An Item Response Theory Approach to the Maintenance of Standards in Public Examinations in England

WHEADON, CHRISTOPHER,BRIAN

An Item Response Theory Approach to the Maintenance of Standards in Public Examinations in England

Christopher Wheadon

Thesis submitted for the Degree of Doctor of Philosophy

2010

Table of Contents

List of tables

List of figures

Acknowledgements

This thesis is the result of many years of study of assessment at both the CEM Centre at Durham University and the Assessment and Qualifications Alliance (AQA). I am indebted to both organisations for financing this work and for allowing me the time (where thanks should also go to Ireni) to complete it.

I would like to thank staff at both institutions for their help, support, guidance and encouragement throughout that time. I am particularly grateful to Carol Taylor Fitz-Gibbon for introducing me to the world of quantitative research, to Peter Tymms and Robert Coe for their supervision. I would also like to thank the staff at Cito for their training in the use of OPLM, Anton Béguin and Qingping He for sharing their ideas, and the Research Committee at AQA for their patient reviewing of early stages of this work.

Further, I should like to thank Melody Charman for her assistance with running my research projects and preparing drafts for this manuscript, Lindsay Simmonds for her proof reading and Debbie Miles for her help in organising its submission.

Finally I would like to thank Michelle Meadows for her support and guidance throughout this project, without which I would surely not have completed this work.

Abstract

Every year outcomes from public examinations in the UK rise: politicians congratulate pupils on their hard earned achievement; the media questions whether this achievement is real; those responsible for administrating the examinations defend their standards; various subject councils and employers decry the failings of candidates with high grades; admissions officers from the elite universities report their struggle with the decrease in discrimination in grades achieved; and academics debate what it means to compare standards from one year to the next. The debate cannot be easily resolved because examination results are put to many purposes some of which are more suited to certain definitions of comparability than others. In procedural terms, however, it should be relatively straightforward to evaluate the strength of the evidence that is put forward on the comparability of standards against various definitions.

Broadly, solely in terms of discrimination, the statistical evidence in the maintenance of standards over time and between qualifications can be evaluated by reference to measures such as model fit, significance and effect size. An evaluation of the literature suggests that predictive statistical models, where employed in the maintenance of standards to meet definitions of cohort referencing, tend to be robust. Beyond discrimination, measures of performance standards are required to support inferences drawn from grades on what candidates can actually do. These are, and have been for many years, underpinned by processes reliant on human judgement. An evaluation of the literature suggests that judgement provides very weak evidence and is subject to unknown bias. The combination of statistical and judgemental

evidence is poorly specified, has no theoretical basis and is therefore impossible to evaluate. If anything more than pure cohort referencing is required from public examinations in the UK there is clearly a need to explore models with a sound theoretical basis whose evidence can be evaluated in terms of model fit, significance and effect size.

The task of maintaining a performance standard can essentially be reduced under test theory to making comparisons between persons that are independent of the items on the basis of which these comparisons are made. Test theory however has been sparingly applied to comparability issues in UK public examinations. This study considers which test theory model would be most suited to the examinations in use in the UK, examines issues of model fit under frequentist and Bayesian frameworks, compares the results from different test equating methods and the practical issues of implementing a test equating design under the given constraints of the UK examination system.

To begin with the Rasch model and the One Parameter Logistic Model were fitted to operational data gathered from examinations in a range of subject domains where marking reliability would not be considered as a potential confound. In this framework the Rasch model requirement of a single discrimination parameter across items appeared overly restrictive. Further, potential issues with model fit were highlighted related to dimensionality, guessing and weak local independence. More complex models were therefore pursued under a Bayesian framework. The Posterior Predictive Model Checking Procedures and Deviance Information Criterion

confirmed that a model which allowed discrimination to vary across items, such as the two-parameter Item Response Theory model, would produce better model predictions. Use of the Multi-Class Mixture Rasch Model suggested that multidimensionality due to a confounding speededness factor could result in misleading inferences being drawn from unidimensional models. The Testlet Response Theory model showed enhanced predictions where weak local independence was correctly specified; however it proved difficult to specify where this weak local independence was expected. When tests from one of the examinations particularly affected by speededness were equated OPLM proved more robust to the confounding speededness factor than the Rasch model.

A Post-equating Non-Equivalent Groups Design was then set up as an experiment using a set of relatively simple Science examinations and candidates at a later stage in their programme of study than those who would take the live examinations in order to understand some of the practical issues involved in equating designs. The study found that item parameters were not stable across samples due to context effects, school effects and maturity effects. These results were partly due to the scale of study, which, though small, still produced reasonably sensible outcomes. It is suggested that more care paid to the context of linking items, their underlying construct, and the sampling of schools would yield more robust results. Finally, a qualitative exploration of views related to test equating designs suggested that teachers, pupils and examiners would not reject the possibility of embedding equating items into live tests.

For examinations where marking reliability is not considered an issue the results reported here suggest that the use of test theory could provide a unified theoretical framework for the maintenance of standards in UK public examinations which would allow the strength of the evidence presented to be evaluated. This would represent a substantial improvement over the current situation in which no comprehensive or coherent evaluation can be made. The time and investment required, however, to introduce such a framework is also substantial. A suitable technical infrastructure is required as well as psychometric expertise. The alternative is to revert to an examinations system that is essentially cohort referenced and focuses on discrimination between candidates in any one year rather than attempting to quality assure, as it cannot do, performance standards from one year to the next.

1. The System of Public Examinations in England

1.1 Overview

This chapter describes the public examination system in England, and outlines some of the major challenges for this system. These include its multiple agency composition, the nature of the assessments and the multiple purposes of the examinations. It then considers what it means to maintain standards within this system and discusses the theoretical and practical difficulties that beset current approaches to maintaining standards. It concludes that these current approaches are inadequately specified and that there is a need to explore alternative models.

1.2 The public examination system in England

1.2.1 GCSEs and A-levels

The public examination system in England is based around the Advanced level (A-level) and the General Certificate of Secondary Education (GCSE). GCSE examinations were introduced in 1988 and are the principal means by which 16-year-olds are assessed at the end of their compulsory education. Nearly all students in state maintained schools study for GCSE examinations. They are available in a wide range of subjects from traditional disciplines such as mathematics to more modern areas such as media studies and photography. Courses leading to GCSEs are intended to be followed for two years although some candidates take them after one year of study.

After gaining GCSEs a proportion of students will stay on at school or college for a further two years to study for A-levels, which act primarily as pre-university examinations. As with GCSEs, A-levels are available in a wide range of subjects encompassing traditional and more modern subject areas. A-levels were introduced in the 1950s and were originally targeted at a fraction of the national cohort. In keeping with government objectives to increase participation in higher education they are now taken by over a third of the national total cohort.

Apart from GCSEs and A-levels there are two other public qualifications worth noting. In 2010 the Diploma introduced a practical element to learning. It represents a portfolio qualification encompassing GCSEs and A-levels as well as specifically designed content in areas such as engineering, and health and beauty. From 2011 all pupils in England will also be expected to have passed Functional Skills qualifications in English, Mathematics and ICT**.** Functional Skills are practical skills intended to allow individuals to work confidently, effectively and independently in life, but they are largely examined through traditional test formats.

1.2.2 How qualifications are developed

The development and delivery of qualifications in England is a complex inter-agency process (Meyer, 2009b). Central government influences policy on education and qualifications through the Department for Children, Schools and Families (DCSF). The DCSF owns the national curriculum in England which applies to all pupils of compulsory school age in state maintained schools. The department decides which subjects are statutory for 14 to 16-year-olds and is responsible for the programmes of study that must be followed in those statutory subjects. Once decisions on examinations policy have been formulated within the DCSF the examination

regulator Ofqual is then charged with overseeing the quality and value for money of the qualifications system, while the Qualifications and Curriculum Development Agency (QCDA) may be commissioned to update programmes of study and subject criteria.

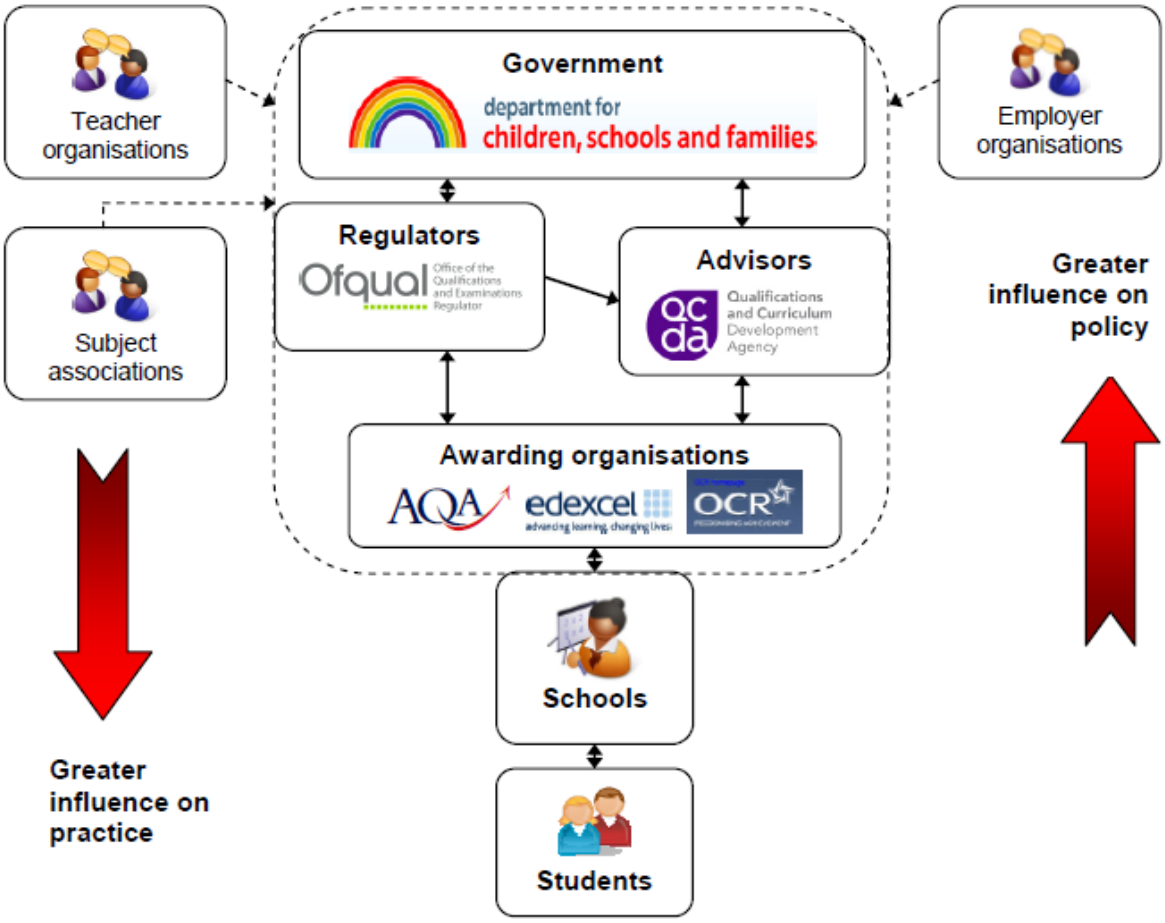


Figure 1.1: Organisations involved in the development of qualifications in England (Meyer, 2009b)

1.2.3 Qualifications criteria and subject criteria

The Qualifications Criteria for A-level and GCSE give broad rules on the structure, assessment and grading of each qualification type. The subject criteria explain the general aims of studying the subject and outline the essential knowledge, skills and understanding that should be present for all qualifications in that subject. The subject criteria indicate the assessment objectives as well as the type of assessment that can

be used within the qualification. The subject criteria for GCSEs and A-levels include descriptions of the standards of achievement that are expected to have been shown by candidates achieving specific grades.

To establish the qualification criteria and the subject criteria, the regulators or QCDA as commissioned by the regulators, consult with teachers and lecturers, subject associations, teacher associations, professional associations, employer organisations, as well as the organisations that will develop and deliver qualifications, which are known as awarding bodies. Once the criteria and the regulations have been finalised and approved by Ofqual, awarding bodies can then use them to develop their qualifications.

1.2.4 Awarding bodies

Any institution from a large company to a small charity may become an awarding body as long as they satisfy certain statutory regulations. Codes of practice govern their operations and procedures and are intended to ensure that candidates get a fair deal irrespective of the awarding body that is delivering the qualification. Awarding bodies, once they have developed their qualifications, market them directly to schools and colleges who are free to choose those qualifications they feel will best meet the needs of their pupils. The largest organisations offering GCSEs and A-levels in England are: the Assessment and Qualifications Alliance (AQA); EdExcel; Oxford, Cambridge and RSA (OCR); the Welsh Joint Education Committee (WJEC); and the Council for the Curriculum Examinations and Assessment (CCEA).

1.2.5 The structure of qualifications

An A-level qualification consists of advanced subsidiary (AS) and A2 units. The AS is a stand-alone qualification and is worth half a full A-level qualification. It normally consists of two units (assessed at the standard expected for a learner half way through an A-level course) that together contribute 50 per cent towards the full A-level. The A2 is the second half of a full A-level qualification. It normally consists of two units (assessed at the standard expected for a learner at the end of a full A-level course) that together are worth 50 per cent of the full A-level qualification. Units are available in January and June and candidates can re-sit units if they wish.

A GCSE qualification consists of between two and four units, which may be available at different test levels known as tiers for different levels of ability. These units may be available in November and February as well as in January and June. Candidates are only allowed to re-sit each unit once.

1.2.6 Assessment modes

The majority of units at GCSE and A-level are externally assessed: they are set, marked, and graded by the awarding body. Most GCSE and A-level qualifications will include however one internally assessed unit. This may be a piece of coursework, a project or a practical assessment. These are marked and graded by teachers in schools and colleges. Traditionally external assessments have favoured the use of extended response items such as essays. Multiple-choice items were in fashion in the 1970s but have fallen into disfavour. More recently, constructed response item formats have tended to predominate. Questions often follow stimuli such as reading passages, diagrams, tables or pictures. Items may be ordered

thematically as well as in some notional order of difficulty. Pre-testing is rare and all items are made available to the public free of charge after use.

1.2.7 How assessments are marked

Most assessments are marked by subject experts following a period of training and standardisation on the mark schemes. Recent developments in technology have allowed an increasing proportion of this marking to be done on screen. While in the past one marker would be responsible for marking an entire script it is now possible, and indeed common practice, for an electronic script to be divided amongst markers. If a script is divided amongst enough markers it is now possible, for constructed response items, to make the assumption of random equivalence between markers (Maris & Bechger, 2007).

1.2.8 Marks and grades

In order to compensate for the variability in difficulty of tests that are presented to candidates over time, candidates' marks are converted into grades. A grade A in one unit taken in one session is primarily intended to be equivalent to a grade A that is achieved on that same unit taken in a previous session. Once a unit has been graded the candidates' raw marks are converted into a scaled score. When a candidate has accumulated enough units they may certificate for that qualification and their scaled score will be converted into a grade on the qualification. The A-level is graded from A* to E as well as fail, and the GCSE is graded from A*to G as well as fail. It is these grades that have public currency and which carry the weight of public expectation.

1.3 Expectations of the public examination system in England

1.3.1 The purposes of public examinations

Newton (2007, 2008) illustrates the variety of purposes to which examination grades can be put, and highlights three key purposes for public examinations:

1. Qualification – in which individuals are judged as equipped to succeed in a certain job, course of instruction or role in life
2. Selection – to provide information for future educational and vocational selection decisions
3. Programme evaluation – in which results are used to judge the success of educational initiatives nationally and locally.

These different purposes suggest different conceptions of comparability of examination grades both between subjects (Coe, 2007) and over time.

1.3.1.1 Conceptions of comparability

Coe (2007) distinguishes three conceptions of comparability. Performance comparability relates to the skills, knowledge and understanding required to achieve a certain grade. Under this conception of comparability it would be important to ensure that candidates who achieved certain grades, from one year to the next, knew, understood and could do the same things. This is reflected by comments such as the following:

> 'I need to receive a consistent message over a number of years as to what a grade A or B represents in terms of knowledge, skills and competencies.' (Binks, 2002, p.6)

It is relatively easy to find evidence that grades awarded today do not, however, relate to the same skills as they did five or twenty years earlier (see, for example, Engineering Council, 2000). As the needs of society have changed so have the curriculum and the examinations. This has led some to conclude that maintaining (or comparing) standards over long periods of time is a futile effort (see, for example, Christie & Forrest, 1980). This conception of comparability clearly relates to the purposes of qualification and programme evaluation as defined above, and suggests the need for scrutiny of candidates' achievements in the process of maintaining standards, which will be discussed later.

Coe's second conception of comparability is that of statistical comparability. Under this conception comparability holds when a typical candidate has an equal chance of achieving a particular level in successive examinations. Under this conception examinations are useful to the extent to which they successfully rank order candidates. If the key purpose of examinations is selection then statistical models such as regression or multilevel models would probably suffice in the maintenance of standards.

Coe's third and final conception of comparability is that of construct comparability. Under this conception two examinations are comparable if performance of typical examinees with the same latent ability leads to the same grade. The construct could be general aptitude, or something more specific such as reasoning. This conception relates to both qualification and selection as it could be

considered that candidates with, say, higher reasoning skills will be better at doing certain jobs and more likely to succeed in further study. This conception of comparability, which implies the use of measurement models or latent trait models, has been lesser explored in the maintenance of standards in the UK.

1.3.1.2 Prioritisation of purposes

If one or other purposes of examination results were to be prioritised then the task of maintaining standards would be a great deal simpler. Newton (2005a) calls this the diktat model. Perhaps the prime exponent of the diktat model is Cresswell (1997, 2000). Drawing on historical precedent and a study of their most prevalent use he argues that the rationing process for future meritocratic educational and vocational selection decisions is the primary purpose of public examinations. In maintaining standards, therefore, we start by establishing the validity of the qualification by design, and then ensure parity of achievement by statistical means. The validity in the design of the assessments ensures we know what candidates can do, the statistics ensure that they are fairly rank ordered within a space of one to two years. This position is underpinned by both theoretical and pragmatic reasoning; it is neither theoretically tenable nor practically possible to award grades in a way that will satisfy the purpose of qualification to a certain performance standard. At the heart of this reasoning is the issue of whether or not, in awarding grades, we can provide any objective measure of performance.

1.4 Awarding grades

1.4.1 The purpose of awarding grades

The purpose of awarding grades is to ensure that candidates receive a fair result regardless of the difficulty of the question paper they have been set. This may mean, for example, trying to ascertain the mark that last year's borderline grade A candidate would have gained on this year's paper. The complex structure of qualifications in England, however, can require certain adjustments to be made to this standard. This may be to bring the relative standards of qualifications offered by different awarding bodies into line, or to ensure, for example, that different routes to a qualification are of the same level of demand (Béguin, Wheadon, Meadows & Eggen, 2007). Two theoretical positions inform the process by which the borderline marks for key grades are determined: judgement and statistics.

1.4.2 The theory behind judgement

Judgemental approaches to determining grade boundaries in England draw on a strong criterion referencing approach. Strong criterion referencing finds its basis in the idea that a standard can be described and made explicit (Cresswell, 2000). Taking a set of observable, well defined qualitative characteristics of work at a certain level a judge is assumed to be able to synthesise an overall judgement of each script. Cresswell (2000**)** criticises this position on a number of levels. Firstly, drawing on Reader Response Theory he argues that we bring our own expectations to texts such as performance descriptions and impose these upon them. It is therefore linguistically naive to believe that performance descriptions can be objective. The reading of scripts generates a frame of reference which may then affect our

interpretation of the scripts that follow but also retrospectively transforms our original understanding of those performance descriptions. Secondly, he argues that it is sociologically naive to believe that pure performance can be distilled from the context in which it is sampled. In this he draws on the literature of experts and novices which suggests that experts develop holistic skills specific to their area of expertise while non-experts need to apply rules to solve problems. The rules being applied by the non-experts can easily be disrupted by unfamiliar contexts in a manner which is unforeseen by the experts. Lastly, he draws on work in artificial intelligence to argue that it is psychologically naive to believe that an objective overall evaluation can be made from all possible quantitative judgements on various dimensions.

The literature on human judgement would seem largely to support Cresswell. Nietzsche, for example, concluded that,

> The falsity of human judgement derives firstly from the condition of the material to be judged, namely very incomplete, secondly from the way in which the sum is arrived at on the basis of this material, and thirdly from the fact that every individual piece of this material is in turn the outcome of false knowledge, and is so with absolute necessity. (Nietzsche, trans. 2004, p. 28)

More recent research has suggested that humans are better at comparative judgement than absolute judgement. Donald Laming's (2004) work suggests that without some physical reference such as a ruler or a scale against which a comparison can be made, five categories of judgement is the pragmatic limit.

Comparison of one examination script with another falls far short of being such a ruler. They are a sample of incomplete linguistic artefacts which display inconsistent performance in the face of contextual difficulties that cannot be perceived by the experts judging them.

The theory is only as good as the data that supports it. On criterion referencing there are plentiful examples of high profile failures. Baird (2007) recounts the introduction of criterion referenced examination into New Zealand in 2004, following a significant teacher training programme to ensure that standards were widely understood. The pass rate in the scholarship examinations dropped to half that of the previous year and there was an outcry over the overall pass rate, as well as variability between subjects. The pass rate for physical education, for example, was 0 per cent. Throughout the 1980s in Britain there was a considerable amount of work done on the notion of grade criteria for public examinations. It became clear, however, that grades awarded by conventional procedures could not adequately be described by the criterion referenced performance descriptions (Baird, 2007).

1.4.3 The theory behind statistics

Statistical approaches to the maintenance of standards in England have been concerned with calculating the probability of a group of examinees achieving the same distribution of grades had they taken a specific unit at a different time or with a different awarding body, for example. In order to calculate this probability a statistical model is required. A fundamental problem is therefore immediately apparent in that no statistical model can claim objectivity, as this extract from the theory of econometrics rehearses,

> The false idol of objectivity has done great damage to economic science. Theoretical econometricians have interpreted scientific objectivity to mean that an economist must identify exactly the variables in the model, the functional form, and the distribution of the errors. Given these assumptions, and given a dataset, the econometric method produces an objective inference from the dataset, unencumbered by the subjective opinions of the researcher.
>
> This advice could be treated as ludicrous... the econometric art as it is practised at the computer terminal involves fitting many, perhaps thousands, of statistical models. One or several that the researcher finds pleasing are selected for reporting purposes (Poirier, 1988, p. 183)

In fitting a statistical model to examination outcomes it is conceivable that we may wish to control for factors such as prior ability, school type, gender, social and economic status, levels of preparation and motivation, as well as a host of other measurable and unmeasurable variables that may impact on examination outcomes. Such a statistical definition has become known as the catch-all definition (Baird, Cresswell & Newton, 2000). Cresswell (2000) reports on the conceptual difficulties of controlling for varying proportions of male and female candidates entering for specific examinations. It could be argued that the association of any of the control variables with examination outcomes reflect bias in those examinations rather than genuine differences in the general ability of the candidates (Baird et al., 2000).

