

## IT'S A LONG, LONG TIME FROM NOVEMBER TO JUNE

### **An investigation into increasing the flexibility of delivery of high-stakes general qualifications in England through an Item Response Theory test-equating approach**

Christopher Wheadon

#### **Summary**

Since 1918, if not before, the maintenance of standards over time in the English examination system has used approaches that assume that large cohorts of candidates sit their examinations at the same time of year, every year, after following a similar programme of study over a similar time period. These assumptions present barriers to the modernisation of the examination system. Firstly, the personalisation policy agenda seeks to deliver a personalised classroom with a personalised examination timetable by 2020. Secondly, the delivery of on-screen assessment is currently being hampered by the limitations on the number of candidates that can be tested on-screen in any centre in any one sitting. Multiple parallel versions of tests would allow longer testing windows, but would pose the standard-setting problem of multiple heterogeneous populations.

Item Response Theory (IRT) test-equating approaches would seem to hold the answer as the parameters that characterise an item do not depend on the ability distribution that characterises the examinees. IRT approaches, however, depend on strong statistical assumptions that do not hold precisely in real testing situations. This research was undertaken to investigate the extent to which the invariance of item parameters would hold for a post-equating non-equivalent group design intended to maintain standards between a June and a November test session. 176 candidates from 5 schools took an anchor test consisting of items from June and November GCSE Science examinations taken by over 100,000 examinees in Physics, Chemistry and Biology. The main finding was that the design of the anchor test and features of the populations tested led to parameter invariance being violated. It is suggested that for high-stakes achievement tests in England, IRT approaches would be more straightforward in the production of multiple parallel versions of tests designed to lengthen assessment windows rather than in the provision of more frequent assessment windows.

#### **Introduction**

Norm referencing to assure year-on-year comparability has provided the basis of statistical guidance on the maintenance of standards over time in the English examination system since 1918 if not earlier (Tattersall, 2007). As candidates sit their examinations at the same time of year every year, following a similar programme of study over a similar time period, the percentage of passes at key grade boundaries in key subjects is expected to be fairly consistent over time. There is little reason to believe that, given the same amount of time to prepare, and the same level of maturity, large numbers of candidates would show any great improvements or deteriorations in performance as a whole from year to year. For the General Certificate of Secondary Education (GCSE), for example, standard practice at the Assessment and Qualifications Alliance (AQA) dictates that the same percentage of

[www.cerp.org.uk](http://www.cerp.org.uk)

candidates is expected to pass any given subject year-on-year (within a limited tolerance range) unless there is compelling evidence to doubt the stability of the cohort. This equipercentile approach has its limitations. It cannot, for example, account for changes that may be due to improvements in teaching and learning, and, if applied rigidly, would never let standards rise or fall. Cresswell's (1996) catch-all definition provides an exhaustive list of all the factors we may wish to control for in considering how pass rates may rise or fall, including everything from the prior achievement of the candidates to the quality of the teaching, but there are practical and theoretical problems with its implementation (Baird, 2007). A practical solution used at Advanced level by all the major English awarding bodies is to control for prior achievement, but this falls far short of the ideal catch-all approach.

As a consequence of the limitations in the statistical information a weak criterion referencing approach is used to maintain standards (Baird, Cresswell, & Newton, 2000). This uses the judgement of subject experts (examiners) to determine whether, given the changes in difficulty of the tests from year to year, they observe any deteriorations or improvements in performance. The logistics of this operation are quite considerable: every year, over five hundred committees of eight senior examiners (on average) are convened to make judgements on GCSE and A level examinations in England (Baird & Dhillon, 2005). Apart from being expensive, research has shown the judgement to be influenced by the question paper difficulty (Good & Cresswell, 1988), inexact (Baird & Dhillon, 2005) and possibly biased in favour of the candidates (Stringer, 2008). For this reason increasing emphasis is being placed on the statistical indicators, but quite apart from the limitations noted above, these are only valid where large relatively stable cohorts take their examinations at fixed points in the year. This situation is now changing.

In 2002 the A level system was restructured so that it became increasingly common for candidates to sit units of their A levels in January after little more than four months of study. It is likely that this trend will continue for the new format A levels launched in 2008. If it does, then this presents a comparability problem between those who sit their units after four months of study with those who sit the same units after a full academic year of study. This problem has proved quite intractable (Eason, 2008a, 2008b). For the GCSE the situation is similar: candidates can now take certain subjects in modules throughout the two year course of study. At present this is limited to three sittings per year, in November, March and June. Some candidates will therefore take the modules after three months of study, others after seven months, others after a year, and still others after two years of study. If candidates can produce a better quality of work after a year than after three months then the same percentage of candidates would not be expected to pass in the November session as the June session, for example. The extent to which they can produce a better quality of work is, however, extremely difficult to quantify.

A further challenge facing the English examination system is the personalisation agenda, represented by the policy positioning in the 2020 Vision (see Gilbert, 2006). A personalised approach to learning requires a personal examination timetable: yet this cannot be delivered while standard setting decisions need to be based on the aggregate performance of large cohorts. Even if personalised learning never becomes a reality, the drive to deliver assessments on-screen is currently being held back by the limitations on the number of candidates that can be tested on-screen in a single sitting. If multiple versions of an assessment were available then the testing window could be lengthened without the security of the assessment being compromised.

