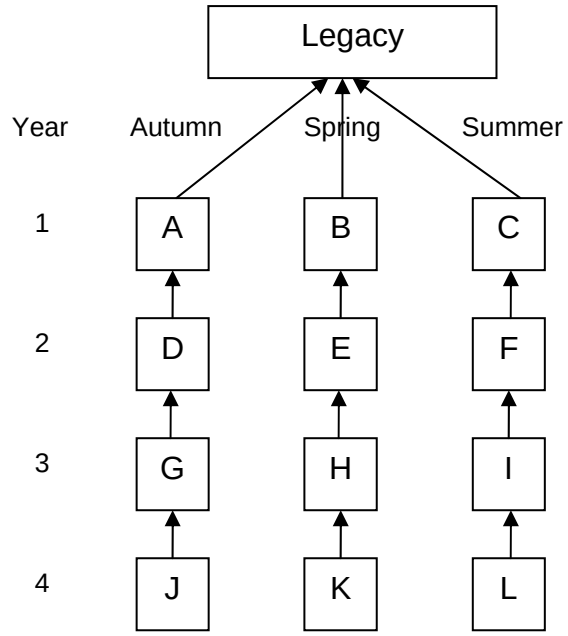
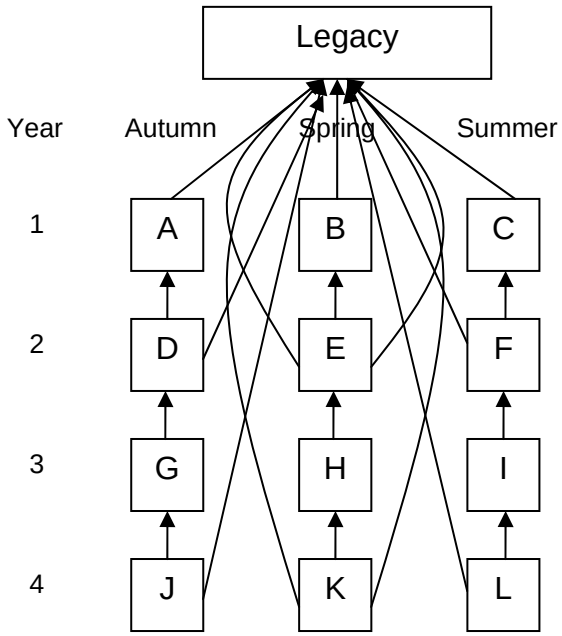


Figure 10: A Double Linking Design



CONTEXT EFFECTS AND THE INVARIANCE OF ITEM PARAMETERS

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

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

The NAEP reading anomaly is clearly a cautionary tale. Under all test equating designs it is now common practice for anchors to be delivered as discrete blocks so that their administration and the time available for their completion can be standardised across different sessions. This approach would be suited to assessment designs that administer blocks of questions around specific stimuli such as a passage of text or a diagram. To accommodate this design e-assessment delivery should therefore be able to facilitate the delivery of discrete blocks within a test, each with its own time limit.

CONCLUSION

It seems clear that IRT designs for maintaining standards over time are feasible, and that developments in e-assessment should be undertaken with the requirements for the chosen design in mind. The range of qualifications over which this approach can be applied will be limited by issues such as dimensionality and the length of response. In the short-term there is most to gain in assessments which can be auto-marked as these can be auto-graded at the same time. There is no reason why these assessments should not be using an IRT test-equating method operationally within the next three years, delivering instant grading, with scope to move to customised testing fuelled by an item-banking, on-demand model soon after.

Much greater care needs to be taken with more complex assessments and the timescales involved will be much longer. Research will be needed to assess issues such as: length of response; nature of mark-schemes; nature of anchors; dimensionality; interactions between question difficulty, time of year and cohort. This research should ensure the safeguards needed are in place before any instant grading is considered. A belt and braces approach will certainly be needed in early stages of this work, with counter-balanced designs validating the IRT approach.