1.5 Maintaining examination standards in practice

The practice of maintaining standards in England draws on both judgement and statistics. Every time an examination paper has been sat and new grade boundaries need to be determined to compensate for changes in the difficulty of that paper, awarding meetings are held. In these meetings a committee of the senior examiners, who have written and overseen the marking of examination papers - known as the awarding committee, or awarders - compare candidates' work from the current year in comparison with work archived from the previous year and in relation to the published descriptors of the required attainment at particular grades (Meyer, 2009a). The committee is also advised of the relevant statistical indicators such as the actual distribution of marks achieved and the details of the entry pattern from year to year, as well as more sophisticated data in the form of a predicted distribution of candidates' achievement (Meyer, 2009a). Their task is to combine their qualitative judgement with the statistical evidence to arrive at final recommendations for the new grade boundaries.

1.5.1 The practice of using judgement to maintain standards

Cresswell (1997) drew on a dataset of grade boundaries that were set purely by a process of qualitative judgement of candidates' work on successive occasions in a broad range of A-level examinations which attracted over 500 candidates. Attempting to explain the variation in outcome over these years he considered the possibility that each cohort was a sample from a population of all candidates who had taken that examination over its lifetime. Applying a standard *z* test he concluded the difference in outcomes between the two years could not reasonably be viewed as

the results of random variations between successive groups of candidates. The distribution of *z* statistics was clearly dominated by extremes at either end. Investigating subgroups of candidates by gender and centre type he was then able to dismiss the possibility that changes were due to variations in the relative proportions of these subgroups. Subsequent work has shown that large variations in the relative proportion of subgroups have little impact on the distribution of expected grades (Eason, 2010). As the entries from one year to the next were relatively stable he could also dismiss the explanation that a subgroup who were previously missing from the sample had entered the examination. Finally, he showed that while there were no year-on-year trends related to gradual improvement or deterioration in the work of candidates there was a clear relationship between changes in marks and the position of the final grade boundaries. The strong suggestion was that the examiners' qualitative judgement takes insufficient account of changes in the difficulty of the examination papers and/or their marking. He concluded that the use of judgement in maintaining standards will lead to year-on-year fluctuations in outcomes that cannot be justified.

This finding is consistent with earlier research that the judgement of subject experts is susceptible to bias dependent on the difficulty of those tests,

> The awarders tended to consider fewer candidates to be worthy of any given grade on harder papers or, alternatively, that more candidates reached the required standards on easier papers. (Good & Cresswell, 1988b, p. vii)

Subsequent research on expert judgement has not been encouraging. In one experiment (Baird & Dhillon, 2005) eight GCSE English and ten A-level Physics awarders were given fourteen mark-free scripts in a seven mark range from around a grade boundary and were asked to rank order them: mean correlations were low at grade A for both subjects (0.16 and 0.21 respectively) and slightly better for GCSE English at grade C (0.42 and 0.20 respectively). Even at their best, in GCSE English grade C, 25 per cent of the judgementally determined boundaries were more than two marks away from the actual boundary. While these differences may seem small, a recent natural experiment related by Stringer (2010) highlights the potential deviation from the true standard that may occur when judgemental evidence is weak. In a recent examination series a change in entry pattern led to a number of A-level committees being provided with statistically recommended boundaries that were higher than they ought to have been. One committee, perhaps because the work they scrutinised seemed to be of a very high standard or perhaps because the marks were so high, spotted the mistake, and triggered an investigation. Other committees were less successful. It emerged that one committee had recommended a grade E boundary of 37 out of 80, which it transpired was 17 marks, or 21.3 per cent of the total marks – from the statistically recommended boundary. If the mistake had not been spotted 14.7 per cent more candidates would have failed the examination than should have. The committee's decision was approximately two grades out.

Laming (2004) suggests that when faced with weak evidence people are unable to resist extraneous suggestion. In this case the extraneous suggestion came from the statistical evidence, which was misleading. It is quite conceivable therefore that committees are susceptible to a range of extraneous influences, some more valid than others, when making their judgement on scripts. These practical problems in

discerning the standards of examination scripts are consistent with judgemental approaches in a variety of disciplines. Laming's comprehensive (2004) review finds examples where judgement, in the absence of clear evidence, is found lacking in the appraisal of art and literature, the identification of children at risk of abuse, eyewitness identification and criminal investigations. Human judgement, he concludes, is much poorer than is generally supposed. In public examinations there is evidence that judgement of scripts is biased in favour of the candidates (Stringer, 2008); that this judgement is unaffected by comparisons with the work archived from previous years (Baird, 2000); and that this judgement is inappropriately affected by the consistency of a candidate's performance (Scharaschkin & Baird, 2000).

1.5.2 The practice of using statistics to maintain standards

Since 2000 predictions based on candidates' prior achievement have been routinely produced to provide statistical guidance in the process of standard setting for A-levels in England. At AQA they have increasingly been used for GCSEs. Currently, the predictions for $Year^{(x)}$ are generated using the following rules paraphrased from Eason (2003):

1. For a given A-Level specification, match each candidate's $Year^{(x-1)}$ A-Level grade with their $Year^{(x-3)}$ mean GCSE grade.
2. For each A-Level subject, subdivide the $Year^{(x-1)}$ A-Level entry population into ten distinct categories based on candidates' mean GCSE grades; the first category containing candidates with the best mean GCSE grades and the tenth category containing candidates with the worst mean GCSE grades.

3. Separately for each of the ten categories created in step 2, generate the achieved Year$^{(x-1)}$ A-Level distribution of grades.
4. For the AQA Year$^{(x)}$ specifications, match candidates with their Year$^{(x-2)}$ mean GCSE grades.
5. Subdivide these candidates into the same ten distinct categories as described in step 2. For each category assume the Year$^{(x)}$ AQA A-Level grade distribution will be the same as that achieved by the Year$^{(x-1)}$ all awarding body A-Level candidates.
6. Calculate an overall predicted grade distribution for the Year$^{(x)}$ A-Level specifications by weighting the category-by-category expected grade distributions by the numbers of candidates in each category in Year$^{(x)}$.

Fitting a surrogate parametric proportional odds model, Pinot de Moira (2008) explored the limitations of the predictions supplied to awarding meetings. She found that the size of the entry, the skew of the independent variable (for example, mean GCSE category) and the actual value of the predicted grade outcome are of importance in assessing the worth of the predictions in an awarding situation; but that predictions for A-levels with entries of over 5000 candidates were accurate to within 2 per cent in every case at grade E and in 87 per cent of cases at grade A. For GCSEs the accuracy was slightly lower. The model assumes of course that outcomes for candidates of a given prior ability should remain the same year on year, and that the baseline remains stable. Where there are doubts regarding the stability of the baseline the baseline measure can be standardised to ensure that stability.

1.5.3 Combining judgement and statistics

As has already been described the standard practice in maintaining standards in public examinations in England is to present awarders with the task of combining their judgement with the statistical indicators. This model has been called weak criterion referencing (Baird et al., 2000). In essence the awarders are required to compare the performance of candidates having taken into account the relative difficulty of the items they have answered. This practice has not been without its detractors,

> No estimate of the cost of these mechanisms and their various procedures to institutions (and the taxpayer) has yet been done but it is likely to run into millions every year. These processes are compounded by assessment models that combine the goals and processes of outcomes and criterion-referencing with remnants of norm-referencing. Not only is this arcane and complicated, even for those inside assessment systems, but the overall effect is to create an ideological, epistemological and technical quagmire around standards and confusion about how to measure validity and reliability. (Ecclestone, 2006, p. 8)

Stringer (2008) describes the practical problems of this combined approach as follows:

> Informing examiners of the average difference in prior attainment between two cohorts cannot help them to make comparisons between

> individual scripts from the two cohorts because the examiners do not know anything about the individuals whose scripts they are comparing. Even if the examiners had access to the necessary information about each of the candidates whose work they scrutinised and a computer to analyse these data, such an underpowered analysis would be unhelpful for quantifying the effect, and possibly unreliable at indicating even its sign. The statistical and judgemental evidence, rather than converging on the same concept of difficulty, point at distinct concepts. (p.3)

Rather than converging on the same solution, the statistics and the judgement may point in different directions or be of a different magnitude. In the short term this may produce the sort of unexplained variation that Cresswell isolated; in the longer term the impact may be more damaging.

Figure 1.2 shows overall AQA A-level final outcomes and predicted outcomes at grade A between 2002 and 2007. Attempting to explain the apparent increase over this period Stringer (2008) considers the role of both predictions and judgement in this increase. He concludes that even when baseline measures for the predictions are standardised, the predictions are cumulatively inflated by small adjustments made every year to the grade boundaries according to the awarders' judgement. It is of course impossible to discount the possibility that the standard of work may actually have improved to this extent over this period, but given the finding that awarders cannot distinguish between scripts within a small range of marks and their tendency to give candidates the benefit of the doubt, Stringer's conclusion

that the current awarding procedures do not adequately maintain standards seems justified. Stringer finds corroboration in Tymms and Fitz-Gibbon who quote a report by SCAA/Ofsted[1] from 1996, which acknowledges that "the emphasis given to awarders' judgement of the quality of candidates' work rather than to statistical data, coupled with a tendency to choose the lower of two scores when there is a decision to be made about setting the minimum mark for a grade, may have allowed small, unintended but cumulative reductions in grade standards in successive years" (SCAA/Ofsted, 1996 cited in Tymms & Fitz-Gibbon, 2001, p.166).

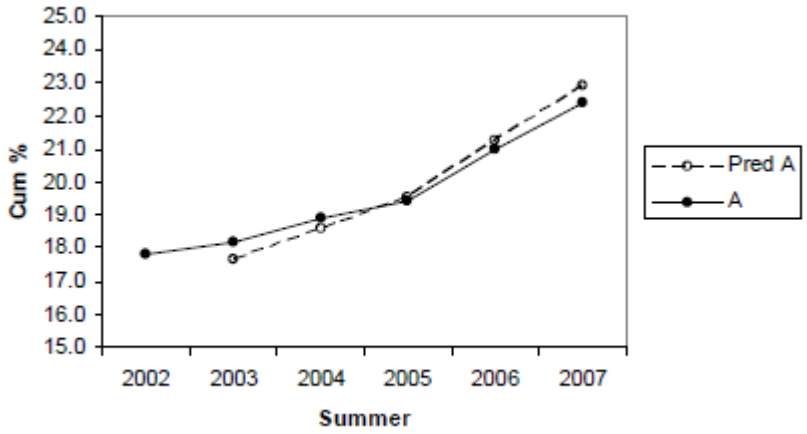


Figure 1.2: Overall AQA GCE final outcomes and predicted outcomes at grade A between 2002 and 2007

The situation is reminiscent of the fatal swim that Laming (2004) recounts in which a coach refused to allow her swimmer to abandon her attempt to swim the English Channel. Despite the advice of the official observer and the skipper of the support vessel that the swimmer was in extreme distress the coach believed that the swimmer would succeed and should carry on. Laming's analysis is that the coach has been watching her swimmer constantly throughout the swim, searching for signs of

[1] School Curriculum and Assessment Authority (SCAA)/Office for Standards in Education (Ofsted)

deterioration in her condition. The very fact that she has been paying such care means she does not notice the very small but continual increases in distress from one instant to the next. The official observer and the skipper of the trawler were not constantly observing the swimmer so the cumulative increases in the distress were more apparent. The swimmer drowned. Year after year, chairs of examiners testify to the fact that their recommended grade boundaries carry forward standards over time in as much good faith as the swimmer's coach. The danger is that the incremental changes have created extreme distress in the system of public examinations.

Even if it is accepted that these increases represent genuine improvements in performance the way in which judgement and statistics are combined may differ between awarding bodies and lead to quite different definitions of standards. The regulators' Code of Practice is silent on how judgement and statistics should be combined (Jones, 2009a). This inadequate specification is not conducive to producing consistent standards (Cresswell, 2010).

1.6 Discussion

It is clear that the public examination system in England carries the expectation that in maintaining standards the actual performance of candidates in terms of their knowledge, skills and competencies is clearly described, considered and maintained. In order to meet this expectation qualitative judgement has been employed despite the weight of evidence that suggests it is unable to make the fine discriminations that are required of it. As a result of incremental year-on-year changes made on the basis of the recommendations derived from this judgement, examination outcomes over time have appeared to increase in a manner which is hard to defend. Reliance on

statistics alone however would open the system up to charges that the process of maintaining standards is a statistical illusion.

Baird, adapting Fawcett's (2005) criteria for evaluation of theories, suggests that a definition of examination standards should ideally meet the following criteria:

1. It should have a theoretical underpinning, referring explicitly to the educational intentions of standards and comparability. The theory should be consistent, as opposed to predicting more than one outcome for any particular case.
2. The definition should be testable and supported by evidence.
3. As with any good theory, the definition should be parsimonious.
4. The definition should be practically useful in our educational culture.

According to Baird (2007), all of the definitions that have been adopted in England have fallen short of these criteria. Cresswell (2010) is less concerned with the theory and has set out a requirement for an agreed measure between awarding bodies for the comparability of standards which is credible and objective. Having shown that both statistical models and judgement are subjective measures it is time to consider whether Item Response Theory could contribute to providing an objective comparison of actual performance standards.

2. Item Response Theory and Test Equating

2.1 Overview

This chapter considers a range of test theory models that may be appropriate for use with public examinations in the UK. It then examines why an equating design is required, the various equating designs available and how they are used in the US and the Netherlands. Finally it considers the equating methods that are available. It concludes that an IRT approach theoretically holds great promise, but that IRT approaches must be tested in the practical context of public examinations in the UK to determine how useful they are.

2.2 The Rasch model and Item Response Theory

The essential problem in grading is deciding the mark that candidates would have gained on previous versions of the same test. This is essentially a problem for test theory which was developed in the 1960s in order to assess performance and achievement across groups in which not all persons had responded to all items. Under this frame of reference, this year's test could be considered a subset of items from a larger pool. The application of test theory, however, is complicated by the existence of two paradigms: the traditional approach represented by Item Response Theory and the models of Lord and Novick (1968) and Birnbaum (1968); and the Rasch model (Rasch, 1960).

2.2.1 The Rasch paradigm

The Rasch model is derived from the epistemic principle that comparisons between objects of interest should be carried out independently of the set of agents which are instrumental for the comparisons, and vice versa (Fischer, 2007). In test theory this means that the comparison between any two persons should be independent of the items on the basis of which this comparison is made (Andrich, 2004). Rasch termed this principle Specific Objectivity (Rasch, 1977). This insight occurred to Rasch following a discussion with Ragnar Frisch, a Norwegian economist and later Nobel Prize winner in which he described how the person parameter had fallen out of one of his mathematical derivations,

> ... Until this point Frisch had only listened politely... on seeing [the elimination of parameters] Frisch opened his eyes widely and exclaimed: " it... was eliminated, that is most interesting!" And this he repeated several times during our further conversation. To which I of course agreed every time - while I continued reporting the main results of the investigation and some of my other work.
>
> Only some days later I all of a sudden realised what in my exposition had caused this reaction from Ragnar Frisch. ...
>
> What Frisch's astonishment had done was to point out to me that the possibility of separating two sets of parameters must be a fundamental property of a very important class of models. (cited in Andrich, 2004, p. 149)

Specific Objectivity is more commonly known as invariance. For dichotomously scored items, the Rasch model resulting from the condition of invariance is:

$$\Pr[correct] = \frac{e^{\beta_n - \delta_n}}{1 + e^{\beta_n - \delta_n}} \tag{2.1}$$

As there is no information in a person's pattern of responses their total score is sufficient for the parameter *beta*. This condition of sufficiency means that equation (2.1) can be rewritten as follows:

Pr {response to i is positive and to j negative, given only one is positive} =

$$\frac{e^{\delta_i - \delta_j}}{1 + e^{\delta_i - \delta_j}} \tag{2.2}$$

which does not contain the person parameter, and characterises comparisons of items which are invariant relative to the locations of persons (Andrich, 2004).

It is the realisation of the concept of sufficiency that Rasch felt was his substantial contribution to the theory of knowledge (Andrich, 2004). Sufficiency is not merely a nice concept, it allows axiomatic, fundamental, measurement, known as additive conjoint measurement compatible with the laws of physics (Andrich, 2004). Sufficiency is not easily realised, however. The Rasch model is extremely restrictive. It requires, among other things, unidimensionality, no guessing, items with the same discrimination and items that perform consistently with respect to variables such as gender, age and education (Fischer, 2007).

2.2.2 The traditional paradigm

In the traditional paradigm a model is chosen to account for data which are given (Andrich, 2004). If the model does not fit more parameters may be added or the

model discarded. If the Rasch model does not fit the data, therefore, then a more general model is fitted:

$$\Pr[correct] = \frac{e^{\alpha_i(\beta_n - \delta_i)}}{1 + e^{\alpha_i(\beta_n - \delta_i)}} \tag{2.3}$$

where *alpha* characterises the discrimination of item *i*. This is known as a two-parameter (2-pl) IRT model. While it may seem reasonable to model the discrimination of items discretely, the loss of sufficiency means that, under Item Response Theory models, certain assumptions regarding person parameters have to be made (Bock & Moustaki, 2007). Further parameters can be added to deal with other restrictions on the Rasch model. Parameters can be added to relax the requirement for conditional independence of responses to items (Wainer, Bradlow & Wang, 2007); for guessing (Lord & Novick, 1968); and for multiple dimensions (Reckase, 1985). In all cases a candidate's summed score is no longer a sufficient statistic.

2.2.3 The Rasch controversy

The requirement for all items to share the same discrimination parameter has led to a great deal of misunderstanding of the purpose of the Rasch model,

> ... the Rasch model... includes only one free item parameter, that for difficulty. ... items that fit a one-parameter model all have the same discrimination parameter ...

> These assumptions about items fly in the face of common sense and a wealth of empirical evidence accumulated over the last 80 years. (Traub, 1983, p. 64)

The Rasch model is not, however, an omnibus method for the analysis and scoring of all sorts of tests. Rather, it is a guideline for the construction or improvement of tests; an ideal to which a test should be gradually approximated, so that measurement can profit from the unique properties of the Rasch model (Fischer, 2007). Having derived his model Rasch found that the second test he attempted to fit to his model showed substantial misfit. Rather than adapting the model, however, he sought to understand the cause of the misfit. He discovered that the increased discrimination of items towards the end of the test which were causing the misfit was due to an unintended speeded dimension to the test. Once this speeded dimension had been removed, he found that the model showed satisfactory fit (Andrich, 2004).

2.2.4 Generalisations from the Rasch model

2.2.4.1 The Partial Credit Model

Most assessments delivered at GCSE and A-level require structured answers. 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.

2.2.4.2 OPLM

The 'One Parameter Logistic Model' (OPLM) relaxes the assumption of equal discrimination between items as it posits that each item belongs to one of a few

classes of items with different discrete rational discrimination parameters. Each item is first assigned to one of the classes by means of a heuristic procedure, then its discrimination is considered as fixed and given (Verhelst & Glas, 1995). Under the 2-pl model item discrimination parameters are free, while in OPLM they are fixed a priori. This allows items with different discrimination values to be modelled while preserving the mathematical and theoretical advantages of the Rasch model derived from the use of total score as the sufficient statistic for ability (Verhelst & Glas, 1995).

2.3 Test equating

2.3.1 Indeterminacy

Rasch and Item Response Theory scales have a location indeterminacy which depends on where the zero point is set. Usually magnitudes of item difficulties are reported relative to the mean calibration of a particular set of items. In order to place two items from two tests on the same scale of difficulty or two cohorts on the same scale of ability a test equating design needs to be in place.

2.3.2 Criteria for equating

Equating, according to the 'crude and intuitive theory of test equating' (Holland, Dorans & Petersen, 2007, p. 173) requires the following:

a) The equal construct requirement. The two tests should both be measures of the same construct (latent trait, skill, ability).

b) The equal reliability requirement. The two tests should have the same reliability.

c) The symmetry requirement. The equating transformation for mapping the scores of Y to those of X should be the inverse of the equating transformation for mapping the scores of X to those of Y.

d) The equity requirement. It should be a matter of indifference to an examinee to be tested by either of the tests that have been equated.

e) The population invariance requirement. The equating function used to link the scores of X and Y should be the same regardless of the choice of (sub) population from which it is derived.

In practice, a) and b) mean that the tests need to be built to the same content and statistical specifications. Requirement c) is a technical concern that, for example, excludes regression as a possible equating technique. Requirement d) is primarily theoretical and hard to evaluate empirically. Lord (1980) shows that property d) only holds if Form X and Form Y are perfectly reliable or Form X and Form Y are strictly parallel. Requirement e) may be just as unattainable in practice (Livingston, 2004) but it is easier to test empirically (Holland et al., 2007). Quantitative measures can be developed that indicate the degree to which equating functions depend on the subpopulations used to estimate them (Dorans & Holland, 2000).

2.3.3 Equating designs

Béguin (2000) distinguishes between two classes of test equating design (data collection procedures). The first class contains designs with a single group or randomly equivalent groups. The assumption of randomly equivalent groups entails

that the groups in the design are drawn from the same population and, as a consequence, they have the same statistical properties. The second class contains designs for which the assumption of randomly equivalent groups may not hold. In these non-equivalent groups the respondents are assumed to be drawn from different populations. It is this latter design which is useful for equating examinations since their performance standard may change from year to year.

For non-equivalent groups to be equated some proportion of the candidates in each group must have taken some proportion of items in common. Figure 2.1 illustrates the anchor test non-equivalent groups design. In addition to their designated test form each group takes the same linking test which is referred to as an 'anchor test'. The anchor test can be internal, in which case the scores on the test contribute to the candidates' overall score or the anchor test can be external, in which case the scores do not contribute to the candidates' overall score.