The fundamental challenge in maintaining standards over time is separating the change in the ability of the cohort from the change in difficulty of an examination paper. When the mean score on an examination increases, this could be due to a more able cohort or an easier examination paper. When large stable cohorts are taught the same curriculum over the same period of time, the latter is more likely to be the explanation. In a modular situation, however, the explanation could be a combination of both factors: an easier paper with a more able cohort or even a harder paper with a much more able cohort. Item Response Theory (IRT) would seem to hold the answer as the parameters that characterise an item do not depend on the ability distribution that characterises the examinees. IRT models performance at a question (item) level in order to separate the characteristics of the population taking a test from the characteristics of the items in that test (Lord, 1980). IRT models free the measurement of ability from dependence on a fixed set of items, and the measurement of item difficulty from dependence on a fixed population (Hambleton, Swaminathan, & Rogers, 1991). For an IRT model to be used to compensate for the variation in candidate performance that is due to the variation in difficulty of a test, however, a test-equating design needs to be in operation and some strong statistical assumptions that do not hold precisely in real testing situations need to be accepted (Kolen & Brennan, 2004). The purpose of this research is to trial an IRT method of test equating in order to examine the extent to which the assumptions of the IRT model can be violated without these violations adversely affecting the results of the test-equating.

### **The implications of violating IRT assumptions for test equating designs**

The assumptions behind the IRT models were at the heart of the controversy over IRT in the UK and beyond in the 1970s, part of a debate that rumbled on into the 1980s (Goldstein & Wood, 1989; McLean & Ragsdale, 1983). The assumption of unidimensionality requires that one ability is measured in a test (Hambleton et al., 1991); yet reality is multidimensional (Goldstein & Wood, 1989). Bejar (1983), however, provided a key clarification of the requirement for unidimensionality; that it is not necessary for a single latent trait to account for the performance of all the items in a test as long as a coherent scale can be constructed (see also Hambleton et al., 1991). IRT methods of test equating have elaborated on this premise, finding that where different dimensions have been found to exist, they appear to share the same equating function, as the same linear composite of latent traits underlies the item responses on both tests. The overwhelming consensus is that IRT methods of test equating are robust to violations of the assumption of unidimensionality within homogenous populations (Harris, 1993). Dimensionality, however, remains an empirical issue to be monitored; and less work has been done on the interactions between population sub-groups and violations of unidimensionality (Brennan, 2008).

For test-equating purposes the assumption of the invariance of item parameters is fundamental. If the parameters of items change when the items are placed in different contexts then the item parameters are not invariant and the conclusions drawn from the equating may be erroneous. A highly publicised example of the failure of this assumption occurred in the equating of tests designed to measure national progress in the US, the National Assessment of Educational Progress (NAEP). Following a major overhaul for the 1986 session the anchor items were administered in tests that differed in length, composition, timing and administration conditions. The results of the equating defied belief: 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 (Beaton & Zwick, 1990). The advice given since then has been to standardise the presentation of items that are used across versions as far as is

possible. This constraint reduces the flexibility of test construction of multiple comparable versions.

### **Choice of test-equating design**

Designing an experiment that tests the extent to which the assumptions of IRT models are violated in a live testing situation and the impact of these violations on test-equating is not a simple matter as a test-equating design is required. This ensures that some proportion of candidates takes some proportion of the same items on any two tests that are to be placed on the same scale of difficulty. If there is no overlap between either the test-takers or the tests then the two tests cannot be placed on the same scale of difficulty (Kolen & Brennan, 2004). As tests in England are currently released for public scrutiny following live use, the items in them cannot be repeated. Even if this requirement were relaxed so that a certain proportion of items were kept secret the candidates who resit the examinations, which can be a substantial proportion of the total cohort, may gain an advantage from studying the items that were asked the first time. With large item-banks these problems can be ameliorated in a variety of different ways (Béguin, 2000), but item-banks are expensive and time-consuming to develop. For the purposes of this research a pragmatic solution was the use of a Post-Equating Non-Equivalent Groups design (PENG) which is used in the Netherlands to maintain standards in national tests (Alberts, 2001).

Under the PENG design, a cohort which is not involved in the live examinations is asked to take a proportion of items from the tests that are to be equated after all the live tests have been administered. The design was executed as follows. Firstly, three experienced GCSE Science examiners were recruited to construct an anchor test consisting of items from the June 2008 and the November 2008 GCSE Science sessions. Candidates from 5 schools were then recruited to take this anchor test in the week following the live November GCSE Science session. These candidates had completed one set of GCSE Science modules in their first year of study and gone on to study further science modules in their second year in order to gain separate GCSEs in Physics, Chemistry and Biology. As such they should have a good knowledge of the curriculum and be motivated to further probe their strengths and weaknesses, although the further teaching in the second year of their GCSE Science could affect the manner in which they respond to items. GCSE Science modules are offered at two tiers (levels), higher and foundation. As the candidates for different tiers may follow different syllabuses in a way which could confound the findings from the study, only higher tier candidates were recruited. The target sample size was informed by empirical work that had shown that for dichotomous items the Root Mean Square Error of Equating (RMSE) using concurrent calibration under OPLM (Verhelst, Glas, & Verstralen, 1995) would be minimal in comparison to the standard error of raw marks for a sample size of 150 (He & Wheadon, 2008). As the test was taken after the items had been used live the security concerns were minimised, but the results of the test-equating would be available to inform the standard-setting decision on the November module.