Where an IRT test-equating approach to maintaining standards over time does prove to be robust, however, the result will be a statistical method of maintaining standards that considers the work produced by candidates and allows for changes in the standard of that work resulting from changes in motivation, teaching and preparation of candidates. Integrating the standard setting approach with IRT methods of test-construction and feedback should produce a virtuous circle of development, appraisal and delivery that ensures the examining community a vital role at the heart of assessment, rather than a judgemental role in standard setting and maintaining in which they are becoming increasingly marginalised.

Chris Wheadon
Tuesday, 13 May 2008

References

Alberts, R. V. J. (2001). Equating Exams as a Prerequisite for Maintaining Standards: experience with Dutch centralised secondary examinations. *Assessment in Education: Principles, Policy & Practice, 8*(3), 353 - 367.

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

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

Baird, J.-A., Cresswell, M. J., & Newton, P. (2000). Would the real gold standard please step forward? *Research Papers in Education, 15*(2), 213–229.

Baird, J.-A., & Dhillon, D. (2005). *Qualitative expert judgements on examination standards: valid, but inexact*. Guildford: Assessment and Qualifications Alliance.

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

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

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

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading (Mass.): Addison-Wesley.

Bode, R. K. (2004). Partial Credit Model and Pivot Anchoring. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch Measurement*. Maple Grove, Minnesota: JAM Press.

Brennan, R. L. (2008). A Discussion of Population Invariance. *Applied Psychological Measurement, 32*(1), 102-114.

Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting*. Thousand Oaks: Sage.

Cook, L. L., & Petersen, N., S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement, 11*, 225-244.

Cresswell, M. J. (1992). *The Scale Of Changes In Public Examination Grade Outcomes: Some Reflections.* Guildford: AQA.

Dorans, N. J., & Holland, P. W. (2000). Population invariance and equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*, 281-306.

Dorans, N. J., Jinghua, L., & Shelby, H. (2008). Anchor Test Type and Population Invariance: An Exploration Across Subpopulations and Test Administrations. *Applied Psychological Measurement, 32*(1), 81-97.

Emons, W. H. M. (1998). *Nonequivalent Groups IRT Observed Score Equating (Unpublished Thesis).* University of Twente, Enschede.

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

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

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144-149.

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

Hanson, B. A., & Beguin, A. A. (1999). *Separate versus Concurrent Estimation of IRT Item Parameters in the Common Item Equating Design*: American Coll. Testing Program, Iowa City, IA.

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

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

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

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

Linacre, J. M. (2004). Rasch Model Estimation: Further Topics. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch Measurement*. Maple Grove, Minnesota: JAM Press.

Linacre, J. M. (2006). WINSTEPS Rasch measurement computer program. Chicago: Winsteps.com.

Linacre, J. M. (2008). Concurrent estimation in Winsteps.

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

Masters, G. N. (1984). Constructing an Item Bank Using Partial Credit Scoring. *Journal of Educational Measurement, 21*(1), 19-32.

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement, 16*(2), 159-176.

Newton, P. (2005). Examination standards and the limits of linking. *Assessment in Education, 12*, 105-123.

Petersen, N., S. (2008). A Discussion of Population Invariance of Equating. *Applied Psychological Measurement, 32*(1), 98-101.

Pinot de Moira, A. (2008). *Statistical Predictions in Award Meetings: How confident should we be?* Guildford: AQA.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Copenhagen: Danish Institute for Educational Research.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Iowa, IA: Psychometric Society.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Stringer, N. (2008a). *An appropriate role for Professional Judgement in Maintaining Standards in English General Qualifications*. Guildford: Assessment and Qualifications Alliance.

Stringer, N. (2008b). *Are we successfully maintaining GCE A Level Standards?*

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). One-parameter logistic model: OPLM. Arnhem: CITO.

Wheadon, C. B., & Beguin, A. A. (2007). *Fear for tiers: are candidates being appropriately rewarded for their performance on tiered examinations?* Guildford: AQA.

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

Wright, B. D. (1979). *Best Test Design*. Chicago: MESA.

Wright, B. D. (1997). *A History of Social Science Measurement*: MESA.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

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