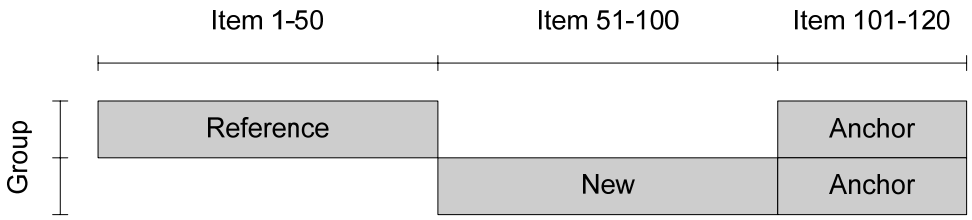


Figure 2.1: Anchor-test non-equivalent groups design (Béguin, 2000, p. 7)

Figure 2.2 illustrates a pre-equating non-equivalent groups design in which the reference form of the test is administered together with subsets of new items. Not all of the new items need to be selected for the new form nor do all items of the new form have to be administered with the reference test. The new items are administered in such a way that the test takers cannot distinguish between the pre-test items and

the items of the actual examination. Again, candidates may or may not be given credit for their answers to the new items.

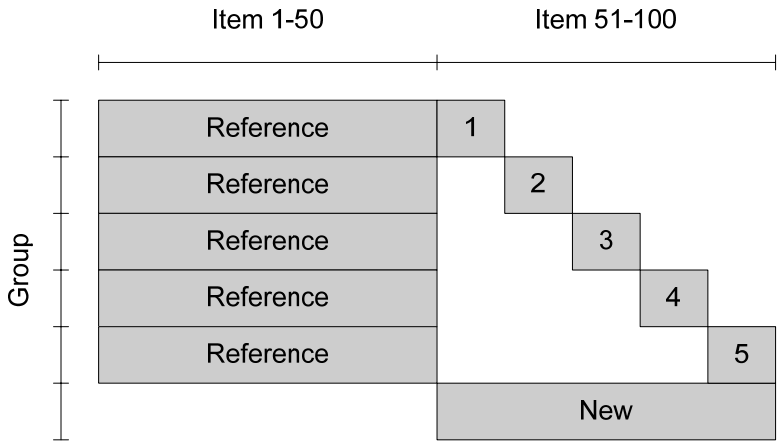


Figure 2.2: Pre-equating non-equivalent groups design (Béguin, 2000, p. 8)

Figure 2.3 illustrates the post-equating non-equivalent groups design in which the reference and new form are presented simultaneously to different linking groups. In pre-equating, the new form is administered before it is to be used operationally, while in post-equating, the new form may have been administered operationally before equating data has been collected. The post-equating design is, therefore, the most secure design as there is no risk of operational items being leaked through the equating design. The major challenge of the post-equating design is to find a suitably motivated cohort who will undertake the equating test forms.

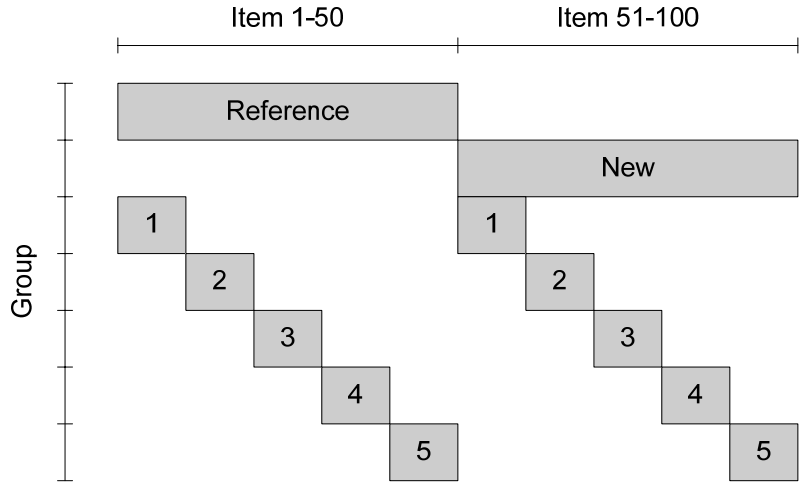


Figure 2.3: Post-equating non-equivalent groups design (Béguin, 2000, p. 9)

2.3.4 Practical issues for test equating designs

In the context of high-stakes testing test security is paramount. In order to disguise items that are being used for equating of pre-testing they are often embedded within sections of scored operational items. This makes the item parameters susceptible to change due to context effects. If they are placed in different positions in the test their relative proximity to other items or the start or end of the test may affect their difficulty or discrimination. There is also pressure to minimise the number of new items that any candidate is exposed to. This may make anchor portions relatively unreliable. As noted above only a proportion of the new items on a test form need to be pre-tested, and security concerns would dictate that this proportion is as low as possible. Without pre-testing of the items, however, problems in their content or level of difficulty cannot be detected and the resulting dataset may be less amenable to equating. Finally, unrepresentative or unmotivated samples undermine equating. Special study data collections may minimise security concerns but appropriate incentives for performance need to be in place.

2.3.5 Some case studies in test equating

2.3.5.1 The National Assessment of Educational Progress (NAEP)

The most extensive application of Item Response Theory at present in the US is the National Assessment of Educational Progress (NAEP), a nationwide survey of educational outcomes reporting to the States and the general public. During each assessment year a probability sample of schools is drawn in the participating states; within each school, a random sample of students is then drawn at each of several grade levels. In group testing sessions, each of the students is randomly assigned a form of a multiform assessment instrument appropriate to his or her grade level. The instrument is constructed by assigning items covering several areas of the curriculum to forms in a balanced incomplete block design in which each area of all forms share an equal number of items in common. There are too few items in each form to support reliable reporting of achievement scores of individual students; rather, the purpose is to give dependable estimates of average achievement at the state and national level. This method of obtaining item data from individual respondents in more than one area of item content to assess outcomes at the group level is called multiple matrix sampling (Bock & Moustaki, 2007). The assessment forms also include a certain number of items held over from previous assessment years so that the results can be expressed on the same scale from year to year (Bock & Moustaki, 2007).

2.3.5.2 The National Assessment of Educational Progress (NAEP) anomaly

The cornerstone of IRT 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 NAEP anomaly in 1986, however, that view was substantially revised (Beaton & Zwick, 1990).

Designed to measure changes over time the NEAP suffers from the tension between keeping its content relevant while following the well-rehearsed maxim that to measure change you should not change the measure. To compensate for changes in the measure deemed necessary to keep content relevant, an IRT test equating design was used. An anchor was constructed that was repeated over time, but 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 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 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. It then becomes the key responsibility of the test agency to monitor the performance of items that are re-used over time for evidence of drift in any of their key parameters.

2.3.5.3 College admissions testing in the US

Two batteries of tests are used for large-scale college admissions testing in the US: the SAT and the ACT. The SAT consists of a number of sections intended to measure developed verbal and mathematical reasoning skills as well as critical reading and writing. Item types are predominantly multiple-choice. The ACT includes a battery of four mandatory multiple-choice tests: English, Mathematics, Reading, and Science. These tests are intended to measure a student's readiness for college and the extent to which a student is prepared to profit from college experience (Schmeiser, 2004).

The following account of the equating procedures used comes from Liu, Harris and Schmidt (2007). The SAT programme uses two types of equating design:

the non-equivalent groups anchor test design and the random/equivalent groups design. New forms are introduced in pairs. The first new form is equated through the NEAT design, while the second new form is equated to the first one through an EG design. These two forms are distributed randomly (spiralled) amongst candidates to ensure equivalent groups in the same administration.

The ACT equating design uses a carefully selected sample of examinees from one of the national test dates and administers a spiralled set of forms to that sample. One of the forms, an anchor form, has already been equated, and serves as the link to their scaled score. The use of randomly equivalent groups allows the use of the relatively simple equipercentile equating methodology. In equipercentile equating, a score on form X of a test and a score on form Y are considered to be equivalent if they have the same percentile rank in a given group of examinees.

Two types of pre-testing are employed by the SAT and the ACT to facilitate the equating designs. Pre-test items can be embedded within operational sections so that examinees are not sure if they are responding to an operational item or a pre-test item. This method provides optimal item statistics for the pre-test items; however the time and energy that the pre-test items require may impact on the examinee's score. Items are not usually appended to the end of a test where they may have less impact on a candidate's scores as fatigue and lack of time may affect the item statistics. The pre-test items do not contribute to the examinees' scores.

Pre-testing may also occur in a separately timed section of the test packaged along with the operational test and given at the same administration. The SAT test book consists of operational sections containing operational items, and a variable section containing either equating or pre-test items. The variable section can appear in any of the sections of the tests, so that test takers do not know which section

contains pre-test items (or equating items) and which sections contain operational items. Use of a separate section ensures the item data is gained from a representative and motivated test taking population working under realistic conditions. Operationally, it is also a very inexpensive way to collect data.

2.3.5.4 Equating in the Netherlands

The following account is extracted from Alberts (2001). The concerns in the Dutch assessment system are similar to those in England. Candidates progressing to university or other forms of tertiary education take a profile of subjects that constitute a diploma. In every subject a single pass/fail cut-off score is set each year to ensure that performance in different years is equally valued and that examination outcomes remain relatively stable year on year regardless of the difficulty of the tests that have been set. Until 1994 these cut-off scores were set to ensure that the same percentage of candidates would pass each subject every year. Concerns over the equivalence of performance standards, however, using this equipercentile approach, led to a series of experimental equatings which revealed that these performance standards were not being maintained.

The post-equating design was very similar to Figure 2.3 above. As there are four different streams of education in the Netherlands, each with its own teaching programme to match the ability and work pace of the pupils, a cohort from a different stream was chosen for the post-equating. To minimize security concerns the equating took place after the examinations had been administered. New cut-off scores were determined and presented to the committees responsible for the examinations. From discussions with these committees, it appeared that they found it difficult to understand how the equivalent cut-off scores were arrived at. They were

also very reluctant to accept the use of within-group comparisons carried out using non-equivalent groups.

A further study followed, and in the face of continuing resistance from the committees responsible for the examinations, the equatings were replicated in three different ways. A first replication made use of pupils from the educational type the exam was meant for, instead of a non-equivalent group, one month before the exam. A second replication re-analysed the data using a different equating method. In a third replication 1000 pupils from the authentic examination population were added to the design. The results revealed that the equatings were indeed robust. The author concludes,

> The results suggest, rather, a tendency in standard setting to use the score distribution and set the cut-off score at an acceptable percentage of passes, without taking into consideration the possibility that the whole population might be performing better or worse than before. The Cito researchers considered that the equating procedure allowed estimation of the difficulty of a particular exam independent of the performance level of the population that took it in the year concerned (Alberts, 2001, p. 364)

As a result, in 1994, a post-equating non-equivalent group design was introduced to operational standard setting procedures for 11 secondary school examinations.

2.3.6 Linking designs

In order to maintain standards over a succession of testing sessions 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 practical considerations such as re-take candidates.

Figure 2.4 illustrates a linkage plan that initially ties three sessions to an established standard. There is no attempt to link tests taken at different times of year (rule 2) and each new link carries forward the new standard (rule 4). It does mean that the standards of the test sessions within any year could slowly drift apart. Figure 2.5 attempts to address this problem by introducing a double link: one link to the previous test one year prior and one to the original reference test. Double links, however, could prove logistically complex.

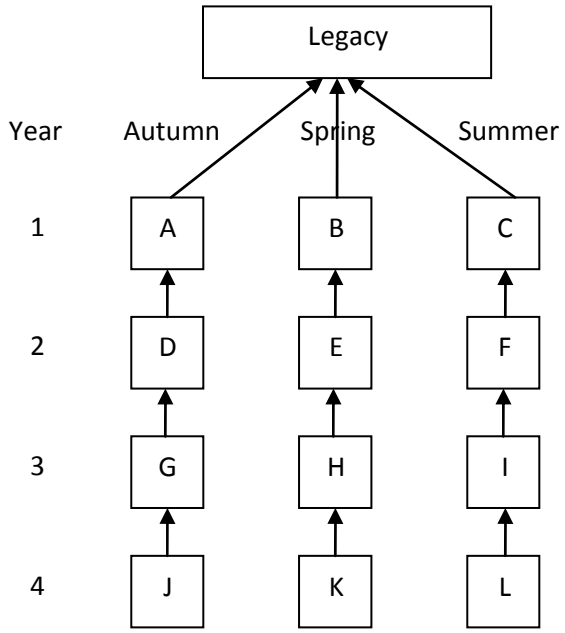


Figure 2.4: A single link plan

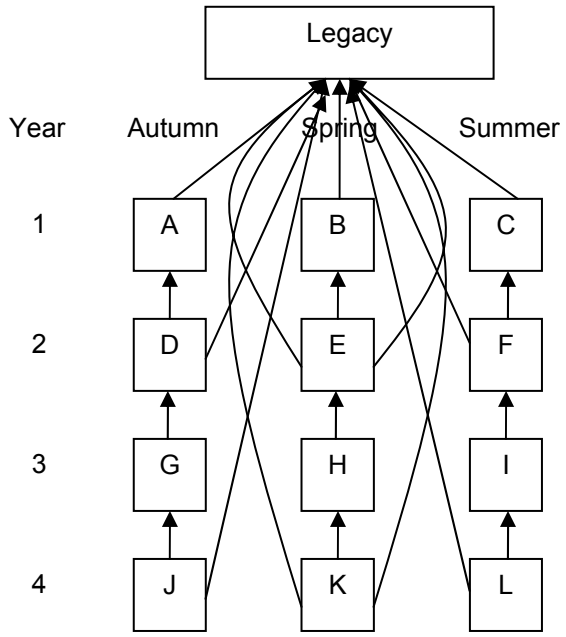


Figure 2.5: A double linking design

2.3.7 Item Response Theory equating methods

Once data has been collected the technical process of equating tests can be undertaken. A distinction can be made between two types of equating methods based on IRT models. In IRT true score equating (Lord, 1980), equivalence is directly based on the latent ability scale. In IRT observed score equating (Lord & Wingersky, 1984) and IRT observed score equating of number-correct scores (Zeng & Kolen, 1995) equivalence is defined in terms of properties of observed scores.

2.3.7.1 IRT true score equating

In IRT true score equating, equivalence is obtained through the expected number correct score under the IRT model, which is also called the number correct true score (Kolen & Brennan, 2004) or true score (Lord & Wingersky, 1984). The true score on a test form is defined as the sum of probabilities of correct answers under the IRT model at given levels of ability. Scores on different forms are considered equivalent if they are associated with the same ability.

IRT number correct score equating consists of three steps:

1. Estimation of the parameters of the IRT model
2. Estimation of the distribution of scores on the form each group were not administered using the parameters of the IRT model
3. Equipercentile equating between the distribution of scores on the form each group were administered and the estimated distribution of scores on the form each group were not administered

Stage one, the estimation of the parameters of the IRT model, can be achieved through separate or concurrent estimation procedures. In separate estimation procedures, the parameters of the IRT models are estimated separately for each form. Using these estimates, the forms are brought onto a common scale via minimisation of some loss function (Béguin, 2000). In concurrent estimation the parameters of the IRT model are directly estimated on a common scale via maximum likelihood. It has been shown that the marginal maximum likelihood (MML) procedure is strongly consistent under fairly reasonable regularity conditions (Keifer & Wolfowitz, 1956). Therefore, standard asymptotic theory on confidence intervals and the distribution of statistics computed using MML estimates directly applies. Little is known of the theoretical properties of separate estimation (Hanson & Béguin, 1999).

2.3.7.2 Rasch equating

Bearing in mind the theoretical limitation noted above, the simplest method of estimating the parameters of the IRT model is to use Rasch equating (Wright & Stone, 1979). In Rasch equating, the separate calibrations of each test produce a pair of independent item difficulties for each linking item. The equating model asserts that each pair of estimates are statistically equivalent except for a single constant of translation common to all pairs in the link. If two tests, A and B, are joined by a common link of K items, then the constant of translation is:

$$Const = \frac{1}{K}\sum_{i=1}^{K}(LinkA_i - LinkB_i) \tag{2.4}$$

where LinkA represents the item difficulty for item *i* in the context of test A and LinkB represents the item difficulty of item i in the context of test B. Then the parameter estimates from test B can be transformed to the scale of test A:

$$TestB' = TestB + Const \tag{2.5}$$

while measures that are common to both tests are represented by:

$$Link = [LinkA + (LinkB + Const)] / 2 \tag{2.6}$$

Before items are placed on a common scale, however, the relationship between the parameter estimates of link items in *LinkA* and *LinkB* should be examined. If the relationship is close to identity then the invariance of the item parameters across the test forms holds. Where this does not hold link items may have to be removed. Rasch equating is the simplest form of equating but has the built-in rigid assumption of equal discrimination between items which may not hold in practice.

2.3.7.3 OPLM equating

The OPLM equating procedure is built on the concept of booklets. Each booklet contains items which may be shared across other booklets. These common items allow scores on the different booklets to be compared. CML estimation is then used to calibrate the difficulty and discrimination parameters across all booklets concurrently. Once the item parameters have been estimated they can then be fixed so that the population parameters can be assessed using the marginal maximum likelihood (MML) method. These population parameters can in turn be used to establish expected score and estimated latent ability distributions.

2.3.8 Evaluating the quality of equating

Equating quality is determined largely by the quality of the tests (including anchors) that are to be equated. Concepts from Classical Test Theory such as item test

correlations, item facilities and item discriminations are important measures of this quality (Holland et al., 2007). When an Item Response Theory model is fitted further checks need to be made on the fit of the model and violations of assumptions such as the conditional independence of the responses and unidimensionality (Kolen & Brennan, 2004). Any violation of these assumptions does not necessarily, however, rule out the use of an IRT model. Béguin (2000), for example, found that the equating procedure based on the Rasch model used in the Netherlands for examinations of language comprehension were robust against violations of unidimensionality and guessing.

Returning to equating theory, it is also important to evaluate how well the tests to be equated operate together. The need for an equal construct between the tests can be evaluated through correlations, differential item functioning statistics and measures of overall model fit (Holland et al., 2007). The reliability of the tests should be compared. The symmetry of the equating function is most easily checked within the Rasch equating through a line of best fit that is drawn through the calibrations of item parameters within each test form (Wright & Stone, 1979). An identity line means the equating transformation for mapping the scores of Y to those of X is the inverse of the equating transformation for mapping the scores of X to those of Y. The further the line departs from this idea the further the symmetry of the equating function will be degraded. Finally the invariance of the equating function to the representation of subgroups can be calculated (Dorans & Holland, 2000).

2.4 Discussion

This chapter has examined the potential of IRT in delivering an objective measure of performance standards for the public assessment system in England. A consideration of the models available suggests that only the Rasch model aspires to provide an objective measure of performance. This aspiration is supported by the principle of specific objectivity: that comparisons between objects of interest should be carried out independently of the set of agents which are instrumental for the comparisons. To achieve this specific objectivity the Rasch model removes the person parameter in comparisons of performance on test items. As a result the ability of persons on a latent trait scale can be compared regardless of the selection of items against which this ability was measured. This would seem to be a powerful argument for the use of the Rasch model in attempting to derive an objective measure of performance. However, the Rasch model is an idealisation, a template for test construction, which is unlikely to hold in practice.

If the Rasch model shows poor fit to the data then the data can be improved or an alternative, more traditional, paradigm can be adopted. Under this paradigm the model is adapted until it shows reasonable fit and delivers successful predictions. Claims can no longer be made, however, for its objectivity. Subjective decisions need to be taken regarding which parameters will be added to the model. Relaxation of the requirement for equal discrimination between items seems an obvious first step; however there are many other parameters that could be considered. Relaxation of the requirement for conditional independence of responses to items may seem equally necessary where tests are organised by theme, and those themes are introduced by specific stimuli.

Once a model has been fitted to the data, its fit and observance of the assumptions of the model need to be examined. Most work within test theory has taken place on unidimensional IRT models. The extent to which the test data support the assumption of unidimensionality needs to be examined. Violations of the assumption of unidimensionality do not necessarily mean, however, that the test equating will not yield accurate predictions.

Apart from the statistical concerns of fitting models to data or data to models IRT test equating requires a test equating design to be in place. These designs need to ensure that high quality, reliable and representative samples of test data are collected. In the US test equating designs are built into the operational procedures for the major college entrance examinations. This ensures that the candidates taking equating items are highly motivated and representative of the population and the process is operationally very efficient. The designs, however, raise significant ethical issues as they may interfere with the performance of candidates on their live tests. There are also security concerns related to the pre-testing of live items under these designs. Further, they may only be feasible for multiple-choice style tests delivered to large, relatively stable populations.

In the Netherlands the ethical and security concerns involved with test equating designs were solved through equating after the live examinations have taken place. While initial studies showed the results of this design to be robust there was substantial resistance to using statistical inferences that had been drawn from a different population. Nevertheless, the strong suggestion is that this post-equating design has delivered an objective measure of performance standards.

Theoretically, an IRT approach seems to hold up against Baird's (2007) criteria for the definition of examination standards. It has a theoretical underpinning

explicitly related to the educational intentions of standards and comparability if that is understood as ranking candidates consistently regardless of the specifics of their assessment session (Lord & Novick, 1968). IRT models are testable and supported by evidence as the model may or may not predict the data (Hambleton, Swaminathan & Rogers, 1991). The definition of IRT is parsimonious as it consists of two postulates: the underlying performance of an examinee on a test item can be predicted by a set of factors called traits, latent traits, or abilities; and the relationship between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function (Hambleton et al., 1991). The final criteria, however, is whether it can practically useful in an educational culture. This is, as yet, untested. The first step in testing the usefulness of IRT is to examine how the data fit the models.

3. Model Fit in a Frequentist Framework

3.1 Overview

From a theoretical perspective IRT appears to offer a framework within which an objective measure of performance standards can be generated. By extracting ability from the specific set of items that examinees have taken it solves the problem, which is currently poorly addressed by expert judgement, of how to adjust test scores for the difficulty of each individual test. In order to implement IRT models, however, and for the inferences derived from these models not to be misleading, some strong mathematical and statistical assumptions must be met. The purpose of this chapter is to examine the methods of assessing how well those mathematical and statistical assumptions are met and to apply those methods to response data from a number of tests.

3.2 Two paradigms of model fit

The study of model fit is complicated once again by the existence of two paradigms: the Rasch paradigm and the IRT paradigm. Under the Rasch paradigm the model is given so the purpose of fit analysis is to consider how the quality of the data can be improved. In building a valid test, for example, the question could be asked whether all the responses stimulated by the test items create a coherent description of the latent variable. Under the IRT paradigm the data is given so the purpose of fit analysis is to consider whether any model fits well enough or whether more complex models need to be explored.

The different paradigms give rise to different emphases in tests of model fit. The Rasch paradigm prioritises descriptive and diagnostic item level statistics over global statistical tests,

> The Rasch model is an idealization, never achieved by real data. Accordingly, given enough data, we expect to see statistically significant misfit to the model. If the current data do not misfit, we merely have to collect more data, and they will! In essence, the null hypothesis of this significance test is the wrong one! We learn nothing from testing the hypothesis, "Do the data fit the model (perfectly)?" Or, as usually expressed in social science, "Does the model fit the data (perfectly)?" Perfection is never obtained in empirical data. What we really want to test is the hypothesis "Do the data fit the model usefully?" And, if not, where is the misfit, and what is it? Is it big enough in size (not "statistical significance") to cause trouble? This is the approach used in much of industrial quality-control, and also in Winsteps (Linacre, 2008, p. 402).