### **Constructing the anchor test**

Two key design decisions were taken regarding the design of the anchor test. The first decision was to create a single anchor test rather than one anchor test for each separate science. This test would comprise items from all three separate sciences (Biology, Chemistry and Physics). Logistically it would have been extremely difficult to recruit schools to undertake

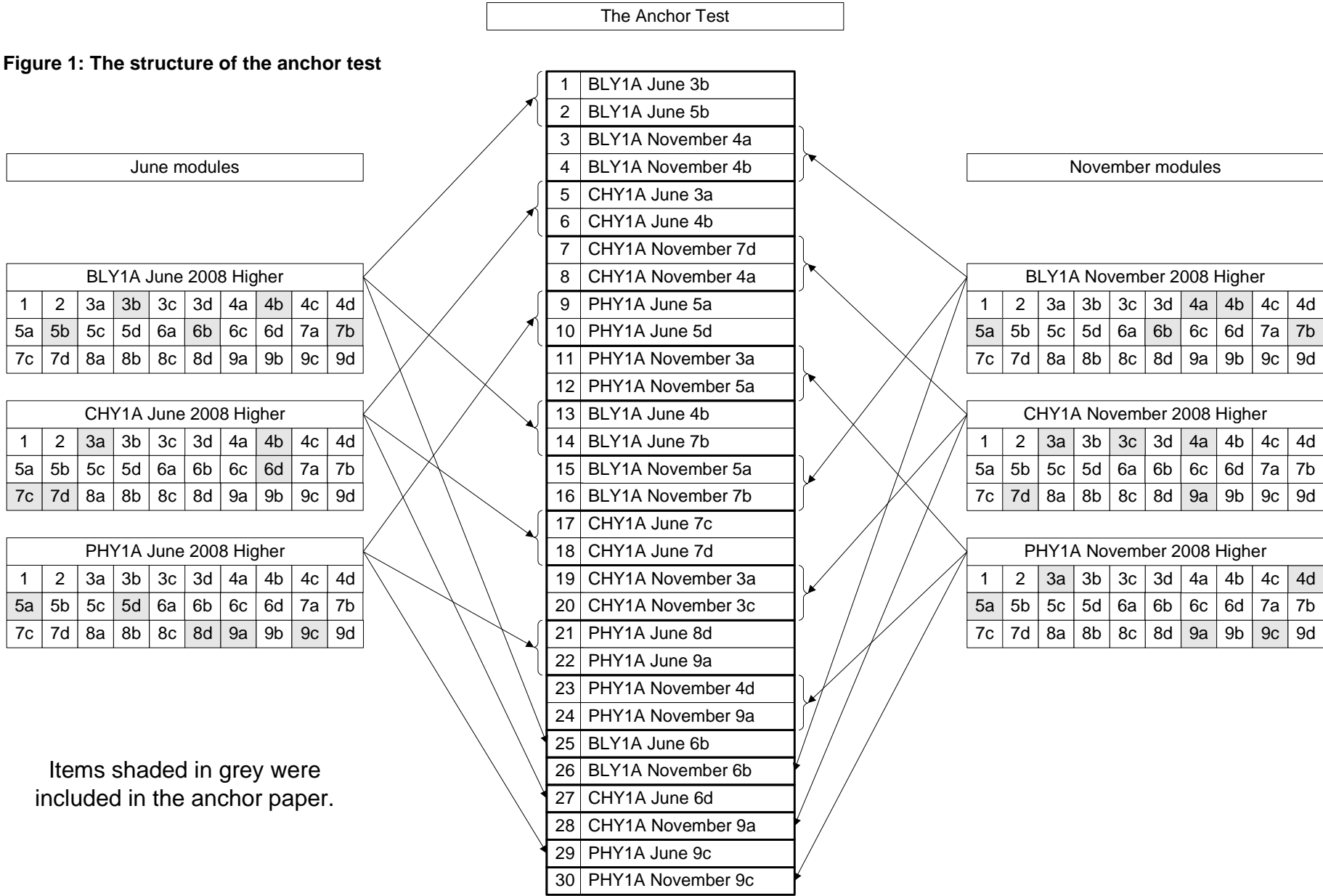
three tests, and to test only a single science would make it harder to generalise from the results. The risks of the chosen approach include: reducing the number of items in each anchor; assuming that the three sciences represent a single construct; and confusing the candidates with an unfamiliar test of 'general science'.

The reduced number of items per subject allowed by this design presented two threats to the equating. The first was that, should any of the items fail, then the equating would be based on very little information. Given the available testing time of 45 minutes it was calculated that candidates should be able to complete 30 items. This meant that each subject would be allocated 10 items; 5 from June and 5 from November. The tests to be equated consist of 30 items, which means the ratio of anchor items between each test and the anchor test is only 1:6, lower than the suggested rule of thumb of 1:5 (Kolen & Brennan, 2004). To manage this risk, the examiners were given indicators of the statistical performance of items from the June session and directed not to include those that had a low discrimination or an item facility of less than 0.3 or greater than 0.7. As the November items had not been pre-tested there was little that could be done regarding the items for that anchor.