The IRT paradigm prioritises statistical tests of global hypotheses that indicate whether more complex models are needed,

> As can be observed from the table, model fit is unsatisfactory. A normal way to proceed would be to fit other models… (van Rijn, Verstralen & Béguin, 2009, pp. 11-12)

Under the IRT paradigm measures of item fit are primarily required because the items may be re-organised into different test forms.

3.3 Assessing model fit

The following two steps are recommended for assessing model fit:

(i) Checking the underlying assumptions such as unidimensionality

(ii) Assessing the agreement between observations and model predictions

There are a large number of potential tests which check the underlying assumptions of the model and the agreement between observations and model predictions so only a small subset will be explored here. A comprehensive description of the tests available can be found in Swaminathan, Hambleton, and Rogers (2007).

3.4 Checking the underlying assumptions of unidimensionality

3.4.1 Linear factor analysis

While multidimensional IRT models have been developed their use is not yet operational. For this reason, testing whether the complete latent space is unidimensional is critical. Under the IRT paradigm a popular approach is the use of linear factor analysis. In this approach,

(i) the matrix of inter-item correlations is obtained

(ii) the percent of variance explained by the largest eigenvalue along with the point where a break occurs in the plot of the eigenvalues, or the scree plot, are examined;

(iii) based on the above considerations a determination is made regarding the dimensionality of the item response data.

This approach has several drawbacks. When the item responses are discrete, the inter-item correlations will be small. The discrete item responses may have non-linear relationships with the underlying ability continuum. Simulation studies have shown that in unidimensional Item Response Theory applications, the largest eigenvalue of the matrix of tetrachoric correlations will typically account for only about 25 to 35 per cent of the variance. Drasgow and Lissak (1983) propose examining the latent dimensionality of dichotomously scored item responses through the second eigenvalue of the tetrachoric correlations matrix of the dichotomous items. This has been implemented in R (R Development Core Team, 2010) by Rizopoulos (2006) through a Monte Carlo procedure which is used to approximate the distribution of the second eigenvalue statistic under the null hypothesis (the IRT model).

3.4.2 Rasch Principal Components Analysis of residuals

If all the data is explained by the Rasch model then the residuals would be random noise, independent of each other. Principal Components Analysis of the standardised residuals identifies characteristics in items which could indicate secondary structures or sub-dimensions within the data (Linacre, 2008). Principal Components Analysis is

not effective, however, if there are two dimensions with an equal number of items, and these are interlaced in difficulty (Tennant & Pallant, 2006).

3.5 Assessing the agreement between observations and model predictions

The second stage in checking model fit is to check the model predictions. Verification of a theory is most directly carried out by examining the predictions made by the theory. In Item Response Theory model predictions can be compared with the observed data. These comparisons can be made at the test or at the item level.

3.5.1 Assessment of model fit at the test level

3.5.1.1 The R0 test

OPLM and IRT models make specific assumptions regarding the distribution of ability as part of MML estimation. These assumptions can be tested using the R0-test (Verhelst & Glas, 1995). In the MML framework, the theoretical distribution of scores is a function of both the item parameters and the parameters of the ability distribution, which have to be estimated from the data, and the observed frequency distribution of the respondents' sum scores will in general not match the predicted frequency distribution. The R0 test measures this deviation.

Since a person's sum score is a sufficient statistic for the ability parameter, the statistic is based on evaluating the difference between the observed and expected score distribution given the MML estimates of the item and population parameters. Let the random variable N_{sb}, with realisations n_{sb}, denote the count of score s in

booklet *b* (in OPLM a booklet refers to a set of items taken by a specific population). The test statistic R0 is based on the differences

$$d_{0sb} = n_{sb} - E\left(N_{sb}|\beta, \mu_{p(b)}, \sigma_{p(b)}\right) \quad (3.1)$$

for b=1,...,*B* and all possible scores *s*, where the ability distribution is normally distributed with mean μ_p and standard deviation σ_p , items have difficulty β and E represents the expected a posteriori (EAP) estimation. For diagnostic purposes, these differences are transformed into standardized binomial variables, scaled deviates and combined into an overall statistic across booklets. As OPLM uses weighted scores not all integers from 0 to S are b possible scores. If, for instance, all discrimination indices are even numbers, all possible scores are even. If the design has *P* normal ability distributions, R0 has an asymptotic chi-square distribution with $\Sigma_b\ S_b^*$ - *P* degrees of freedom, where S_b^* is the number of possible scores in booklet b.

A significant result for the R0 test could be due to misfitting items or incorrect assumptions about the ability distribution. If further investigations reveal the significant result is due to the former, then CML estimates of the item parameters can be passed as fixed parameters to the MML analysis (Verhelst & Glas, 1995).

3.5.1.2 The R1m test

The R0 test does not reveal whether misfit is due to incorrect assumptions concerning the ability distribution or differences in discrimination between the items or item functioning across sub-groups. In a CML framework, no assumptions concerning the ability distribution need to be made, which means that it is possible to consider global item fit separately. The persons taking a test are divided into (a maximum of four) subgroups according to their score level. The subgroups will be indexed $q = 1,...,Q$. Let the random variable $M_{ijs|b}$, with realization $m_{ijs|b}$, denote the

number of observations having score *s* and a response in category *j* of item *i* in booklet *b*, and define $d_{ijs|b}$ as the difference between this number and its MML expected value, that is,

$$d_{ijs|b} = m_{ijs|b} - E\left(M_{ijs|b}|\beta, \mu_{p(b)}, \sigma_{p(b)}\right) \tag{3.2}$$

Once again they are transformed into standardized binomial variables, scaled deviates. To construct a global test statistic, first a set of vectors d_{1bq} with elements $d_{1bq}(i,j)$ is constructed. These elements are defined by,

$$d_{1bq}(i,j) = \sum_{w(s,b)\in G_{bq}} d_{ijs|b} \tag{3.3}$$

In order to ensure an asymptotic chi-square distributed statistic, the differences d_{0sb} between observed and expected score frequencies as defined above must be taken into account. Therefore a second set of vectors d_{0bq}, is defined as,

$$d_{0bq} = \left(d_{os_1b}, \ldots, d_{os_Qb}\right), (s_1, \ldots, s_Q \in G_{bq}) \tag{3.4}$$

The set of vectors d_{bq} is then defined as

$$d'_{bq} = \left(d'_{obq}, d'_{1bq}\right), (q = 1, \ldots, Q_b; b = 1, \ldots, B) \tag{3.5}$$

Using this definition, it can be shown (Verhelst & Glas, 1995) that the statistic

$$R_{1m} = \sum_b \sum_q^{Q_b} d'_{bq} W_{bq} d_{bq} \tag{3.6}$$

where W_{bq} is the estimated asymptotic covariance matrix of d_{bq}, has an asymptotic chi-square distribution with degrees of freedom given by

$$df(R_{1m}) = \sum_b \left(S_b^* - Q_b + Q_b \sum_{i\in I_b} m_i \right) - 2P - (B-1) - \sum_{i=1}^k m_i \tag{3.7}$$

and S* represents the number of possible scores.

3.5.1.3 Comparisons between the observed score distribution and the predicted score distribution under the Rasch model

Within the Rasch paradigm a more descriptive approach to comparisons between the observed score distribution and the predicted score distribution is suggested by Lord (1980) and enabled by the recursive algorithm derived by Lord and Wingersky (1984). The algorithm determines the conditional distribution of number correct scores from the probability of incorrect responses to every item for any given level of ability. Once a predicted conditional number correct score distribution has been obtained, the marginal number correct score distribution can be obtained by summing over the examinees at their estimated ability values. The observed number correct score distribution can then be compared with the model predicted number correct score distribution and the results displayed graphically (Swaminathan et al., 2007). The observed and expected distributions may also be compared statistically by employing the traditional chi-square test using the statistic:

$$\chi^2 = \sum_{i=1}^{m} \frac{(f_{oi} - f_{ei})^2}{f_{ei}} \tag{3.8}$$

where f_{oi} and f_{ei} are observed and expected frequencies. The statistic is distributed as a chi-square with m-1 degrees of freedom, where m is the number of score groups.

While there is no software available to obtain the marginal number correct score distribution the details on the Lord and Wingersky recursive algorithm given in Kolen and Brennan (2004, pp. 181-184) allowed the author to develop his own program within R (R Development Core Team, 2010). This code is archived here (https://github.com/cbwheadon/predicted_scores).

3.5.2 Assessment of model fit at the item level

3.5.2.1 Residual analysis

The most useful tool for comparing model predictions and what was actually observed at the item level are, under the Rasch paradigm, the Item Characteristic Curves (ICC) or, under the IRT paradigm, the Item Response Functions (IRF). The ICC illustrates the estimated or predicted probability of a correct response at any trait level. This probability can be compared with the proportion of correct answers achieved by examinees at that trait value. Under the Rasch model (and OPLM) the number correct score is a sufficient statistic for the trait theta so this comparison can be done directly. Under IRT models few individuals will have identical trait estimates except in the case of the one-parameter model. In this case artificial trait intervals must be constructed (Swaminathan et al., 2007).

3.5.2.1.1 Rasch mean squares

3.5.2.1.1.1 Calculation

The discrepancy between the observed and expected frequencies can be analysed using chi-square item fit statistics. According to the Rasch model, for each observation, there is an expectation and a model variance of the observation around that expectation. So,

$$z_{ni} = x_{ni} - p_{ni} \tag{3.9}$$

where z_{ni} is the residual, x_{ni} is the observed response of person n on item i, p_{ni} is the probability of a correct response of person n on item i.

As the residuals will have a different variance the residuals are typically standardised so that each score residual is divided by its standard deviation. For dichotomous responses this is equivalent to:

$$y_{ni} = \frac{(x_{ni} - p_{ni})}{(p_{ni}(1 - p_{ni}))^{1/2}} \tag{3.10}$$

As the residuals will sum to zero, they are then usually squared. Early studies revealed that these statistics were sensitive to outliers particularly on tests that have a wide range of item difficulties and person abilities. To counteract this sensitivity to outliers weighted version of the fit statistics were developed.

The unweighted Infit calculation and weighted Outfit calculations for Winsteps (Linacre, 2008) are as follows.

Two observations: Model p=0.5, observed=1. Model p=0.25, observed =1.

$$Outfit = \frac{\frac{(1-0.5)^2}{0.5 \times 0.5} + \frac{(1-0.25)^2}{0.25 \times 0.75}}{2} = 2 \tag{3.11}$$

$$Infit = \frac{(1-0.5)^2 + (1-0.25)^2}{(0.5 * 0.5) + (0.25 * 0.75)} = 1.86 \tag{3.12}$$

The off-target observation has less influence on the Infit statistic.

3.5.2.1.1.2 Interpretation

The mean square statistics reported by Winsteps can be interpreted as chi-squares. As the degrees of freedom vary the chi-squares are usually divided by their degrees of freedom so general guidelines can be set. Consequently their expected value is close to 1.0. Values greater than 1.0 (underfit) indicate unmodelled noise or other sources of variance in the data which degrade measurement. Values less than 1.0 (overfit) indicate that the model predicts the data too well. Overfit is not necessarily a problem, but it reveals that summary statistics, such as reliability measures, may be inflated. This overview of fit is detailed in Figure 3.1.

Outfit	**Interpretation**
>2.0	Distorts or degrades the measurement system.
1.5 - 2.0	Unproductive for construction of measurement, but not degrading.
0.5 - 1.5	Productive for measurement.
<0.5	Less productive for measurement, but not degrading. May produce misleadingly good reliabilities and separations.

Figure 3.1: Interpretation of parameter-level mean-square fit statistics (Linacre, 2008)

The likelihood of these mean squares can be calculated and reported as t-statistics or z-statistics, with critical values set that have equal Type I error rates across a variety of conditions, including sample size (Smith, 2004). Simulation studies (Smith, Schumacker and Bush, 1998) have shown that the standardised fit indices have more consistent distributional properties in the face of varying sample size than do the mean square statistics.

As well as these general guidelines for fit, the combined use of infit and outfit can lead to quite detailed diagnosis of problems. As outfit is more sensitive to unexpected observations by persons on items that are relatively very easy or very hard for them (and vice-versa) then it is related to issues such as guessing and carelessness. High infit requires a closer look at the validity of the test as it is less easy to explain why candidates of the same ability appear to perform erratically on an item.

3.5.2.1.2 Parallel Item Response Functions: M tests

3.5.2.1.2.1 Calculation

The *M* tests of item fit (available in OPLM) are based on the rationale developed for the Rasch model by Molenaar (1983) - hence the generic name of the tests. The *M* tests are based on a comparison of the probability of a correct answer by different score groups (high, medium, low) on each item with the observed proportions. The probabilities, denoted by $\Pi_{i|s}$ are given as reasonable approximations of the item characteristic curves as are the CML derived probabilities $\hat{\Pi}_{i|s}$.

$$M_i = \sum_{s \in L} (p_{i|s} - \hat{\Pi}_{i|s}) - \sum_{s \in H} (p_{i|s} - \hat{\Pi}_{i|s}) \tag{3.13}$$

.

3.5.2.1.2.2 Interpretation

In general, the graph of $\hat{\Pi}_{i|s}$ will be an S-shaped curve whose steepness depends on the discrimination index a_i. However, if a_i has too large a value, one can expect a typical pattern of deviations of $p_{i|s}$,the observed proportions p from their predicted values. If the scores are partitioned in a low group (L), a medium group (O), and a high group (H), the statistic will tend to be positive if the discrimination index is set too high. If the discrimination index is set too low, the *M* statistic will tend to be negative, suggesting an upward adaptation. The raw index *M* is divided by a suitable function of the *i* (estimated) item parameters, such that the resulting statistic follows asymptotically the standard normal distribution. For further details, see Verhelst and Glas (1995). There are various forms of the *M* tests depending on how the scores are partitioned into score groups. Polytomous items are dichotomised in order to calculate the *M* statistic.

3.5.3 Additional considerations for polytomous items

When considering the fit of polytomous items, poor fit may be explained by disordered categories or disordered category thresholds. A higher category should imply more of the latent variable; if it does not then the category will exhibit large misfit. Apart from differing in order, categories may also differ in their probability of being observed. This results in disordered thresholds. This does not necessarily degrade measurement, but implies that a category discriminates across a very narrow range of the latent variable (Linacre, 2004b). If there are very few observations then the estimation of the category parameters can only be approximate. As the Rasch method of test equating depends on the estimation of the category parameters poor estimation of these parameters can pose technical difficulties.

3.6 Method

3.6.1 Design

Rasch and OPLM models were fitted to a selection of GCSE tests. Then, some summary statistics were calculated in order to verify that the items appeared to contribute to a coherent measurement instrument. Model fit was then investigated using the following steps:

(i) Routine analysis

Firstly, a routine examination of classical indices such as facility values (p-values), item total correlations and the distribution of scores at both test and item level.

(ii) Unidimensionality

Secondly, the dimensionality of the tests was examined using:

(a) Drasgow and Lissak's (1983) linear factor analysis approach as implemented by Rizopoulos (2006). All items were dichotomised and a maximum sample of 1,000 candidates was used to minimise processing time.

(b) Principal Components Analysis of residuals (PCAR) as implemented in Winsteps

(iii) Test level measures of fit

Then test level measures were obtained using:

(a) R0 and R1M tests (Verhelst & Glas, 1995) as implemented in OPLM

(b) Graphical comparisons between the observed score distribution and the predicted score distribution as suggested by Swaminathan et al. (2007) and implemented in R (R Development Core Team, 2010). The person parameters under the Rasch model were estimated using the MML procedure from eRm (Mair & Hatzinger, 2007). A maximum sample of 1,000 candidates was used to minimise processing time.

(iv) Item level measures of fit

Finally misfit was examined at the item level using:

(a) Standardised Infit and Outfit statistics as implemented in Winsteps

(b) M-statistics as implemented in OPLM

(c) IRFs from both Winsteps and OPLM

3.6.2 Components

Thirteen tests were selected so that a variety of item-types, response lengths, subject areas and difficulties were selected. They were also chosen with test equating in mind, so they have common items between levels and a coursework element common to both tiers that can be used for cross-validation purposes. Tests with longer response items such as essays or tests with optional items were excluded as these introduce assumptions about marking and choice that do not hold.

3.6.2.1 Science (Biology, Chemistry, Physics)

The Science tests have two primary objectives. The first is to assess candidates' knowledge and understanding of science and how science works. The second is to assess the application of their skills, knowledge and understanding of science and how science works. At foundation tier the candidates answer 5 matching items (four pieces of information matched to four stimuli) and 16 multiple choice items (with four response categories and only one correct answer). The test is divided into 9 sections, each preceded by a stimulus. The stimulus may be in the form of a graph, a table, a paragraph, or some combination of all three. At higher tier candidates answer 2 matching items and 28 multiple choice items.

3.6.2.2 Mathematics

Mathematics assesses: use and application of mathematics; number and algebra; shape, space and measures; handling data. The foundation tier has 63 and 56 items in Papers 1 and 2 respectively, both with a total mark of 100. The higher tier has 47 items and 50 items in Papers 1 and 2 respectively, both with a total mark of 100. For

all papers the items vary from single mark items through to four mark items. Very few of the items are multiple-choice.

3.6.2.3 Geography

For Geography, candidates are expected to: show knowledge of places, environments and themes at a range of scales from local to global; show understanding of some specified content; apply their knowledge and understanding in a variety of physical and human contexts; select and use a variety of skills and techniques appropriate to geographical studies and enquiry. Paper 1 comprises a series of short answer items and two structured items on the United Kingdom. The paper also includes one or more items based on a UK Ordnance Survey map. Both tiers have a maximum mark of 75, with 33 items on the foundation tier and 28 items on the higher tier. The maximum mark for an item is 6 for both tiers. Paper 2 comprises four sections. Section A comprises a series of short answer items taken from: The European Union; The Wider World; Global Issues. The remaining sections each comprise a structured item on one of those same three areas. Both tiers have a maximum mark of 120; the foundation tier has 47 items while the higher tier has 31 items. The maximum mark for an item is 6 on the foundation and 9 on the higher. No items are multiple-choice.

3.6.2.4 Mathematics Functional Skills

Mathematics Functional Skills aims to assess how well candidates demonstrate their mathematical skills in a range of contexts for a range of purposes. The items therefore embed the mathematics within authentic contexts. Paper 1 is comprised of 30 short response dichotomous items, some of which are multiple-choice.

3.7 Results

3.7.1 Classical test statistics

The alpha coefficients of the longer tests were, with one exception, above 0.8 (Table 3.1). This implies that: the tests are long enough; reasonably well-targeted; and the proportions of error variance are low relative to systematic differences in the abilities being measured. The alpha coefficients of the shorter Science tests were low, and low relative to the Mathematics Functional Skills paper of a similar length. The means of the Science tests are not much higher as a proportion of the total mark than the Mathematics Functional Skills test, but the standard deviations are relatively low. This suggests the Science tests do not discriminate as well as the Mathematics Functional Skills test. The differences in coefficient alpha on these shorter tests could be due to: poor discrimination; guessing; construct irrelevant variance; other clear dimensions than the main latent variable. As the Science tests are entirely multiple-choice the most obvious explanation is that the items are liable to guessing.

3.7.2 Classical item statistics

The classical item statistics highlight a wide range of item-test correlations and item difficulty. This suggests that a discrimination parameter in the IRT model would improve its fit. The minimum values of the item facilities for the Science tests are relatively high; this would suggest that they are indeed liable to guessing and may show poor fit to both the Rasch and the OPLM models. If this is the case then a more complex model may be preferred.

Table 3.1: Classical test statistics

Test	Level	N	Items	Max Score	Mean	SD	Grade C / Level 2 boundary	Alpha
Biology	Foundation	6902	21	36	26.06	4.74	31	0.68
	Higher	11283	30	36	24.76	5.94	23	0.79
Chemistry	Foundation	4679	21	36	19.32	6.11	25	0.69
	Higher	7935	30	36	22.58	6.25	20	0.80
Physics	Foundation	7636	21	36	22.80	5.83	29	0.71
	Higher	10412	30	36	23.05	4.98	21	0.72
Mathematics Paper 1	Foundation	10000*	63	100	53.31	19.63	70	0.94
	Higher	10000*	47	100	48.97	21.83	28	0.94
Mathematics Paper 2	Foundation	10000*	56	100	52.03	19.66	69	0.93
	Higher	10000*	50	100	51.20	20.37	30	0.93
Geography Paper 1	Foundation	1192	33	75	32.98	9.28	42	0.79
	Higher	2042	28	75	45.37	10.49	37	0.83
Geography Paper 2	Foundation	1176	47	120	50.03	14.48	67	0.87
	Higher	2044	31	120	59.80	17.32	48	0.89
Mathematics Functional Skills	Level 2	15907	30	30	19.56	6.39	19	0.89

*Samples of 10000 were taken from larger populations

Table 3.2: Classical item statistics

		Correlations			Facility	
Test	Level	Minimum item-test	Maximum item-test	Average inter-item	Minimum	Maximum
Biology	Foundation	0.15	0.54	0.11	0.12	0.98
	Higher	0.20	0.49	0.12	0.39	0.91
Chemistry	Foundation	0.24	0.54	0.11	0.22	0.82
	Higher	0.18	0.52	0.12	0.30	0.89
Physics	Foundation	0.26	0.52	0.12	0.25	0.95
	Higher	0.14	0.45	0.08	0.10	0.98
Mathematics Paper 1	Foundation	0.14	0.68	0.21	0.07	0.97
	Higher	0.23	0.68	0.26	0.08	0.91
Mathematics Paper 2	Foundation	0.23	0.68	0.20	0.03	0.98
	Higher	0.13	0.68	0.22	0.10	0.99
Geography Paper 1	Foundation	0.12	0.53	0.11	0.10	0.87
	Higher	0.10	0.63	0.14	0.23	0.95
Geography Paper 2	Foundation	0.14	0.56	0.14	0.07	0.99
	Higher	0.26	0.66	0.20	0.05	0.84
Mathematics Functional Skills	Level 2	0.20	0.62	0.21	0.25	0.96

3.7.3 Unidimensionality

A one factor model specified for the Principal Components Analysis accounted for a large proportion of variance for the Mathematics tests. Figure 3.2, for example, highlights clearly that there is one dominant factor in the higher tier paper for Mathematics Paper 1. The first factor accounts for over nine times as much of the score variability as the second factor. Variance in the test scores in the other tests is less clearly dominated by a single factor model. The low factor scores could be due, however, to a non-linear relationship between items and the underlying latent trait.

The simulations of the factor structure reveal that most of the tests have a substantial second factor that is not predicted by the Rasch model. This is particularly apparent for the Geography papers. Paper 1 on both tiers is split into two sections. The first section is a traditional test of knowledge and understanding; while the second section requires candidates to undertake practical exercises using an Ordinance Survey map. The factor analysis reveals a significant second factor that is not predicted by the Rasch model for both of the tiers. Paper 2 is split into several sections but each section is a more traditional test of knowledge and understanding. Neither tier shows a significant second factor.

A comparison between Figures 3.2 and 3.3 illustrates an interesting point regarding dimensionality. Both of the factor structures have a significant second factor that is not predicted by the Rasch model. The size of the second factor in the Mathematics test is proportionally much smaller than the size of the second factor in the Geography test. Clearly, the second factor may be more disruptive to measurement for the Geography test than for the Mathematics test.

The factor analysis generally agrees with the Rasch Principal Components Analysis of the residuals (Table 3.3). A high proportion of the variance explained by

the principal Rasch measure tends to correspond with a high proportion of the score variability being explained by a single factor model. The one exception is for Mathematics Functional Skills, the reasons for which will be explored later.

Table 3.3: Tests of Unidimensionality

Test	Level	Principal Components Analysis	Second Eigenvalue			PCAR Variance explained by measures (per cent)
			Observed	Simulated	p	
Biology	Foundation	0.28	1.08	0.84	0.10	0.47
	Higher	0.24	1.93	0.72	0.01	0.30
Chemistry	Foundation	0.22	0.81	0.54	0.01	0.36
	Higher	0.24	0.96	0.69	0.01	0.38
Physics	Foundation	0.25	1.16	0.59	0.01	0.46
	Higher	0.20	1.50	1.91	0.75	0.39
Mathematics Paper 1	Foundation	0.40	3.50	2.06	0.01	0.60
	Higher	0.43	2.11	1.28	0.01	0.57
Mathematics Paper 2	Foundation	0.39	3.53	2.84	0.02	0.53
	Higher	0.41	3.43	2.68	0.04	0.59
Geography Paper 1	Foundation	0.21	1.75	0.74	0.01	0.46
	Higher	0.26	1.40	0.75	0.01	0.52
Geography Paper 2	Foundation	0.25	2.35	2.60	0.59	0.50
	Higher	0.30	1.39	1.21	0.50	0.44
Mathematics Functional Skills	Level 2	0.41	1.42	0.84	0.03	0.35

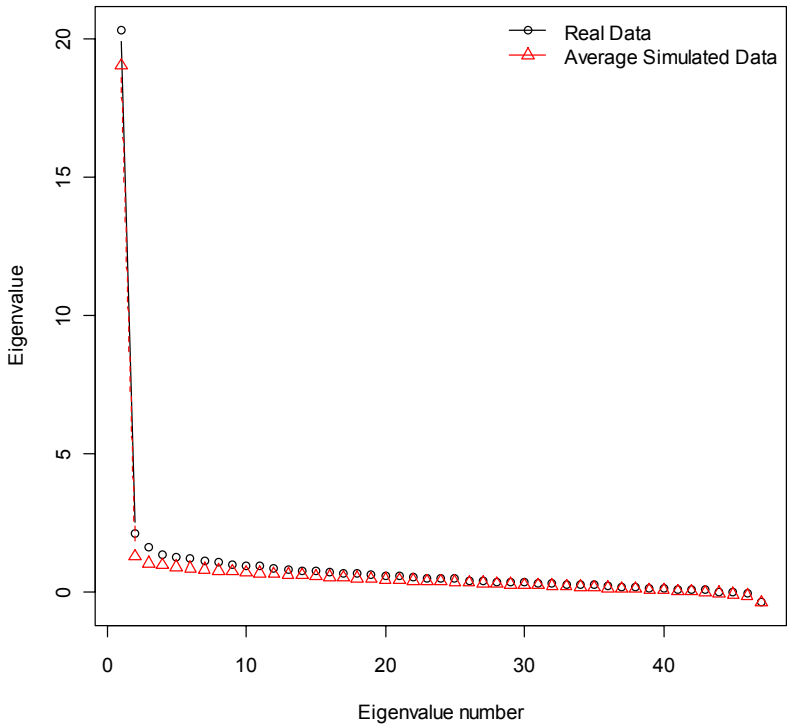


Figure 3.2: Observed and simulated factors for Mathematics Paper 1 higher tier

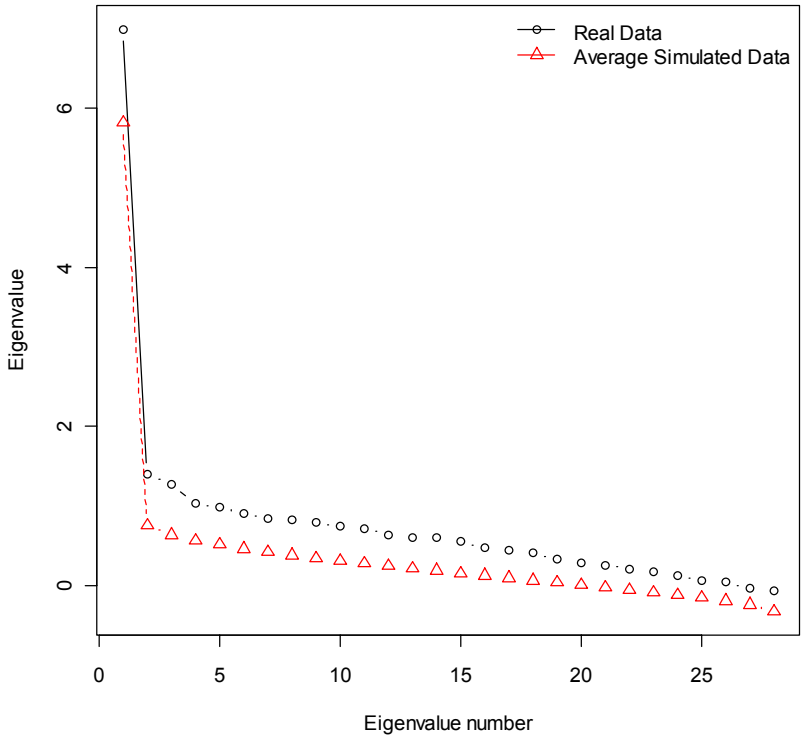


Figure 3.3: Observed and simulated factors for Geography Paper 1 higher tier

3.7.4 Test level measures of fit

As the statistics of interest reported in Table 3.4 have an asymptotic chi-square distribution, and as the sample sizes differ, Cohen's *w* (1988) was used as a measure of effect size to produce some meaningful comparisons:

$$w = \sqrt{\frac{\chi^2}{N}} \tag{3.14}$$

Cohen defines a *w* of 0.1 as a small effect size, a *w* of 0.3 as a medium effect size and a *w* of 0.5 as a large effect size. The results are reported in Table 3.4.

Comparisons between the observed number correct score distribution and the predicted number correct score distribution (according to equation 3.1) are reported in the MML columns of Table 3.4. For the shorter tests, Biology, Chemistry and Physics, the deviation appears statistically significant in 4 out of 6 cases, although the effect size is low. The results for the other tests are not statistically significant, although the effect sizes for the Geography tests appear higher than for the other tests. This could be related to the presence of a substantial second factor in the tests, as noted in the previous section, or due to the use of longer constructed response items in these tests. Generally, however, the suggestion is that the person parameters estimated under MML for the Rasch model would produce a reasonable replication of the observed distribution of scores for these tests.

Visual inspection of the difference between the observed number correct score distribution and the predicted number correct score distribution, however, revealed an interesting trend. The Biology, Chemistry and Physics papers all displayed a more acute peak in the observed score distribution than the Rasch model predicts. Figure 3.4 illustrates this deviation for the Biology foundation tier paper. There is therefore a higher probability of achieving the mean score, and a slightly lower probability of achieving scores higher than the mean, than expected under the

Rasch model. This could be due to any number of factors: guessing; violation of conditional independence of item scores; or the significant presence of a second factor that explains the variance in the test scores. Figure 3.5 shows the more successful modelling of the observed score distribution under the Rasch model for the Mathematics higher tier Paper 1.

The effect sizes derived from the R0 test correlate well (0.7) with the effect sizes derived from equation 3.1, which suggests that neither the different model used (OPLM rather than Rasch) nor the different statistic tell an entirely different story. The R0 test under OPLM does show, however, a statistically significant deviation between the observed and the expected score distributions for all but two of the tests and higher effect sizes. The Geography tests again appear to give most cause for concern.

Finally, the R1M statistics suggest that in every case the severity of the misfit invalidates the use of a person's sum score as a measure of ability. Under the IRT paradigm it would be time to investigate more complex models; under the Rasch paradigm, however, the source of this misfit is more interesting.

Table 3.4: Test level measures of fit

			Rasch					OPLM					
			JML	MML									
Test	Level	N	Items Outfit mnSq > 1.3	Chi Square	df	p	w	R0	df	w	R1M	df	w
Biology	Foundation	6902	2	63.92	35	0.00	0.10	199	103	0.17	652	248	0.31
	Higher	11283	0	38.92	35	0.30	0.06	321	97	0.17	1414	242	0.35
Chemistry	Foundation	4679	0	47.27	35	0.08	0.10	102*	81	0.15	478	221	0.32
	Higher	7935	2	34.02	35	0.52	0.07	476	97	0.24	2164	242	0.52
Physics	Foundation	7636	0	47.45	35	0.08	0.08	142	93	0.14	522	238	0.26
	Higher	10412	2	53.36	35	0.02	0.07	865	101	0.29	2548	246	0.49
Mathematics Paper 1	Foundation	10000	17	102.58	99	0.38	0.10	1413	276	0.38	11615	677	1.08
	Higher	10000	9	93.08	99	0.65	0.10	413	266	0.20	3622	667	0.60
Mathematics Paper 2	Foundation	10000	8	67.03	99	0.99	0.08	1409	253	0.38	7425	654	0.86
	Higher	10000	10	71.63	99	0.98	0.08	838	281	0.29	4243	682	0.65
Geography Paper 1	Foundation	1192	0	61.13	74	0.86	0.23	310	189	0.51	784	478	0.81
	Higher	2042	0	92.50	117	0.95	0.21	4768	206	1.53	5174	507	1.59
Geography Paper 2	Foundation	1176	2	63.03	74	0.82	0.23	664	287	0.75	1578	760	1.16
	Higher	2044	1	87.86	117	0.98	0.21	281*	321	0.37	933	798	0.68
Mathematics Functional Skills	Level 2	15907	6	28.90	29	0.47	0.04	154	83	0.10	796	199	0.22

*p>0.05

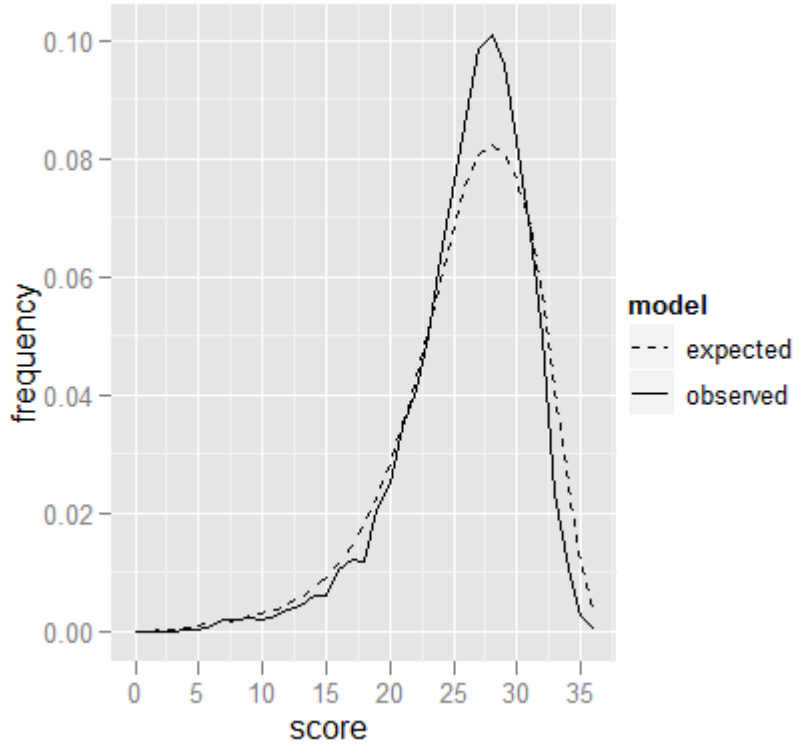


Figure 3.4: Observed and expected score distributions for Biology foundation tier based on trait-estimates for the Rasch model

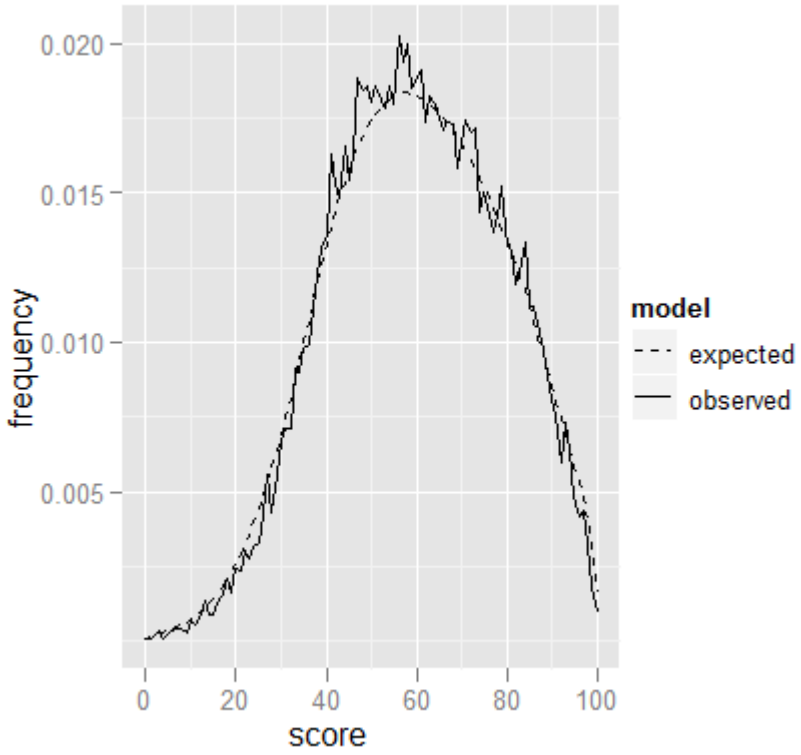


Figure 3.5: Observed and expected score distributions for Mathematics Paper 1 higher tier based on trait-estimates for the Rasch model

3.7.5 Rasch person-item maps

Poor discrimination could be the reason for the acute peak in the score distribution and the particularly low coefficient alpha for the foundation tier Biology test. Figure 3.6 shows the Rasch item-person map for this paper. The solid dots represent the item location while the hollow dots represent the Rasch-Andrich thresholds, the point on the latent ability scale at which each category has the same probability of being observed. With only one exception the item difficulty is lower than the modal person ability. This will result in poor discrimination amongst the higher performing candidates and reduce the information available at higher levels.

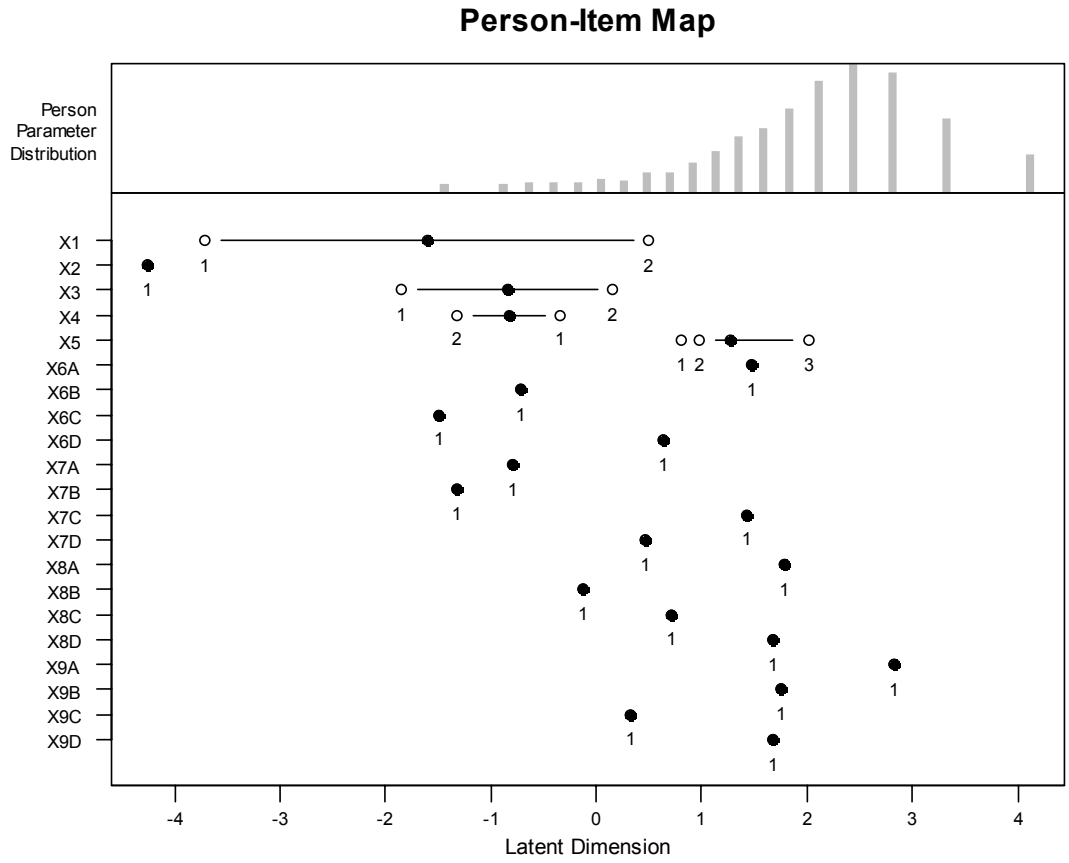


Figure 3.6: Person-item map for Biology foundation tier

While the other Science papers are better targeted, their lack of information compares unfavourably with the Geography and Mathematics papers. Figure 3.7 illustrates the wide range of discrimination achieved by the higher tier Geography Paper 2. Most item locations are close to the modal level of ability, but the category thresholds spread out across the ability range. This paper contains far more information on candidates at all levels of ability.

3.7.6 Item measures of fit

3.7.6.1 Mathematics Functional Skills

While the alpha coefficient was high for Mathematics Functional Skills, the variance explained was low and there were a number of misfitting items. The OPLM measures appeared relatively low, however, in comparison with the other tests, suggesting a reasonable fit. The further investigation at item level revealed that the same items were identified as most misfitting items by both the Outfit Mean Squares and the M statistic when the Rasch model was fitted (Table 3.5). The indices from the Classical Test Theory model reveal that three of these four most misfitting items are too easy ($p>0.7$) for this population, which could explain the low item-test correlations. The mean squares values are not exceptional and could be explained by poor discrimination. This suggestion is corroborated by the OPLM M3 statistic which is positive in each case, suggesting that a shallower item characteristic curve would show better fit.

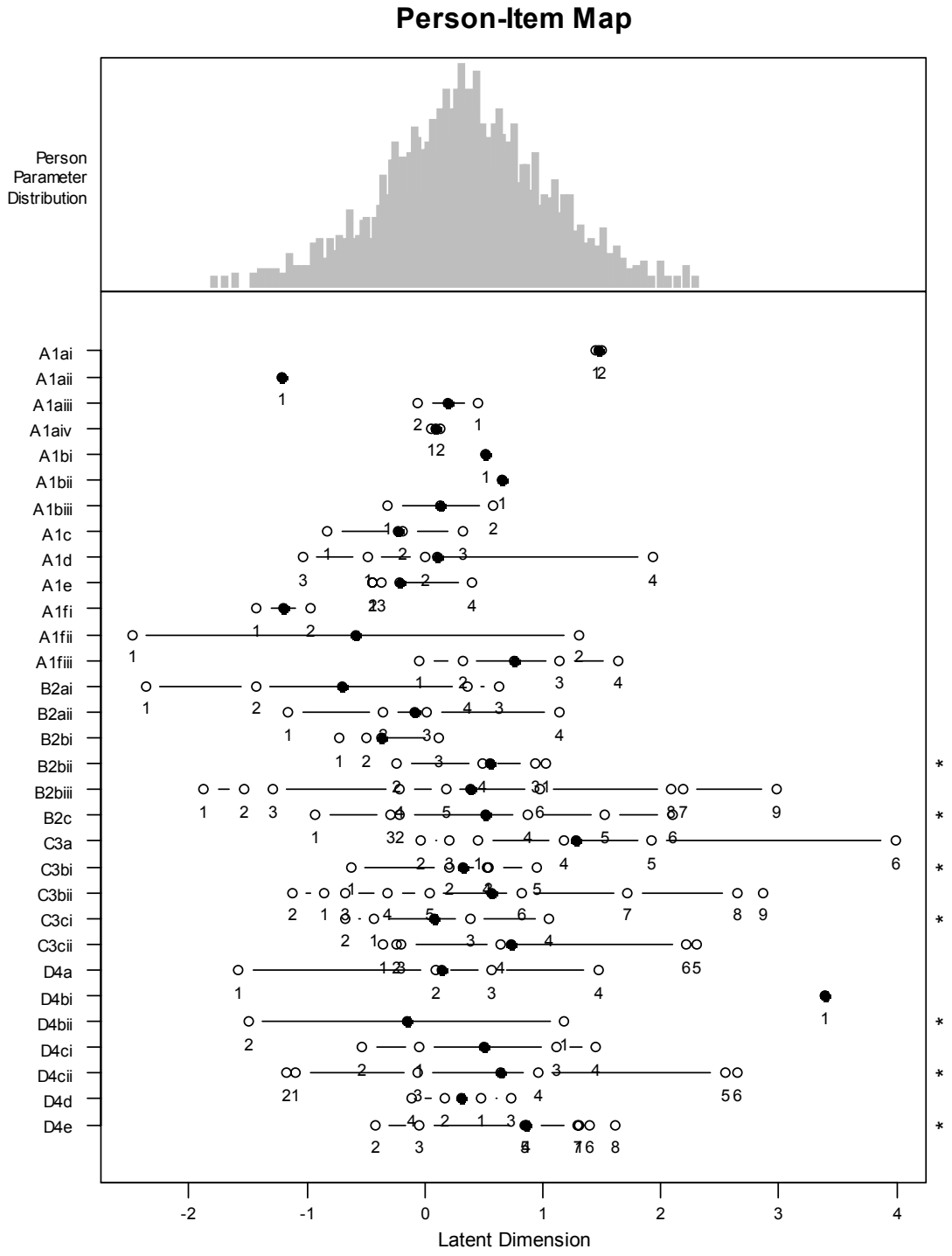


Figure 3.7: Person-item map for Geography Paper 2 higher tier

Table 3.5: Misfitting items for Mathematics Functional Skills

			Model						
	CTT		Winsteps		OPLM (Rasch)		OPLM (OPLM)		
Item	Facility	Item-test	B	Outfit (MnSq)	B	M3*	A	B	M3*
28	0.49	0.35	1.02	1.49	0.99	35.14	1	0.49	-4.64
11	0.81	0.26	-0.98	1.65	-0.95	26.98	1	-1.07	-3.42
2	0.90	0.21	-1.86	1.90	-1.80	21.49	1	-1.81	-3.05
4	0.79	0.31	-0.84	1.37	-0.81	20.58	2	-0.35	8.31

When OPLM was fitted the M statistic for three of these items fell relative to the other items so that items 28, 11 and 2 all became among the best fitting rather than the worst. Item 4, however, remained the worst fitting item under OPLM. The better fit for the other items is illustrated visually for item 28 in Figures 3.8 and 3.9. The observed performance is depicted as the solid black line while the modelled performance is depicted as the solid blue line with its associated confidence intervals. A cross shows agreement (or fit) of the modelled and empirical item performance while a dot shows deviation. The flexibility of OPLM in allowing the discrimination of the items to vary means that, in this case, a model which fits the majority of ability levels can be fitted to the data.