The second threat from reducing the number of items in each anchor was that the anchor test as a whole or any of the anchors within it may no longer be representative of the construct as a whole. If one construct is under-represented in a multi-dimensional test then the equating results may not be stable (Kolen & Brennan, 2004). This threat was considered initially to be minimal as a Principal Components Analysis of Residuals (PCAR) (Linacre, 2004) had revealed most (60 to 70 per cent) of the variance in the modules could be explained by a single factor. To further reduce this threat the examiners were asked to work individually on selecting items for their own subject, but in collating the test as a whole to consider whether, where they were aware that they had not tested a specific skill in their subject, it was tested by another subject. On review of the final test they felt there was a broad range of skills included, and did not feel the need to replace any of the original items they had selected. While the anchor test as a whole should have provided good information, the individual anchors may not, however, have been representative of the construct as a whole. The content, as opposed to the skills, being tested in each anchor certainly only represented a small sample of the total content of the syllabus.

The second design decision taken was to allow examiners to choose one item from each group of four within which they are normally presented. Groups of four items follow a stimulus which can vary from one sentence to a paragraph with accompanying figures and tables. This decision was taken to examine the impact of this change in the context on the parameters of the items. Removing the items from the context facilitates test creation as examiners are able to choose the best single items which offer the greatest skill and content coverage but the change in context can have a substantial effect on the item parameters causing the test equating to fail.

To ensure that the change in item difficulty caused by the removal of context was not confounded with that due to the unrepresentative nature of the cohort taking the anchor test, a session was planned after the test-equating had been done in which the examiners could identify reasons for any changes in the relative difficulty of items. To ensure that the order in which items were taken would not systematically affect their performance, items from the different subjects were distributed evenly throughout the test (Figure 1). As the time limits for



the anchor test and the live tests are generous the risk of the items performing differently in the anchor test than in the live tests due to their ordering was considered low.

As one of the risks facing the study was the motivation of the candidates and how prepared they were for the trial in comparison with the live tests they had taken up to a year earlier, a questionnaire was devised to accompany the test (Appendix B). This attempted to gain an insight into the motivation and the knowledge of the candidates. Unfortunately only a relatively small proportion of the candidates completed the questionnaire, but some useful information was still gleaned from it.

## Anchor test results

176 candidates from 5 centres took part in the trial in the week following the live November 2008 examination session. Table 1 illustrates the number of candidates from each centre provided for the trial, and the date when the candidates had taken their live GCSE modules. Unfortunately one centre, contrary to the advice given, used candidates who had just taken the live November 2008 test in the trial and used some foundation tier candidates. The foundation tier candidates may have been taught in a structurally different manner which would introduce confounding factors to the study. The candidates who had just taken the live test may have been subject to fatigue, poor motivation and would introduce confounding factors as they were up to a year younger than the other participants in the trial. All the candidates from this centre were therefore excluded from further analyses. Two other candidates were excluded, one who achieved a near perfect score despite not having a GCSE science mark on record and one who skipped most of the items. The exclusion of these candidates left a sample size of 123, which was smaller than hoped for, but the effect of the exclusions on the item parameters was minor.

**Table 1: Number of participants in the trial and the date when these candidates had undertaken their live GCSE modules**

Centre	Live Session	Trial Candidates
A	Mar-08	41
B	Nov-07	42
C	Nov 07 / March 08	16
D	Nov-08	51
E	Nov-07	26
Total		176

From the June and November live tests a random sample of 10,000 fifteen year olds were taken from the total entries, summarised in Table 2. A sample was required due to restrictions in the software. It may seem an odd decision to sample only fifteen year-olds, as sixteen year-olds took the anchor test, but the sixteen year-olds in the live test session were re-taking the examinations in their second year of study and therefore comprised a less homogenous group.

**Table 2: Entries for the Science tests<sup>1</sup>**

	June 2008			November 2008		
	15 yr olds	Total	Proportion of 15 yr olds	15 yr olds	Total	Proportion of 15 yr olds
Biology	20,086	31,052	64.69%	63,860	85,736	75.10%
Chemistry	15,391	23,993	64.15%	55,937	73,049	77.31%

## The quality of the anchor test

The Pearson's Product Moment Correlation between the ranks of candidates on the trial and a rank derived from the average of their live GCSE science module scores was .65, which is reasonable given the reliability of the trial test (Coefficient alpha = .63) and the live tests (Coefficient alpha = .72, .74 and .77 in June 08, for example). This provides some reassurance that the anchor test was testing the same construct as the separate Science tests.

The questionnaire accompanying the test attempted to ascertain how motivated and prepared the candidates felt for the trial. Unfortunately only 44 candidates responded, but of those three quarters indicated that had the results of the test counted, it would have made no difference to their motivation in answering the majority of the topics. In terms of revision the picture was more mixed. Three quarters of the candidates said they would have performed better if they had revised hormones and oral contraceptives which required knowledge of the function of particular chemicals, for example, while only four candidates felt they would have benefitted from revision on a question involving a bar chart.