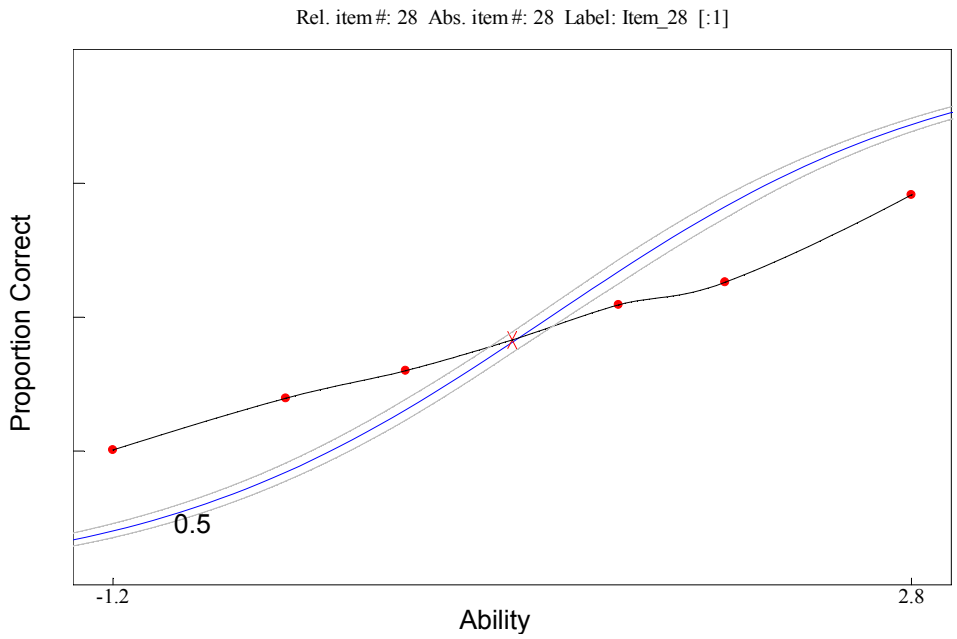


Figure 3.8: Item 28 fitted with the Rasch model[2]

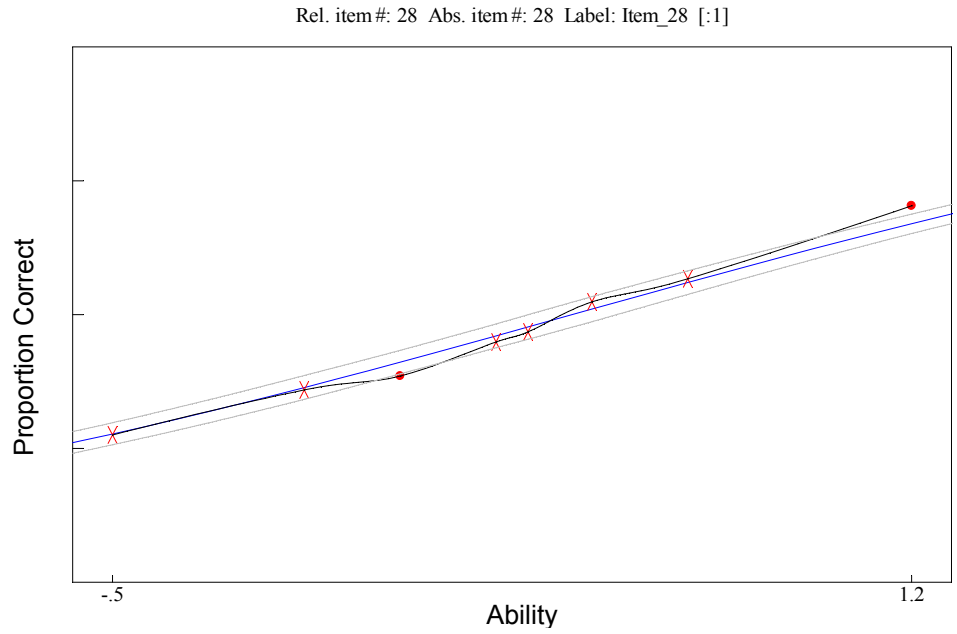


Figure 3.9: Item 28 fitted with OPLM

[2] The observed performance is depicted as the solid line intersected by dots and crosses while the modelled performance is depicted as the curve with its associated confidence intervals. A cross shows agreement (or fit) of the modelled and empirical item performance while a dot shows deviation.

Under OPLM, therefore, only item 4 (from these 4 items) remains a concern. Item 4 asks for the difference in temperatures between a maximum temperature of 11 degrees C and a minimum of -3 degrees C. An inspection of the ICC reveals that while the most able and the least able perform well on this item, those of middling ability perform less well than expected. The item does not represent a difficult concept, and the concept can presumably be drilled into the less able. For those of medium ability it may be assumed that they will get this item right, when they could have benefited from practice.

The Rasch Principal Components Analysis suggests that this item is central to the first contrasting dimension in the test, and that all these most misfitting items load negatively on the main Rasch factor (Table 3.6). It is worth considering, therefore, whether the cause of the misfit is the substantial second factor that can explain the variance in the test scores which was highlighted in Table 3.3.

Item 28 asks for 8786 to 1 significant figure, an area of notorious confusion, and there is a dip in performance amongst those of middling ability similar to that observed for item 4. Item 16 asks for one perspective of a 3D model, and again shows the same empirical response curve, with a dip in the middle. Item 11, however, which requires the candidates to name a solid from its net, shows the most able at an unexpected advantage and a different shape to the empirical responses. It would be hard to argue, therefore, that there is any clear dimension represented by the first contrast other than poor modelling of the responses by the Rasch model. The high proportion of misfitting items (6 out of 30) could, in this way, explain the discrepancy between the factor analysis and the Rasch Principal Components Analysis of residuals. A single factor can explain the variance in the test scores, but this single factor is poorly accounted for by the Rasch model.

It is worth also returning to the other most misfitting item under the Rasch model, item 2, which requires simple addition of two numbers. The table in which it is presented (Figure 3.10) does not make it sufficiently clear whether the number of pupils in each cell needs multiplication. While OPLM models the item better than the Rasch model, it is the poor presentation of the item that has resulted in its poor discrimination. There is no doubt that OPLM models the empirical response curves better than the Rasch model, which is useful if the data is assumed to be given. However, adjusting the discrimination parameter can hide quality problems with the items. If test quality is the most important reason for the model fitting, then the rigid assumptions of the Rasch model, in this case, are useful.

Table 3.6: Principal Components Analysis of residuals: first contrasting factor

Item	Loading	B	Infit (MnSq)	Outfit (MnSq)
4	-0.35	-0.84	1.18	1.37
28	-0.34	1.02	1.31	1.49
11	-0.34	-0.98	1.24	1.65
16	-0.31	-0.23	1.14	1.28

The table shows the number of pupils in a school.

	Number of pupils
Year 7 to Year 11	983
Year 12 and Year 13	275

What is the total number of pupils in the school?

Figure 3.10: Item 2 from Mathematics Functional Skills

3.7.6.2 Geography Paper 1 higher tier

The assumption of unidimensionality for Geography Paper 1 for both foundation and higher tiers was not supported. The hypothesis was that the practical component of this paper loaded on a different factor. This is borne out by the Rasch Principal Components Analysis of the residuals. Table 3.7 shows the loadings on the main Rasch factor for those items based on a map extract. The consistently negative loadings suggest that the responses to these items comprise a separate and coherent dimension in the data.

Table 3.7: Rasch Principal Components Analysis of residuals: loadings on the main Rasch factor for items based on a map extract

Item Id	Max Mark	Loading	Difficulty	Infit	Outfit
6 a i	1	-0.17	-0.24	1.02	1.05
6 a ii	1	-0.17	0.93	1.07	1.09
6 a iii	4	-0.32	0.56	1.08	1.09
6 a iv	4	-0.12	0.01	1.03	1.03
6 b i	6	-0.32	-0.10	1.07	1.09
6 b ii	4	-0.27	-0.19	1.03	1.04

The Rasch analysis revealed no misfitting items, although one 3 mark item displayed disordered categories. The candidates obtaining three marks on this item showed slightly less ability overall than the candidates obtaining two marks. Even OPLM, with its more powerful measures of item fit found significant levels of misfit in only 7 out of the 75 response categories. In all but two instances the misfit was caused by very few responses being observed in specific categories. Missing response categories can be an issue for test equating using separate calibration, as the

basis of the equating is the category thresholds. When there are few responses in a category the thresholds need to be estimated.

One dichotomous item showed misfit according to OPLM. This item showed excellent fit according to the Rasch fit analysis, although it did load negatively on the main Rasch factor according to the Principal Components Analysis. It is unlikely, however, that a single misfitting item would degrade any use of the model.

3.7.6.3 Physics higher tier

The Physics higher tier paper was one of the Science tests in which the modelling of the observed score distribution was poor (p=0.02). The classical item statistics revealed some demanding items with low item-test correlations. Neither the Rasch nor OPLM could model these items adequately as they showed response levels little above chance even for the brightest candidates.

Under OPLM the worst fitting items were not these items that did not discriminate. Rather, the item characteristic curves of the worst fitting items all appeared to have a lower asymptote suggestive of guessing. The comparison between the observed and expected scores for the item reproduced in Figure 3.12 under OPLM is illustrated in Figure 3.11. This is obviously quite a difficult item as it only appears to discriminate at the very highest level of ability. For all the other levels of ability the proportion of observed scores correct is fairly stable at between 0.3 and 0.4. The item requires candidates to select the relevant information from the table and enter it into the equations presented. It is actually quite hard, using the equations presented, to get a result of either £1.50 or £1500 so unsurprisingly the middle two options attracted 70 per cent of the responses; the other values may also seem implausibly high or low. Either of the middle options would therefore seem a

good bet. Neither OPLM nor the Rasch model explicitly includes a guessing parameter so they cannot model this item, or the others displaying this same pattern, accurately.

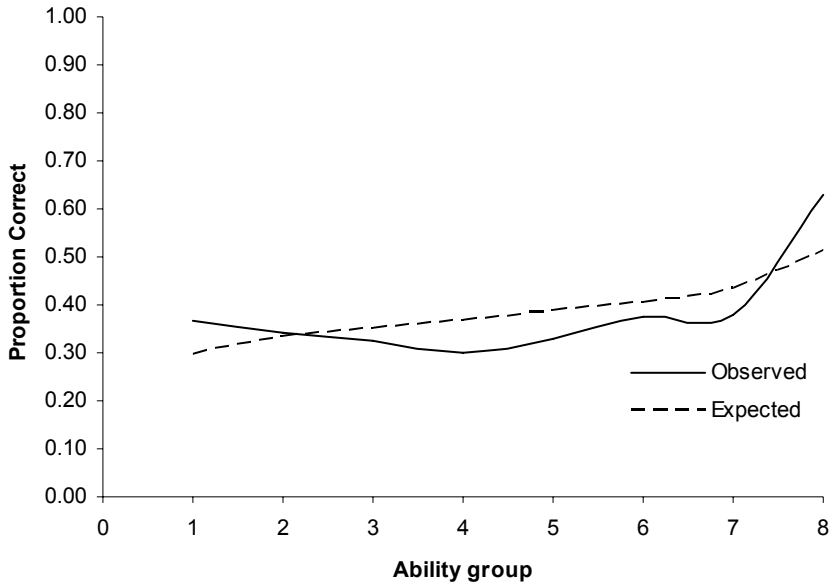


Figure 3.11: Observed and expected scores for Item 7C from Physics higher tier

The table compares data for two types of lamp.

	Filament lamp	**Compact fluorescent lamp (CFL)**
Cost	80p	£3.00
Efficiency	0.2	0.8
Expected life	1000 hours	8000 hours

Energy transferred = power x time

(kilowatt-hour, kWh) (kilowatt, kW) (hour, h)

Total cost = number of kilowatt-hours x cost per kilowatt-hour

Electricity costs 15p per kWh

7C What will be the cost of using a 100W filament lamp during its expected life?

1 £1.50

2 £15.00

3 £150.00

4 £1500.00

Figure 3.12: Item 7C from Physics higher tier

In an attempt to quantify exactly how much guessing affects the modelling of the Physics paper the expected scores (derived from OPLM) of the lowest ability group were plotted against their observed scores (Figure 3.13). Where the expected score is lower than the observed score this could be due to guessing. At least four items, all of them relatively difficult, displayed this pattern. Guessing could, therefore, contribute to the poor modelling of the observed score distribution, but it appears to be a small effect in a small number of items.

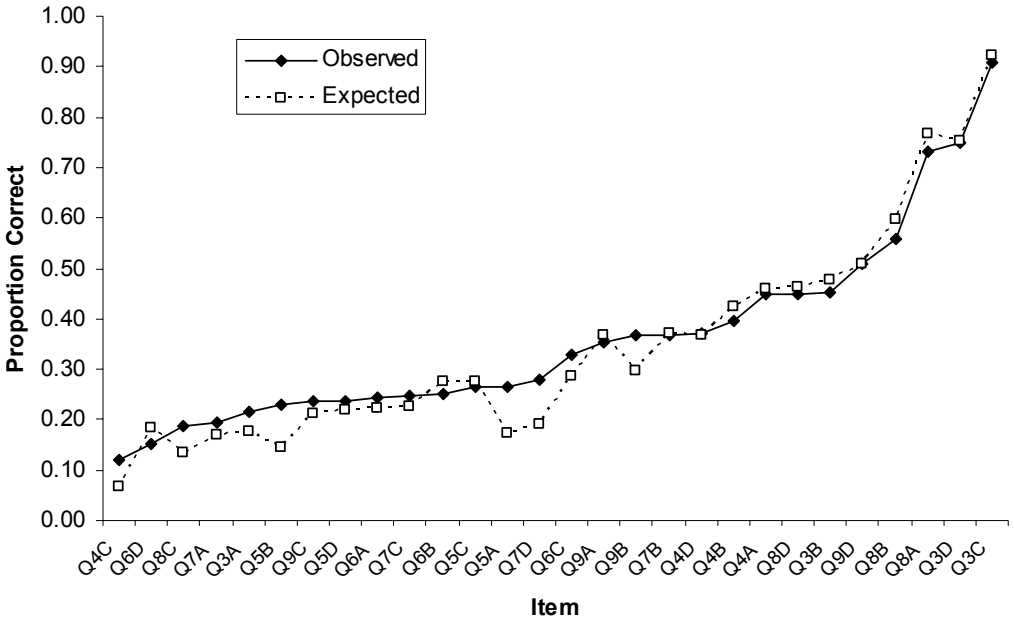


Figure 3.13: Expected performance compared to observed performance for the lowest ability group on Physics higher tier.

3.7.6.4 Mathematics Paper 1 foundation tier

Although the reliability and the variance explained for the Mathematics Paper 1 foundation tier was high (as it generally was for all the Mathematics papers) there appeared to be a large proportion of misfitting items. Some of the misfit could be due to the second dimension highlighted in Table 3.3 although both the factor analysis and the Rasch Principal Components Analysis reveal a substantial dominant factor. The paper does, nevertheless, appear to test a wide range of skills and concepts. Candidates are expected, for example, to perform tasks which vary from basic calculation to extrapolating from 3D diagrams to justifying the use of the median. A number of the most misfitting items were very easy; in this case the high outfit mean square is characteristic of carelessness.

There is a further potential explanation for the misfit: the Mathematics mark schemes allow multiple routes to a mark in order to reward positive achievement wherever possible. Evidence of mathematical worth, regardless of the answers given, can be rewarded with marks. The various categories under which marks can be awarded are detailed in Figure 3.14.

Category	*Reason*
M	Method marks are awarded for a correct method which could lead to a correct answer.
A	Accuracy marks are awarded when following on from a correct method. It is not necessary to always see the method. This can be implied.
B	Marks awarded independent of method.
M dep	A method mark dependent on a previous method mark being awarded.
B dep	A mark that can only be awarded if a previous independent mark has been awarded.
ft	Follow through marks. Marks awarded following a mistake in an earlier step.
SC	Special case. Marks awarded within the scheme for a common misinterpretation which has some mathematical worth.
oe	Or equivalent. Accept answers that are equivalent. eg, accept 0.5 as well as ½

Figure 3.14: Categories for Mathematics marks

Figure 3.15 shows the empirical category probability measures for the item in Figure 3.16. Category measure 1 discriminates poorly and shows disordered thresholds. Part of the reason for this could be the two different ways in which a score of 1 can be achieved. A mark of 1 could be achieved for calculating a third of six hundred pounds or for decoding the words of the item into mathematical symbols. These are potentially two different facets of mathematical ability. A score of 2 was highly unlikely, which suggests that there were very few candidates who

could write out the calculation but then fail to perform it. For those who did, however, the safety net of mathematical worth was there to catch them.

The partial credit model assumes that more of a category implies more ability; this incline in difficulty in the category marks is not reflected in the thinking in the Mathematics mark schemes. A third mark of three, for example, may be gained simply by adding the correct units to a complex calculation. One solution that is often used is to collapse the categories, particularly where there are few candidates in a category or the category does not appear to discriminate well. Ideally, were the information available, each mark could be modelled separately.

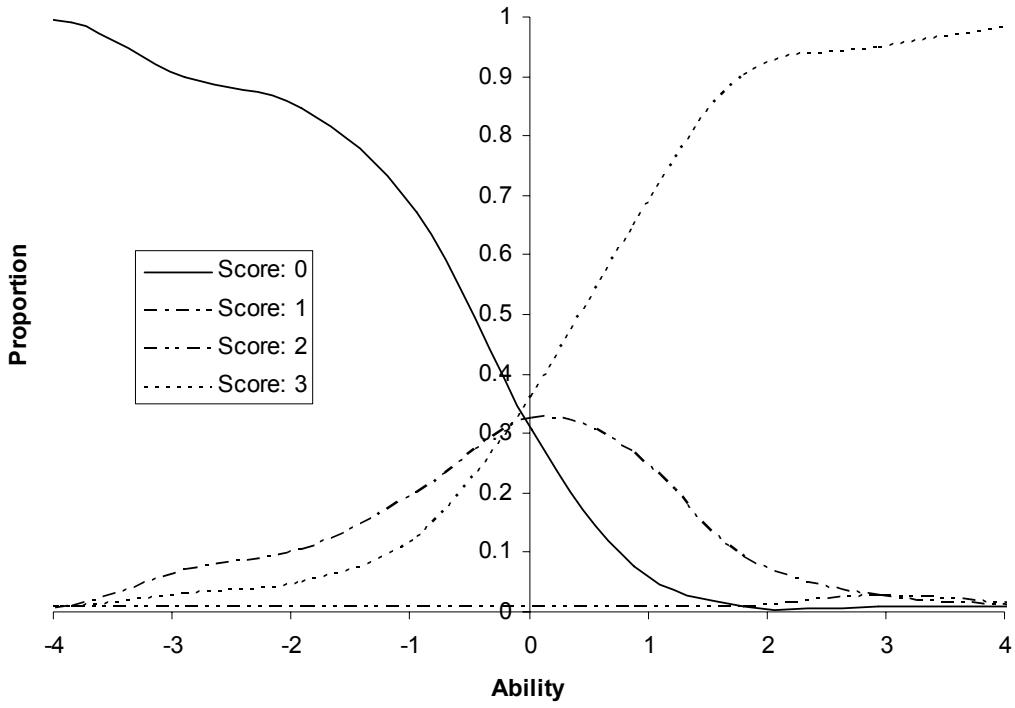


Figure 3.15: Empirical category probability curves for a GCSE Mathematics item with method marks

23 Dylan wants to buy this computer.

The price of £600 is reduced by $\frac{1}{3}$ in a sale.
He then pays $\frac{1}{4}$ of the sale price as a deposit.

How much is the deposit?
You **must** show your working.

.....

.....

.....

.....

Answer £ ..... *(3 marks)*

23	200	B1	
	[600 – (their 200)] ÷ 4	M1	
	100	A1	

Figure 3.16: The GCSE Mathematics item and mark scheme modelled in Figure 3.14

3.8 Discussion

The purpose of this chapter was to examine the statistical and mathematical assumptions that lie behind Rasch and IRT models and consider how well they are met for a selection of GCSE tests. Failure to meet these assumptions could mean that the conclusions drawn from the models are erroneous. Tests of model fit from both the descriptive Rasch paradigm and the restrictive statistical IRT paradigm were used, but they could necessarily only represent a small subset of the available tests. Tests of model fit are designed to have power against specific model violations. Molenaar's M tests, for example, indicate whether the discrimination of an item is set too high or too low. Perhaps the biggest omission from the tests applied here is

any checking of the invariance of item parameters across sub-groups of examinees. These subgroups may be based, for example, on gender, ethnicity or language groups. Although it is recognised that this testing is important the categorical variables on which such an analysis must be based are difficult to obtain, and often represent variables that are easy to collect rather than those which would be of specific interest and may explain a high proportion of variance. Some subgroup analysis based on levels of ability is attempted in a later chapter.

The underlying assumption of unidimensionality was examined using linear factor analysis. This approach does have limitations, as it is subjective, and relies on a linear relationship between items and the underlying latent trait. The magnitude of the eigenvalue of the second factor was generally greater in the observed data than in the simulations derived from the unidimensional models. This implies that the unidimensionality assumption is unlikely to hold.

Under the Rasch paradigm the next step would be to reorganise the tests so that they measure clearly separated unidimensional constructs. In Geography, for example, the test of map skills could be separated from the tests of knowledge and understanding and reported on a separate subscale. Several practical concerns present themselves in this respect. Firstly, it is not always possible to separate a coherent second factor in the data. Secondly, if a coherent factor is separated it too may prove to be multidimensional. There is no reason to believe that map-reading, for example, does not involve elements of different skills and abilities. Thirdly, in public examinations, subscales have no currency. The value of reporting them separately is therefore diminished. This approach, therefore, may have more use in a formative assessment setting.

Under the IRT paradigm the next step would be to fit a multidimensional IRT model. These models, however, are not yet operational. So, while they may be of theoretical interest, they are at present of little practical use. The third possibility is to proceed with the fitting of IRT models and to rely on the results of simulations that suggest whether or not the specific uses to which the models will be put are robust to violations of unidimensionality. The strong suggestion is, for example, that test equating is robust to violations of unidimensionality.

The assumptions regarding distributions of ability under MML were examined using both statistical tests and more descriptive measures. For the shorter tests there appeared to be substantial deviations between the observed score distributions and the expected score distributions based on trait estimates for the Rasch model. For the Science tests, for example, the acute peaks in the score distributions were not predicted by the model. There could be many reasons for this, which include: guessing; violation of the conditional independence of responses to items; or the measurement of different dimensions causing scores to regress to the mean. Where there are deviations between the observed score distributions and expected score distributions then conclusions based on these frequency distributions could be misleading. For example, if the Rasch model is used to measure classification accuracy or consistency, the proportions of candidates achieving different scores is critical. In this scenario the interpretation of results based on the Rasch model could be problematic.

At the item level various issues with fit were highlighted. In Mathematics Functional Skills the axiom of monotone increasing Item Response Functions did not seem to hold for a number of items. Carelessness or drilling of the less able could explain this finding. A poor and confusing stimulus appeared to be responsible in

one case. In the Physics test poor distractors were responsible for misfit in at least one case and quite probably some others. In Mathematics the mark schemes resulted in uneven distributions of scores between mark categories and the associated disordered item thresholds. Again, according to the Rasch paradigm, the next step would be to examine the quality of the misfitting items with a view to improving their quality. More carefully worded questions, better quality stimuli, better distractors and more coherent mark schemes could improve the quality of items and improve the fit of the Rasch model. According to the IRT paradigm, categories could be collapsed and guessing or carelessness parameters added to the model.

It is certainly possible and valuable to pursue the Rasch paradigm to a certain extent. Item writers can receive training on sources of difficulty in test items, for example, and better item quality screening can be put in place. However, without pre-testing, it is inevitable that items will vary in difficulty and discrimination. For this reason the Rasch model would seem overly restrictive for operational practice without routine pre-testing. OPLM would appear to be an attractive alternative as it allows discrete item discrimination parameters in the model. All the tests, however, showed substantial levels of misfit to the OPLM model. In particular, the Science tests proved difficult to model. It was not possible, however, within these analyses, to isolate the reason for that misfit. Two obvious hypotheses are worth dismissing: that the misfit is due to guessing, or that the misfit is due to violations of the conditional dependence of the item responses. The purpose of the next chapter is to fit more complex models in order to test these hypotheses.

4. Model Fit in a Bayesian Framework

4.1 Overview

The previous chapter highlighted potential issues with the fit of the Rasch model and OPLM for certain GCSE tests. The assumption of a single discrimination parameter for the Rasch model appeared overly restrictive for all of the tests measured; and poor test and item fit for a number of the tests highlighted potential issues with dimensionality, guessing and weak local independence. In the Rasch paradigm these issues would be addressed through a reorganisation of the data and improvements made to the quality of the test items. This approach, while it has much to recommend it, has certain limitations. Difficulty will always be hard to control without pre-testing, tests will always be multidimensional to a degree, guessing will always be possible on multiple-choice items and weak local independence may be inevitable when groups of questions follow a stimulus. This chapter will instead, therefore, pursue the IRT paradigm and fit more complex models to the data in a Bayesian framework.

4.2 Why use Bayesian estimation?

Bayesian estimation procedures for IRT models were first proposed by Swaminathan and Gifford (1982). Recently, the approach has been adopted to the estimation of IRT models with a correlational structure of latent abilities (de la Torre, 2009), multi-level structures with clustering of respondents by background variables (Fox & Wyrick, 2008), multiple raters structured by covariates such as training (Mariano &

Junker, 2007) and local dependence among response categories (Bolt, Cohen & Wollack, 2001). All of these applications use a Markov Chain Monte Carlo (MCMC) algorithm in order to make Bayesian inferences. The motivation behind these investigations of Bayesian inference and MCMC is that the complex dependency structures that are being modelled require the evaluation of multiple integrals to solve the estimation equations in an MML or Bayes modal framework (Patz & Junker, 1999). These problems are avoided in an MCMC framework.

Apart from being able to estimate more complex models, the MCMC framework also offers powerful tests of model fit that do not depend on asymptotic analysis. Under IRT models the possible number of responses (2^I for a test with I binary items) is sufficiently large for even a moderately large number of items that the standard $\chi 2$ test of goodness of fit does not directly apply (Sinharay, 2005). The usual approach, therefore, to the fit of IRT models is to investigate whether the model can explain various summaries (or collapsed versions) of the original data (Glas & Falcon, 2003). In a Bayesian MCMC framework the posterior predictive model checking (PPMC) method (Guttman, 1967; Rubin, 1984) provides an appealing alternative because of its simplicity, strong theoretical basis, and intuitive appeal. The method primarily consists of comparing the observed data with replicated data (those predicted by the model) using a number of test statistics.

A further advantage of Bayesian modelling is that, through simulation, it can address whether the misfit of a model has substantial practical consequences for the intended applications of the model. It is possible that discrepancies between the test data and predictions from a model are of no practical consequences. Only a p-value or a diagnostic plot rarely provides any insight about practical consequences of misfit (Sinharay, 2005).

4.3 Bayesian procedures

Bayes' theorem provides a representation of the conditional probability of one event given another (i.e., A given B, or the probability of measles given a certain test result) in terms of the opposite conditional probability (i.e., B given A, or the probability of a particular test result given the presence or absence of the measles). In terms of Bayes' theorem, information on model parameters such as item difficulty, item discrimination and examinee ability is reflected in the relative likelihood of particular parameter values for the model given the observed item response data. The opposite conditional probability is the probability of the item response data (B) given the model parameters (A). As item and ability parameters are continuous values, not discrete events the "probability" of their occurrence can be expressed as continuous probability density functions. The goal of MCMC analysis is to reproduce the joint posterior density, in other words the probability density functions of the model parameters given the data (Kim & Bolt, 2007).

If ω denotes the vector of parameters, then the posterior distribution of ω given the observations or data is

$$\pi(\omega|y) = \frac{L(y|\omega)\pi(\omega)}{\pi(y)} \tag{4.1}$$

where y is the dataset (the set of item responses of N examinees to n items). In the IRT context, $\omega = [\vartheta\ \xi]$ where ϑ is the vector of trait parameters and ξ is the vector of item parameters (Swaminathan et al., 2007). The joint posterior distribution of item and trait parameters contains all the information about these two sets of parameters and takes into account the uncertainty in both item and trait parameters (Swaminathan & Gifford, 1982).

4.4 Posterior Predictive Model Checking (PPMC)

A posterior predictive distribution is a replicate set of observations conditional on the distribution of model parameters given the observed data. In a Bayesian framework, Markov Chain Monte Carlo (MCMC) techniques can be used to generate these sets of replicate observations. Once generated, any discrepancy statistic of interest can be calculated across these replicate observations and compared against the same statistic in the observed data. If a large percentage of the discrepancy statistics observed for replicated data sets (say 95 per cent) exceed those in the observed data, or a large percentage are lower than those in the observed data, then it would seem that the feature in the data corresponding to the statistic is not being replicated by the model (Kim & Bolt, 2007).

The discrepancy statistics used depend on the structural information required from the model. A basic requirement of all test models is that they preserve the rank order of individuals. For this purpose the comparison of observed and expected scores produced in the previous chapter provides a useful descriptive measure (Béguin & Glas, 2001). At the item level, the following tests are suggested (Sinharay, 2005; Sinharay, Johnson & Stern, 2006):

1. The point biserial correlation coefficients show the extent to which items are consistent with a test, and are closely linked to the discrimination indices in the 2- and 3- parameter IRT models.

2. Item-interdependence caused by a test not being truly unidimensional can be tested using the odds ratio (Chen & Thissen, 1997). The odds ratio is represented as,

$$OR = \frac{n_{00}n_{11}}{n_{01}n_{10}} \tag{4.2}$$

where *n* represents the number of examinees obtaining a given sequence of examination scores and the subscripts identify the pattern (e.g. n_{10} indicates the number of examinees answering the first item in the pair correctly and the second item incorrectly).

3. The Mantel-Haenszel statistic which compares odds ratios across slices of ability (Sinharay, 2005). The comparison across slices of ability gives it more power than the odds ratio.

4.5 Exploring different models

Under the IRT paradigm poor model fit leads to an exploration of more complex models. This approach risks naïve empiricism as the number of model parameters could be increased until the model 'fits' (Feyerabend, 1988). It is a well known theorem of mathematics that an *N*-degree polynomial can fit *N* data points exactly (provided none is exactly on top of any other). Just as well-known is Occam's razor: the fewer adjustable parameters required to explain something the better. The challenge therefore is to use the simplest model possible that captures the information of interest and fits well enough. If a model overfits then all the noise and idiosyncrasies in the data will be modelled and the predictions degraded (Hitchcock

& Sober, 2004). If it underfits, however, then the model is no longer an adequate representation of the data and the predictions will again be degraded.

Any assessment of the relative worth of different models is complicated when more complex models are nested inside simpler models. The more complex model will always show better fit, but at the cost of some simplicity of interpretation. In an IRT context, for example, a two-parameter IRT model will show better fit than a one-parameter IRT model, but the Item Response Functions are no longer parallel. This means that the order of difficulty of items may vary according to ability. A mathematical estimation of the compromise between fit and complexity was formulated by Akaike (1973) who showed that an unbiased estimate of the predictive accuracy of a model can be obtained by combining a measure of fit with a measure of simplicity.

Akaike's procedure is as follows. A model is first fitted to the data at hand and then the fitted model is used to predict new data drawn from the same underlying distribution. The fit of the model to these data are estimated using the logarithm of the likelihood. This process is then repeated: data are drawn, the likeliest member L(M) of the model M determined, and L(M) evaluated in predicting new data. The average (expected) fit of L(M) to new data defines M's predictive accuracy. Given certain assumptions, an unbiased estimate of the predictive accuracy of model M can be represented by,

$$M \cong \log[\Pr(Data \mid L(M))] - k \tag{4.3}$$

where k represents the number of adjustable parameters. For the complex model to have higher estimated predictive accuracy the data must fit the data sufficiently better to compensate for the loss in simplicity it represents. While Akaike's formulation applies to parameters estimated using maximum likelihood estimation,

analogous measures have been developed for use with other forms of estimation. Simulation studies have shown that the Bayesian model selection method, the Deviance Information Criterion (DIC), based on a posterior mean deviance and a penalty for model complexity, appeared to be stable and accurate in model identification (Jones, 2002).

4.6 Beyond the Rasch model

There are many models that could be fitted to the data examined in Chapter 3. Four models suggest themselves, however, from that analysis. The 2-parameter IRT model (Birnbaum, 1968; Lord & Novick, 1968) would address the poor fit of the Rasch model that was due to different levels of item discrimination. The 3-parameter IRT model (Birnbaum, 1968; Lord & Novick, 1968) would address the issue of poor fit due to guessing that could be an issue for the Science tests. The Multi-Class Mixture Rasch Model (MMRM) for Test Speededness (Mroch, Bolt & Wollack, 2005) could explain some of the poor fit due to a speeded dimension appearing late in the tests. All of the tests suggested a second substantial factor could explain some of the variance in the test scores. It was not always obvious what that factor was. Finally, a Testlet IRT model (Wainer, Bradlow & Wang, 2007) may explain poor fit due to weak local independence. This is particularly an issue for the Science tests that are explicitly designed with a testlet structure.

4.6.1 The 2-parameter item response model

Critics of the Rasch model argue that the parallel Item Response Functions of the Rasch model restrict its application to relatively homogenous items (Hambleton,

Swaminathan & Rogers, 1991). This may explain the relatively poor fit of the Rasch model to GCSE tests noted in the previous chapter. This restriction is relaxed by the two-parameter model which has the form:

$$P(U_i = 1|\vartheta, a_i, b_i) = \frac{e^{a_i(\vartheta - b_i)}}{1+e^{a_i(\vartheta - b_i)}} \tag{4.4}$$

The main element that distinguishes this model from the one-parameter model is the parameter a_i which is called the item discrimination parameter. The a_i parameter is proportional to the slope of the Item Response Function at the point b on the ability scale. Steeper slopes represent higher discrimination.

4.6.2 The 3-parameter item response model

The relatively high item facilities for the GCSE Science tests noted in Chapter 3 suggested that guessing may be responsible for some part of the poor model fit for these tests. For multiple-choice style tests the guessing or pseudo-chance level parameter c_i can be added to the model. This represents the lower asymptote of the Item Response Function.

$$P(U_i = 1|\vartheta, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{e^{a_i(\vartheta - b_i)}}{1+e^{a_i(\vartheta - b_i)}} \tag{4.5}$$

4.6.3 The Multi-Class Mixture Rasch Model (MMRM) for test speededness

All of the GCSE tests in the previous chapter appeared to have a significant second factor which could explain some part of the variance in the test scores. This factor could be valid, related to a different skill, for example. The factor could, however, represent construct irrelevant variance due to, for example, a speeded dimension appearing towards the end of a test.

While GCSEs and A-levels are not intended to be speeded, little work has been done to examine whether there is enough time to complete all the questions within a given time limit. Inevitably, some candidates will fail to complete a test. This means that item parameters for items at the end of the test may be overestimated compared to their difficulties if estimated earlier (Bolt, Cohen & Wollack, 2002). In the context of test equating this could lead to erroneous conclusions if anchor items are placed towards the end of one test form but towards the beginning of another. Several models have been proposed to address problems regarding speededness (Bolt et al., 2001; Mroch et al., 2005; Wollack, Youngsuk & Bolt, 2007; Yamamoto & Everson, 1997). These models address speededness effects through the introduction of latent examinee classes that are distinguished by individual differences.

Under MMRM, multiple latent classes are distinguished by the end of test item locations at which their responses become speeded (if at all) (Mroch et al., 2005). The MMRM assumes that examinees belonging to the same latent class experience common item difficulties for items at the end of the test. Under the MMRM, the probability that an examinee *j* answers an item *i* correctly is written as follows:

$$P\left(U_{ij}=1|\vartheta_{gj},b_{ig},g\right)=\frac{e^{\left(\vartheta_{gj}-b_{ig}\right)}}{1+e^{a_i\left(\vartheta_{gj}-b_{ig}\right)}} \tag{4.6}$$

where,

U_{ij} is the 0/1 response of examinee j to item i,

ϑ_{gj} is the latent ability parameter of examinee j in class g

and b_{ig} is the difficulty parameter for item i in class g.

This equation is similar to that for the Rasch model, with the key difference being the subscript g which indexes each latent class. Equality constraints are placed on the difficulty parameters for all unspeeded items across classes as well as for all speeded items across classes. This means that there exist two difficulty parameters for each potentially speeded item, one for its speeded condition and one for its non-speeded condition. Further, an ordinal constraint is applied to each speeded item such that the speeded item difficulty is always higher than it is for the same item in the non-speeded condition. As a result, an examinee's response pattern can be said to exhibit effects of speededness when the relative difficulties of items at the end of the test are higher than the relative difficulties of items at the beginning. The model has proved useful in recovering item parameters from simulated speeded test data, although it tends to underestimate the ability of examinees affected by test speededness (Mroch et al., 2005).

The MMRM does, however, have several drawbacks (Wollack et al., 2007). The designation of which items will be modelled as speeded is arbitrary and must be done in advance of the modelling. The model is sensitive to examinees whose performance on end-of-test items is appreciably worse than on the rest of the test; therefore, it requires examinees to have achieved a certain level of performance prior to becoming speeded. Consequently, the mixture model is biased against identifying low-ability candidates who have run out of time. The mixture model approach is also extremely time-consuming.

4.6.4 Testlet Response Theory (TRT)

A testlet is defined as a group of items that may be developed as a single unit that is meant to be administered together (Wainer et al., 2007). An example of a testlet is

the traditional reading comprehension item type in which the examinee is presented with a passage and a bundle of items related to that passage. The GCSE Science tests presented in Chapter 3 have a testlet structure. Each group of four items is arranged around a single stimulus. The stimulus may be a reading passage, a table of data, a figure, or any combination of these three. Other GCSEs and A-levels use thematic groupings of items that to some extent could be described as testlets. The problem with modelling tests based on a testlet structure is that the items within testlets may exhibit conditional dependence (Wainer et al., 2007). This violates an essential underlying assumption of Rasch and IRT models. If the testlet items are modelled as conditionally independent the amount of information in the test will be overestimated. In fit analysis, testlets may be misdiagnosed as poor fit or multidimensionality if the testlet structure is not explicitly modelled.

The testlet model in a Bayesian framework is given by Wainer et al. (2007) as follows. Under Bayesian modelling two-parameter dichotomous items are modelled as,

$$P(Y_{ij} = 1) = logit^{-1}(t_{ij}) \tag{4.7}$$

where Y_{ij} is the response of examinee *i* on item *j* and t_{ij} is the latent linear predictor of score. The extra dependence due to testlets is modelled by extending the linear score predictor t_{ij} from its standard form,

$$t_{ij} = a_j(\vartheta_i - b_j) \tag{4.8}$$

where a_j, b_j and ϑ_i have their standard interpretations as item slope, item difficulty, and examinee proficiency, to

$$t_{ij} = a_j(\vartheta_i - b_j - \gamma_{id(j)}) \tag{4.9}$$

with $\gamma_{id(j)}$ denoting the testlet effect (interaction) of item j with person i that is nested within testlet d. Items within the same testlet d for a given an examinee j share the effect $\gamma_{id(j)}$ in their score predictor.

Different approaches have been taken to testlets in the past. One approach, described in Moreno and Segall (1997) was to remove the testlet structure from the test design. The Armed Services Vocational Aptitude Battery used a traditional reading comprehension design. When the battery was transformed to computer based administration a decision was made to adapt the reading comprehension section in order to reduce conditional dependence and facilitate computer adaptive testing. The traditional passage was about 300 words long and was accompanied by five items. The first consideration was to retain the passage at its current length but simply ask a single question. This seemed very inefficient. Instead, they asked a single question but of a much shorter passage which was about 120 words long. The new items showed an increase in their correlation with word knowledge and correlated only 0.38 with the prior form of paragraph comprehension items. The shortening of the passage solved the conditional dependence but changed the construct being measured.

A Rasch approach to the problem may be to model each testlet as a single polytomous item. For example, a reading comprehension passage followed by four questions would be marked as a single polytomous item with a score from 0 to 4. This approach, however, does not allow the discrete modelling and quality control of each item within that testlet. Further, under the IRT paradigm, where patterns of responses contribute information to the ability parameter of each examinee over and above their summed score, this approach leads to a substantial loss of information (Wainer et al., 2007). An alternative IRT approach is to use a multidimensional IRT

model. A multidimensional IRT model, however, suggests the presence of disparate coherent constructs in a test which should be reported as separate subscales. This description would not seem to suit the Science and reading comprehension tests described above. The advantage of the TRT model is that it allows the requirement of conditional independence to be relaxed in a unidimensional IRT framework (Wainer et al.).

Although designed under the IRT paradigm, TRT models can prove useful in the test design process rather than simply improving the fit of models. The amount of dependence that testlet structures introduce is an interesting topic for test designers. Further, TRT models have been used to model internal structures within tests such as speededness (Wollack et al., 2007). Speededness can be described by greater conditional dependence than would be expected within each half of a test. The major disadvantage with TRT models is that they are computationally intensive and time-consuming.

4.7 Method

4.7.1 Application 1: The Multi-Class Mixture Rasch model (MMRM) for test speededness

4.7.1.1 Design

To examine the impact of test speededness on the estimation of item parameters the Multi-Class Mixture Rasch Model (MMRM) was estimated for a number of tests. For each test an arbitrary decision was taken on the number of latent classes to distinguish and the number of items that would be designated as potentially speeded. All items that were not designated as speeded were constrained across classes to the

same difficulty. It was assumed that the items are answered in order and that when a given item response is speeded, responses to all subsequent items are also speeded. For the longer tests only the initial section of the test and the designated speeded items at the end of the test were modelled in order to facilitate computation.

A Markov Chain Monte Carlo (MCMC) algorithm was used to estimate the parameters of the MMRM model. WinBUGS software (Spiegelhalter, Thomas, Best & Lunn, 2003) was used for this purpose as well as code written by the author. This code is archived here (https://github.com/cbwheadon/predicted_scores). 1000 iterations were sampled from the Markov chain and 500 iterations were used as burn-in, leaving 500 iterations sampled from the posterior distribution to use as estimates of model parameters. Convergence of the MCMC solution was determined by inspecting plots of sampling histories for estimated parameters. For the ability and item difficulty parameters, the means of the sampled values (after burn-in) were used as estimates of the parameters. Analysis of the latent classes was undertaken using the medians of the sampled values (after burn-in).

4.7.1.2 Priors

The prior distributions were as follows:

b_{ig} ~ Normal(0, 1)

ϑ_{gj} ~ Normal(μ_g, 1)

$\mu_{\vartheta 1}$ ~ Normal (0, 1) (for the unspeeded class)

$\mu_{\vartheta 2} \ldots \mu_{\vartheta k}$ are functions of the unspeeded and speeded item parameters

c_j ~ Categorical (π_1, π_2, … π_k), where c_j = {1, 2, …, k} is a class membership parameter

(π_1, π_2, … π_k) ~ Dirichlet (1, …, 1).