Initial screening of the item parameters revealed that the last item in the anchor test, testing the application of knowledge of electricity, had a negative item total score correlation. This item had a positive item total score correlation in the live test, but the facility was very low. As it was located at the end of the anchor test, the obvious explanation is that the motivation of the trial candidates was flagging by this point. It was therefore excluded from further analysis. One item from the live Chemistry test in June and one from the live Biology test in June, neither of which was acting as an anchor item, were excluded from the analyses due to negative item total score correlations.

## Assessing the quality of each anchor

### (i) Context

The quality of each anchor to each live test is critical if the results of the test equating are to be stable. As the number of anchor items is sparse, an OPLM (Verhelst et al., 1995) approach was used to model the item parameters. OPLM is essentially a two parameter IRT model which, in allowing discrete discrimination parameters, provides more flexibility in describing the data and therefore better model fit. As suggested by Béguin (2000), concurrent estimation of the parameters of all forms in each subject triplet (June, Anchor, November)

<sup>1</sup> Figures are not given for Physics as the equating was not undertaken. This is explained later in this report.



was used and the fit statistics for the anchor items inspected for the marginal populations (June, Anchor, November).

It is apparent that whereas the majority of the Chemistry and Biology anchor items show good fit to the model, the Physics items in the trial performed differentially in the live tests. Table 3 shows how the expected scores of candidates in the trial, derived from the OPLM model of the item for both marginal populations, were substantially lower than expected for the question illustrated in Figure 2. The final column in Table 3 represents the difference between the expected item facilities and the observed item facilities for each ability group. On presentation of this evidence the examiners were quickly able to explain why this pattern occurred on a number of the Physics items. The stimulus to each set of items presents data that can be used to answer the items that follow. Some of the difficulty in the items lies in matching the right data to the right item. In this particular item the critical information for the item asked in the trial is the number of watts rather than the life or cost of the lamps – these data are required to answer the other items in the series that were in the live test but not in the trial. It seems that while the items are not explicitly linked, they can be answered using a process of elimination. As there is only one item in the anchor paper on each stimulus there are no other items present to help eliminate the irrelevant data; in some cases this makes them more difficult. The additional information effectively performs a similar role to distracters in a typical OTQ question.

**Table 3: Observed scores and expected scores derived from OPLM for the question in Figure 2**

Ability	Number of Candidates (N)	Observed Score (O)	Expected Score (E)	Observed - Expected (O - E)	Scaled Observed - Expected (O-E) / N
Low	38	8	11.1	-3.1	-0.08
Medium	42	9	19.8	-10.8	-0.26
High	43	22	29.2	-7.2	-0.17

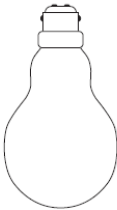
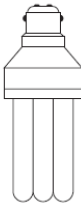

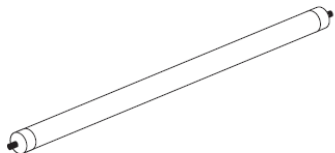
For Physics, therefore, in its current design the context of the item includes not only the stimulus, but the items associated with the stimulus, even when these items are not interdependent. Changing the context by removing the items from their neighbours causes the parameters of the items to change. This would, if overlooked, produce misleading results in test-equating. This question paper design presents potential problems for test equating, therefore, because the need to present blocks of items together conflicts with the requirement to make the anchor tests representative of the constructs in the tests being linked. If a single block of items linked to a single context is used the anchor is more likely to be unrepresentative of the live test. There are more complex IRT models available to model small groups of items as testlets (see for example Wainer, Bradlow, & Wang, 2007) but little work has yet been done on the equating of testlets.

**Figure 2: An anchor item from Physics. Question 5A was presented in the trial without the subsequent questions 5B to 5D.**

### QUESTION FIVE

The diagram shows information about four types of electric lamp.

Each lamp produces the same amount of light energy in the same time.

 <p><b>100 watt filament lamp</b> Average life = 1000 hours Cost = £0.50</p>	 <p><b>20 watt energy-saving lamp</b> Average life = 10 000 hours Cost = £3.00</p>
 <p><b>10 watt LED spotlight</b> Average life = 60 000 hours Cost = £30.00</p>	 <p><b>15 watt fluorescent tube</b> Average life = 5000 hours Cost = £5.00</p>

5A Which lamp is the most efficient?

5B Which lamp would get the hottest when it is working?

5C Which lamp would be the cheapest to run for 1000 hours?

5D You want a lamp that will provide light for 60 000 hours. You realise that you may have to buy more than one lamp to last this long. Which type of lamp would work out the cheapest to buy?

### (ii) How Science Works

While some items taken out of context put the trial candidates at a disadvantage, there were a number of items that appeared to advantage the trial candidates. As the total test score is a proxy for ability in the model, there will always be a balance between positive and negative differential item functioning. Regardless of the relative change in item difficulty that may be caused by the developed ability of an older population or by the presentation of questions in isolation rather than in blocks of four, the absolute performance of the trial candidates on one Physics item was quite impressive (Table 4). All high ability candidates answered this item correctly. The examiners identified this question as a 'How Science Works' (HSW) item, assessing scientific literacy rather than specific knowledge of Physics (see Appendix A). Their explanation for the relative advantage of the trial candidates over the live candidates on these items was that as these candidates had continued to study science their scientific literacy would have improved. This argument was supported by the questionnaire data, as the

proportion of candidates who felt they would have benefited from revision on HSW items was generally low; it is possible, of course, that even when taking the live examinations the candidates feel little need to revise HSW.

**Table 4: Observed scores and expected scores derived from OPLM for a HSW question**

Ability	Number of Candidates (N)	Observed Score (O)	Expected Score (E)	Observed - Expected (O - E)	Scaled Observed - Expected (O-E) / N
Low	38	28	12.3	15.7	0.41
Medium	42	36	21.3	14.7	0.35
High	43	43	30.5	12.5	0.29

Whereas the trial candidates were generally at a slight advantage on HSW items the picture on factual recall items was mixed. On one Chemistry item requiring knowledge of the periodic table the trial candidates appeared at a disadvantage while a Biology item on respiration and the role of sports drinks put the trial candidates at an advantage. The examiners confirmed that the Biology item was covered in more depth later in the Science syllabus whereas the Chemistry item was not. According to the questionnaire responses the candidates would have preferred to have revised both topics: nearly three quarters felt they would have done better had they revised the periodic table and nearly half had they revised respiration. Both items were subsequently excluded from the equating due to the differential functioning. The candidates' fears were not an absolute guide to differential functioning: on leaching and smelting they suffered no disadvantage in comparison to their younger counterparts, but three quarters felt they would have done better had they revised this topic.

The use of a population one year older in the trial reveals aspects of multi-dimensionality related to skills and factual recall that are not immediately apparent in an analysis of the live test results. The multi-dimensionality is only revealed in the interaction of different populations with the items. In the context of the trial the risk to the test-equating is apparent. If the anchor to the June test was entirely composed of HSW items on which the trial candidates enjoy a relative advantage, and the anchor to the November test was entirely composed of factual recall items, on which they possess no consistent advantage, then no differential item functioning would be apparent and the November test would appear to be much harder than the June test. Luckily this was not the case, with every anchor, even after the removal of the misfitting items, including between one and two HSW items and at least two non HSW items.

This finding has implications for test-equating more generally. If the same mixture of HSW and factual recall items were presented to candidates in March as in November the difficulty of the items relative to each other would change. This violates the assumption of the invariance of the item parameters. One solution may be to separate these sets of items and calibrate them separately; but not all HSW items functioned differentially, and the classification of an item as HSW can seem arbitrary - the definition of scientific literacy is inevitably subjective.

## The test equating

Items were removed by comparing the probability according to the estimated model of achieving a correct answer as a function of score group, and its 95% confidence intervals,

with a plot of the corresponding proportions calculated directly from the data. Where the item parameters for the anchor items lay outside the confidence intervals for the modelled parameters for the majority of the ability of the populations modelled they were not retained. Only one Physics anchor item remained for the anchor test to November link so equating was not pursued. The anchor items remaining for Biology and Chemistry are summarised in Table 5. Where an anchor item was excluded it was not included in the anchor test as a discrete item as a matter of expediency even though it may have shown good model fit when modelled on the trial population alone.

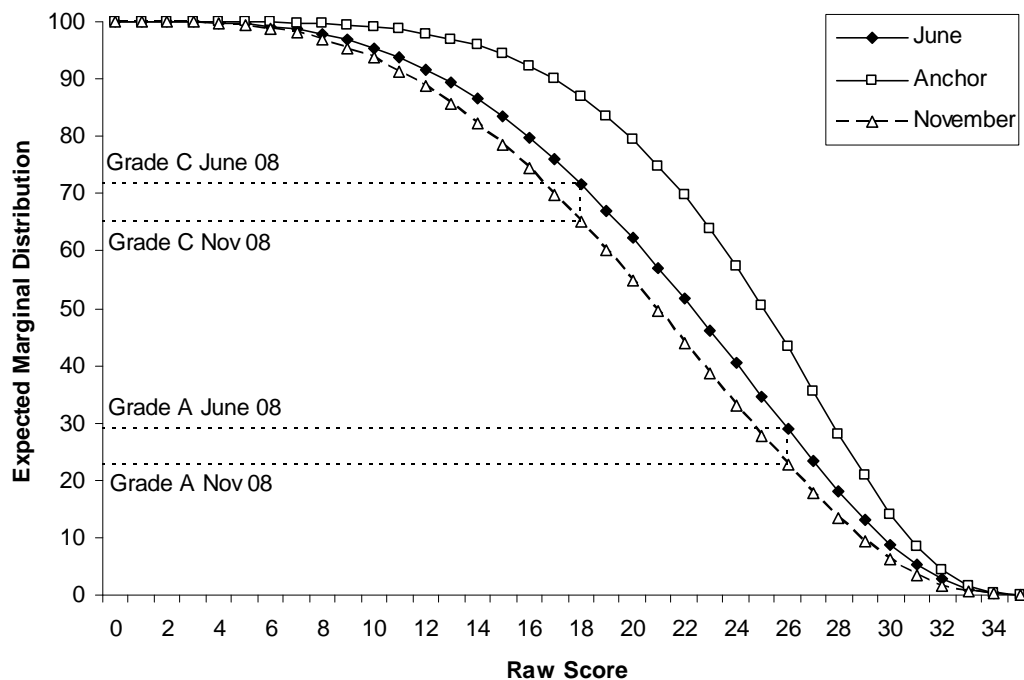
**Table 5: Number of items used in the equating**

	June 2008		November 2008		Trial
	Live	Anchor	Live	Anchor	
Chemistry	29	5	30	4	28
Biology	29	4	30	5	28

If candidates are as prepared for the November examination as they are for the June examination the same percentage of candidates could be expected to achieve each grade in November as in June. Distributions were therefore produced of the performance of fifteen year olds in June 2008 and fifteen year olds in November 2008 as the basis of the comparison with the test equating. While prior achievement measures were not available for the fifteen year-old cohort in November, previous analysis had shown that the prior achievement of cohorts in November differs little from cohorts in June.

Once the item parameters as well as an estimate of the distribution of the person parameters were produced using the marginal maximum likelihood (MML) estimation procedure based on the data in the design, an estimate of the cumulative distributions were determined for each marginal population for each test. Figure 3 illustrates how these expected distributions can be used in equipercentile equating between the marginal populations. In this example, the grade C boundary set in June 2008 produced a pass rate of 71.60 per cent for fifteen year-olds. The closest match on the expected cumulative distribution created from the sample of fifteen year-old candidates entered in June is 71.67 per cent. Reading across and down, the expected cumulative distribution for the November population on the June test is 65.11 per cent.

**Figure 3: Equipercentile equating between marginal populations on the June 2008 Chemistry live test**



The results from the test equating, summarised in Table 6, suggest that the performance of candidates in November 2008 was worse than in June 2008. These results suggest between five and eight percent fewer passes should be achieved at the key grade boundaries, equivalent to a single mark in each case. While some caution must be exercised with this finding, given the multidimensionality already noted, the consistency of the findings across grades and subject areas suggests that the result is not simply an accident produced by the specific combination of anchor items used.

**Table 6: Results from the test equating**

		June 08 (15 year olds) Cum %	Expected score June 08 Cum %	Expected score Nov 08 Cum %	Difference between June and November (%)	Grade Boundary (Equipercentile)	Grade Boundary (OPLM)
Biology	Grade A	32.10	29.82	22.56	-7.26	28	29
	Grade C	72.00	70.36	61.40	-8.96	23	24
Chemistry	Grade A	28.50	29.00	22.62	-5.88	29	30
	Grade C	71.60	71.67	65.11	-6.49	23	24

## Discussion

This research set out to understand some of the difficulties inherent in moving to an IRT approach to the maintenance of standards over time which would offer more flexibility in test delivery than current systems. The invariance of item parameters is the cornerstone of IRT, and the findings here illustrate how difficult it is to ensure that this invariance is not violated. Features of the populations and the tests both caused parameters to vary to a degree such that the equating for one out of six equating links failed. In a live situation the failure of any one link would not be conceivable as it would result in candidates not receiving grades from their examination.

If equating is to be considered for tests that have a characteristic testlet design, then it is clear that testlet design must be respected in the equating. This greatly reduces the efficiency of equating designs and potentially amplifies the effect of nuisance factors that may be introduced by the particular choice of stimulus. More needs to be known about how these nuisance factors could impact on test-equating and how they can be minimised without the validity of the tests being compromised.

A second issue highlighted by the research is multidimensionality in tests that, based on prior analysis, had appeared to consist of a single main factor or dimension. Items testing scientific literacy were relatively easier for the more mature post-equating sample than they were for the younger populations taking the tests. This is less of an issue for the provision of multiple parallel versions of a test than it is for equating designs seeking to maintain standards over time. If the balance of skills tested by items included in anchors changes from one session to the next then the test equating results may be unstable. This balance involves subjective decisions, in this instance, on the extent to which any item tests scientific literacy as distinct from other scientific skills. There may also be other dimensions: reality is indeed multi-dimensional; separating the dimensions in order to measure them is not a simple process.

The results from the test equating itself suggest that the cohorts taking examinations after three months of study are not as well prepared as those taking the examinations after a full academic year. This is of little surprise, yet the difference between the cohorts is extremely difficult to quantify in an operational context. These findings alone do not seem enough to justify suppressing the cumulative percentage pass rates every November by, say, five per cent, compared to the previous June for example. We cannot be confident that this relationship will remain stable and doesn't vary according to syllabus or aspects of the tests themselves.

It would appear on the basis of these results that the provision of multiple parallel versions of a test is more straightforward than any linking design that seeks to maintain standards over time. Multiple parallel versions would facilitate the delivery of a test over a longer test window allowing larger cohorts to be tested onscreen. Maintaining standards over time through test-equating is potentially less robust as the ability profiles of candidates change as they move through a syllabus. Where tests consist of more homogenous skills content then equating would be more likely to be robust. The maintenance of standards is not an exact science, however, and it may be preferable to live with the assumption that skills profiles do not change within a year to the assumption that ability and skills show no improvement throughout that year.

Christopher Wheadon  
June 2009

## 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.
- 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., & 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.
- Béguin, A. A. (2000). *Robustness of Equating High-Stakes Tests*. University of Twente.
- Bejar, I. (1983). *Achievement Testing: Recent Advances*. Beverley Hills: CA: Sage.
- Brennan, R. L. (2008). A Discussion of Population Invariance. *Applied Psychological Measurement*, 32(1), 102-114.
- Cresswell, M. J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgement and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues*. Chichester: John Wiley and Sons.
- Eason, S. (2008a). *GCE Information and Communication Technology (5521/6521), Conflict of unit standards between the January and June examination series*. Guildford: AQA.
- Eason, S. (2008b). *Perceived conflict between GCE Unit awarding outcomes from the January and June Examinations Series - a worked example based on AS Psychology B (5186)*. Guildford: AQA.
- Gilbert, C. (2006). *2020 Vision: Report of the Teaching and Learning in 2020 Review Group*. Nottingham: Department for Education and Skills.
- Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, 42, 139-167.
- Good, F. J., & Cresswell, M. J. (1988). Grade Awarding Judgements in Differentiated Examinations. *British Educational Research Journal*, 14(3), 263-280.
- Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of Item Response Theory*. Newbury Park, California: Sage.
- Harris. (1993). *Practical Issues in Equating*. Paper presented at the The Annual Meeting of the American Educational Research Association.
- He, Q., & Wheadon, C. B. (2008). *The effect of sample size on item parameter estimation*. Guildford: AQA.
- Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking*. New York: Springer.
- Linacre, J. M. (1998). Detecting Multidimensionality: Which Residual Data-type Works Best? *Journal of Outcome Measurement* 2(3), 266-283.
- Linacre, J. M. (2004). Mapping multidimensionality. *Rasch Measurement Transactions*, 18(3), 9990-9991.
- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum.
- McLean, L. D., & Ragsdale, R. G. (1983). The Rasch Model for Achievement Tests - Inappropriate in the Past, Inappropriate Today, Inappropriate Tomorrow. *Canadian Journal of Education*, 8(1), 71-76.
- Stringer, N. (2008). *An appropriate role for Professional Judgement in Maintaining Standards in English General Qualifications*. Guildford: Assessment and Qualifications Alliance.
- Tattersall, K. (2007). A brief history of policies, practices and issues relating to 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.
- Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). One-parameter logistic model: OPLM. Arnhem: CITO.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet Response Theory and Its Applications*. Cambridge: Cambridge University Press.



## Appendix A: A Physics ‘How Science Works’ question

### QUESTION EIGHT

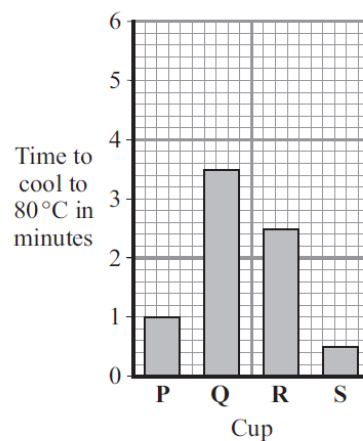
This article was in a business newsletter.

#### *Cofihouse* names ‘Insu-cup’ as its cup supplier

*Cofihouse* chose Insu-cup because of its design and product benefits. *Cofihouse* evaluated the cups from different suppliers and it concluded that Insu-cup gave the best results. It is a superior product in terms of preserving the taste of *Cofihouse* coffees and keeping the coffee hot. The material it is made from is environmentally friendly.

- 8A** *Cofihouse* measured the time that it took water at 90 °C to cool to 80 °C in the cups **P**, **Q**, **R** and **S**, from different suppliers.

They displayed their results in a bar chart.



They used a bar chart because . . .

- 1 both variables are continuous.
- 2 one variable is categoric.
- 3 one variable is independent and the other is a control variable.
- 4 one variable is continuous and the other is dependent.

## Appendix B: Questionnaire responses

These questions are not part of the test, but we would like to find out how you might have performed on the test if you had revised for it, or if you knew that it counted towards your GCSE.

Please tick the boxes below to indicate which questions you think you would have done better on with revision or if it counted; please tick all that apply.

Topic area	Subject	HSW	I would have done better if I'd revised	I would have done better if it counted towards my GCSE
Hormones and oral contraceptives	Biology	HSW	34	10
Leaching and smelting	Chemistry		33	12
LDLs and HDLs	Biology		32	11
Groups of the periodic table	Chemistry		32	14
Benzene	Chemistry		29	12
Hydrocarbons	Chemistry		28	11
Electricity (using the formula)	Physics		28	16
Uranium (using the formula)	Physics		26	15
Reflex actions	Biology		25	11
Alkanes	Chemistry		25	12
Reactivity of elements	Chemistry		22	11
The efficiency of lamps	Physics		20	13
Nuclear power stations	Physics		19	10
Carbohydrates in a sports drink	Biology		19	12
Electricity and power stations	Physics		17	11
Spit-roasts	Physics	HSW	15	10
Copper and recycling	Chemistry		15	9
Whooping cough	Biology	HSW	11	10
Solar cell panels	Physics		10	11
Smoking and disease	Biology	HSW	9	16
Vitamin C	Biology		7	11
Insulation	Physics	HSW	7	11
Drug trials	Biology		7	10
Quarries	Chemistry	HSW	5	10
Infections in maternity wards	Biology	HSW	5	9
Bar charts	Physics	HSW	4	12
Total			44	44

	TRUE	FALSE
I think I have done as well on the test as I did last year	16	22