4.7.1.3 Components

Speededness is only really an issue for item parameter estimation if the items are located in different places on different test forms. Common items are used as a measure of comparative performance between tiers on certain GCSE tests. For certain test forms, these common items are located towards the end of the foundation tier and towards the beginning of the higher tier. This design is typical of both Mathematics and Science. As the speededness is therefore likely to affect the item parameters of the foundation tier papers in these subjects, these tests were selected for analysis. As the analysis is computationally intensive samples of 1000 candidates were taken from each test. The analysis was replicated across two separate samples. Due to the complexity of the models the responses were all dichotomised. An algorithm was used for this dichotomisation which split the responses around their median value.

4.7.2 Application 2: The testlet model

4.7.2.1 Design

To examine the amount of dependence that testlet structures introduce, a 1-parameter, 2-parameter, 3-parameter and 2-parameter testlet model were fitted to a number of tests which use a testlet design. This allowed model fit to be examined using the Deviance Information Criterion (DIC) (Spiegelhalter, Best, Carlin & van der Linde, 2002). The model with the smallest DIC is estimated to be the model that would best predict a replicate dataset of the same structure as that currently observed. If the DIC is lower for a 2-parameter model than the 2-parameter testlet model, for example, then it may be concluded that the testlet structure is not a key feature of the data.

DIC is intended as a generalisation of Akaike's Information Criterion (AIC) (1973) and caution is advisable in the use of DIC as it is still experimental (Spiegelhalter et al., 2003). For this reason a number of PPMC checks were then undertaken. The PPMC checks reveal the extent to which the models preserve specific features of the original dataset.

WinBUGS software (Spiegelhalter et al., 2003) was used to estimate the models using Markov Chain Monte Carlo (MCMC) algorithms. For each model, one chain of length 10,000 was run, with the first 4,000 iterations burnt in and a thinning rate of 6. This approach yielded 1,000 simulated datasets. Non-informative priors were used and confirmation of convergence obtained through visual inspection of the sampling history.

As the computation was intense and the simulated datasets large (1,000 simulations of 1,000 candidates on 21 items yields 21 million responses) the statistical package R (R Development Core Team, 2010), which includes optimised routines for calculations on large matrix style data, was used to drive WinBUGS using R2WinBUGS (Sturtz, Ligges & Gelman, 2005). Apart from improving computational efficiency this approach also allows for better error trapping and direct exploration of results through descriptive plots, which were programmed using ggplot2 (Wickham, 2009).

4.7.2.2 Priors

The likelihood parameters are assumed to have normal prior distributions. They are specified as:

$\vartheta_{jg} \sim Normal\ (0,1)$

$a_i \sim \log\left(Normal(\mu_a, \sigma_a^2)\right)$

$b_i \sim \log\left(Normal(\mu_b, \sigma_b^2)\right)$

$\gamma_{jd(i)} \sim Normal\left(\mu_{\gamma(d)}, \sigma^2_{y(d)}\right)$, d = 1, …,K

4.7.2.3 Components

The higher tiers of the GCSE Science tests described in Chapter 2 were chosen for analysis as these used specifically designed testlets. The higher tier uses 7 distinct testlets, while the foundation tier only uses 4 so the higher tier was chosen for the analysis. Due to the computational intensity noted above a sample of 1,000 candidates was taken from each test. The mean and standard deviation of the summed scores of this sample were checked for representativeness before the modelling was undertaken. Due to the complexity of the models the responses were all dichotomised. An algorithm was used for this dichotomisation which split the responses around their median value.

4.8 Results

4.8.1 The Multi-Class Mixture Rasch Model (MMRM) for test speededness

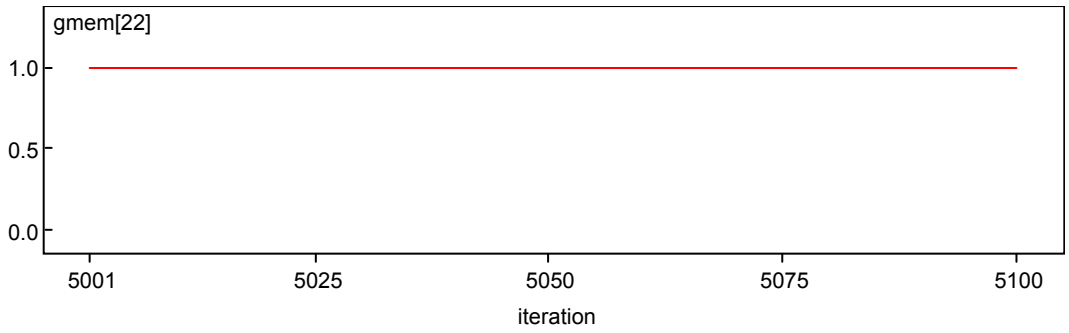
4.8.1.1 Convergence

The main parameter of interest is the group membership parameter which assigns candidates to a particular latent class. Traces of the samples were used to estimate the point at which convergence had been reached. Figure 4.1 shows the traces between iterations 5001 to 5100 and the response patterns they are modelling. The classes run from 1 (unspeeded, or under no time pressure) to 7 (speeded or under time pressure).

For examinee 22 there was no doubt regarding the classification. With a strong finish of three correct answers it seems safe to assume that this candidate was not under time pressure. As long as the underlying assumption of the model that the candidate attempts the final questions in order is taken as given, the classification as unspeeded seems secure. It cannot be known, however, whether time pressure caused the candidate to skip two of the last six items.

For examinee 23 the pattern is less secure. Ending with two wrong answers despite a relatively strong start could place the candidate in anything from class 1 to 5. The median value of class 4 seems a reasonable approximation, although the 2.5 per cent to 97.75 sampling intervals are 1 to 4. Finally, examinee 3 has a strong start so the final 5 incorrect answers would seem to suggest time pressure. Although the median class value is 5, which would seem correct, the 2.5 per cent to 97.75 sampling intervals are between 1 and 7.

As the median values seemed to represent a fair representation of the data convergence was assumed. It is, of course, impossible to assert that convergence has been reached; it is much easier to state when it has not been reached (Spiegelhalter et al., 2003). Other parameters were also checked to ensure that no residual upward or downward trends in the sampled values remained.



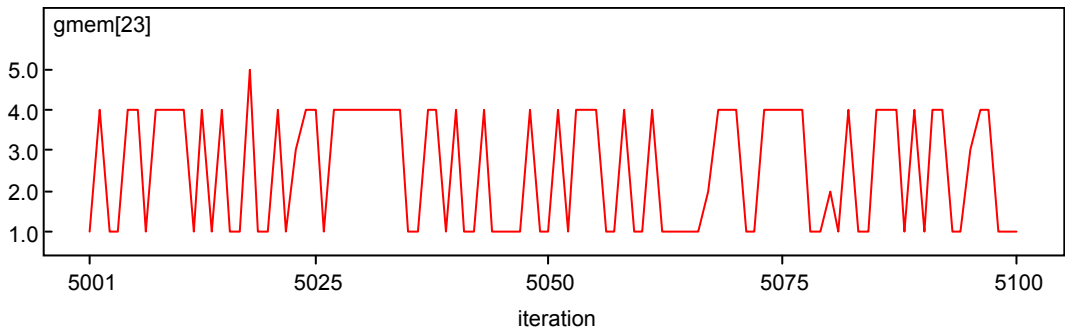
Examinee: 22

Pattern:

1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,0,1,0,0,1,1,1,0,1,1,1,1,0,1,1,1,0,0,1,0,1,1,1,1,1,1,1,

1,1,1,1,1,0,1,1,0,1,0,0,1,0,0,1,1,1

Median Class: 1



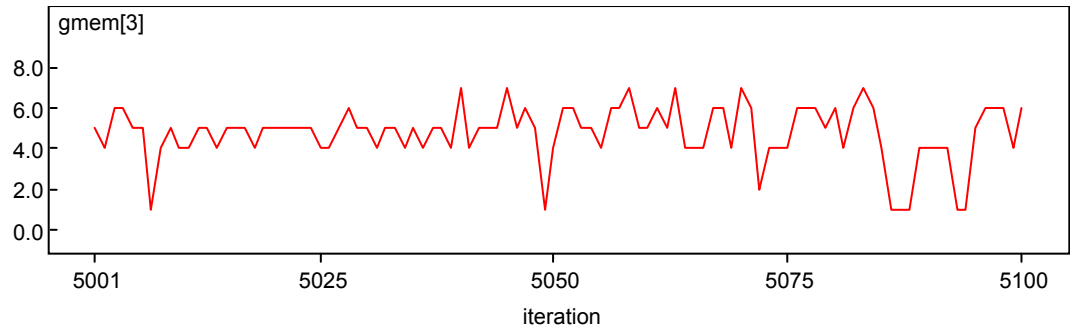
Examinee: 23

Pattern:

1,1,1,1,1,1,1,1,0,0,1,1,1,1,1,0,0,1,1,0,1,1,1,1,1,0,1,1,0,1,1,1,1,1,1,0,1,1,0,1,0,0,1,1,0,

1,1,1,1,1,0,1,1,1,1,1,0,1,0,1,1,0,0

Median class: 4



Examinee: 3

Pattern:

1,1,1,1,1,0,1,1,1,1,1,1,1,1,0,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,0,0,

1,1,0,1,1,1,1,1,1,1,1,1,0,1,0,0,0,0,0

Median class: 5

Figure 4.1: Sampling traces for the latent class parameter for iterations 5001 to 5100: Mathematics Paper 1 foundation tier

4.8.1.2 Stability of samples

The stability of estimations of speeded class was checked across the three samples of 1,000 candidates taken from the Mathematics tests. Tables 4.1 and 4.2 show that the estimated proportion of candidates classified as unspeeded seems relatively consistent for both papers. The last classes, which distinguish the point at which time pressure begins, are the most unstable. There is clearly uncertainty in determining the exact point at which time pressure begins for a candidate. This is not unreasonable, as time pressure is likely to build rather than be an automatic switch from unspeeded to speeded. It would appear, however, that the time pressure builds from 5 and 6 items from the end on Paper 1 and 6 to 7 items from the end on Paper 2. All further analysis is based on the first sample.

Table 4.1: Class membership compared across three samples (N=1,000) candidates for Mathematics Paper 1. 1 = Unspeeded, 7 = speeded 6 items from the end.

Class	1	2	3	4	5	6	7
Sample 1	558	17	36	84	180	125	0
Sample 2	543	32	14	119	16	276	0
Sample 3	546	2	0	54	382	16	0

Table 4.2: Class membership compared across three samples (N=1,000) candidates for Mathematics Paper 2. 1= Unspeeded, 7 = speeded 6 items from the end.

Class	1	2	3	4	5	6	7
Sample 1	457	22	30	24	13	434	20
Sample 2	450	24	13	32	11	89	381
Sample 3	489	15	13	32	11	407	33

4.8.1.3 Effect of speededness

Table 4.3 summarises the proportions of candidates that were identified as belonging to the unspeeded class. The Science tests showed the lowest propensity for speededness, with ability estimations showing little evidence of decline at the end of the test. Higher proportions of candidates taking the Mathematics tests showed a decline in ability estimations consistent with running out of time.

Table 4.3: The proportions of candidates that were identified as 'unspeeded'

Test	Level	Designated Unspeeded	Designated Speeded	Unspeeded Class	Total (all classes)	Proportion unspeeded
Biology	F	17	4	1000	1000	1.00
Chemistry	F	17	4	621	1000	0.62
Physics	F	17	4	858	1000	0.86
Maths 1	F	12	6	558	1000	0.56
Maths 2	F	12	6	457	1000	0.46

Tables 4.4 to 4.7 show the numbers and the mean ability estimates for each of the (un)speeded classes. For the Mathematics tests the ability declines in line with the speeded class with only one exception. This is to be expected for two reasons. Firstly, candidates have to show a certain level of ability for a decline in ability estimation throughout the test to be perceptible. Secondly, these ability estimates are determined by candidates' performance on the whole test. The true ability estimate of the speeded candidates will be higher than suggested here. Figures 4.2 and 4.3 show the ability estimates calibrated on their responses to the first 20 items. These figures suggest that only the most able candidates can complete Paper 1 in time, while candidates of all ability struggle to complete Paper 2 in time.

For Biology, no candidates appear to be under time pressure. For the other two Science tests the majority of the candidates show no decline in their ability estimation on the designated speeded portion of the test. For these tests there are a small proportion of able candidates who appear to run out of time or energy.

Table 4.4: Mean ability by speeded class: Mathematics Paper 1 foundation tier

Class	n	theta
1	558	0.33
2	17	-0.58
3	36	-0.53
4	84	-0.94
5	180	-1.09
6	125	-1.14
7	0	NA

Table 4.5: Mean ability by speeded class: Mathematics Paper 2 foundation tier

Class	n	theta
1	457	0.28
2	22	0.05
3	30	-0.13
4	24	-0.12
5	13	-0.63
6	434	-0.96
7	20	-0.19

Table 4.6: Mean ability by speeded class: Physics foundation tier

class	n	theta
1	858	0.71
2	36	-0.81
3	26	-0.61
4	80	0.15
5	0	NA

Table 4.7: Mean ability by speeded class: Chemistry foundation tier

class	n	theta
1	621	0.20
2	127	-0.55
3	193	-0.13
4	38	0.53
5	21	0.87

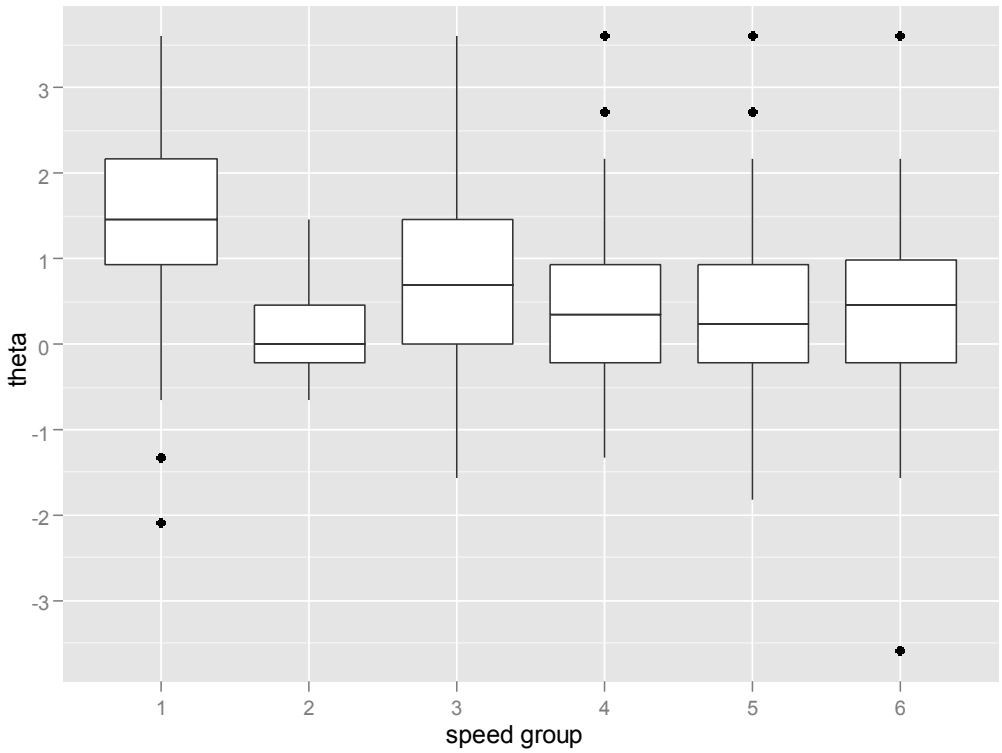


Figure 4.2: Ability estimations on the first 20 items across speeded classes: Mathematics Paper 1 foundation tier

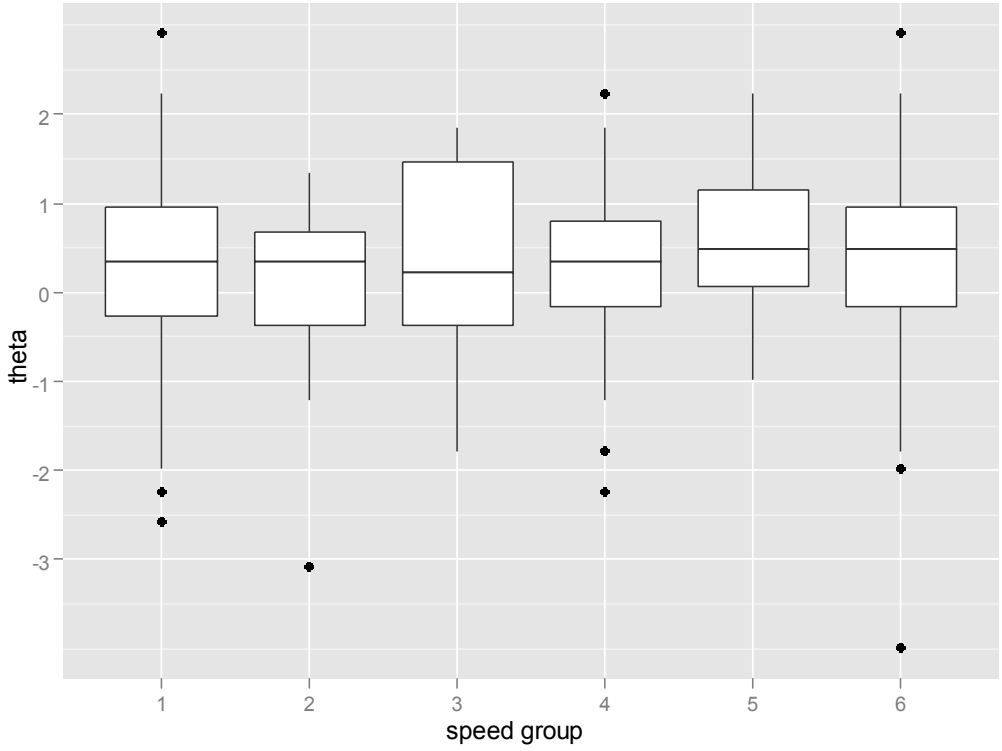


Figure 4.3: Ability estimations on the first 20 items across speeded classes: Mathematics Paper 2 foundation tier

The item parameter estimations by class are shown in tables 4.8 to 4.11. These estimations are drawn from the posterior distribution of the class estimations and are therefore based on larger samples than the summary of the median values of the classes in tables 4.4 to 4.7 suggest.

Extreme values for a number of the Mathematics Paper 1 items in their speeded condition are suggestive of a high proportion of missing data. Candidates appear to be able to solve the first designated speeded question, 6 from the end in its speeded condition, but from thereon in they are very unlikely to get a question correct in a speeded condition. For Mathematics Paper 2, the values do not decline in line with their speeded condition. This suggests that candidates running out of time do not finish the test in a linear fashion, but choose which items to complete. The difficulty parameters for the second, third, fourth and six items in their speeded condition are relatively close to their unspeeded condition. The Science tests share this same erratic pattern.

Overall, therefore, it would seem that a substantial proportion of candidates run out of time on Mathematics Paper 1. Nearly half of the candidates have an unexpectedly low probability of answering the last five questions correctly. On Mathematics Paper 2 a large proportion of the candidates appear to struggle with the relatively difficult penultimate item due to its positioning in the test.

Most of the Science candidates are unaffected by speededness. In Physics a small proportion find 2 out of the last 4 questions disproportionally difficult, while in Chemistry a slightly larger proportion appears to run out of time on 2 out of the last 4 questions.

Without any external information it is difficult to determine the extent to which the item parameter estimation has been purified in its unspeeded condition. The next chapter will consider evidence from the equivalent items on the higher tier and attempt to make some judgement on how much the item parameters can be considered purified. It is also, of course, possible, that the decline in ability estimations could be due to a dimension other than speededness emerging late in the test. This could be an analogous construct such as fatigue, or a change in content area.

In an equating context there are two possible courses of action that could be prompted by these findings. The first course of action is to exclude the items that display considerable different item parameter estimations under their two conditions. This has the disadvantage of reducing the item parameter information available to the equating. The alternative is to select a sample of candidates for the equating based on their speeded class. Only candidates whose ability estimates do not decline towards the equating portion of the test would be used. This has the advantage of using all the available item parameter estimates, but reduces the sample size. The former solution would appear to be most appropriate for Mathematics Paper 1 and the Science tests as a high proportion of candidates are classified as unspeeded, and a high proportion of items show a considerably raised parameter estimation in their speeded condition. A combination could be used for Mathematics Paper 2; candidates in classes 1 and 2 with item 2 removed would offer a relatively stable solution.

Table 4.8: Item parameter estimations across classes: Mathematics Paper 1 foundation tier

Item	1	2	3	4	5	6
Unspeeded	0.27	2.27	2.07	0.57	-0.09	3.81
Speeded	0.53	5.41	60.66	2.83	9.28	11.58

Table 4.9: Item parameter estimations across classes: Mathematics Paper 2 foundation tier

Item	1	2	3	4	5	6
Unspeeded	1.64	1.48	0.95	-0.02	3.40	-0.70
Speeded	26.34	1.77	1.16	0.08	64.89	0.00

Table 4.10: Item parameter estimations across classes: Physics foundation tier

Item	1	2	3	4
Unspeeded	2.15	-0.72	0.57	-0.05
Speeded	2.86	5.61	8.44	0.31

Table 4.11: Item parameter estimations across classes: Chemistry foundation tier

Item	1	2	3	4
Unspeeded	0.83	0.46	1.15	0.07
Speeded	7.02	0.60	5.20	0.56

4.8.2 The testlet model

4.8.2.1 Convergence

Convergence was estimated visually from the sample traces of the parameter estimation. Evidence of stabilisation was taken to mean the absence of any upward or downward trend in the parameter values. It is much easier to say, however, whether convergence has not been achieved than with any certainty that convergence has been achieved (Spiegelhalter et al., 2003)

4.8.2.1.1 The one and two-parameter models

Convergence appeared to have been reached within 500 iterations for the one and two-parameter models. The standard deviation for most of the beta parameters was within about 0.1 while the standard deviation for the theta parameters was generally within 0.5. These levels of precision are similar to those achieved under ML or MML estimation.

4.8.2.1.2 The two-parameter testlet model

Figures 4.4 to 4.6 show the sampling history for three of the parameters for the Chemistry foundation tier paper under the testlet model from iterations 6001 to 7000. For all three parameters the sampling does appear to have converged as there is no upwards or downwards trend remaining. The beta parameters appear to be estimated with reasonable precision, with the beta parameter for item 1, for example, oscillating within a band of -0.8 to -0.6 with a standard deviation of 0.08. The theta parameter for person 1 is less precisely estimated, oscillating within a band of 0 to 2, with a standard deviation of 0.56. The eta parameter for that same person is estimated with the least precision, oscillating within a band of -2 to +1.5, with a standard deviation of 0.91. The same pattern was repeated across items and persons. The Biology paper showed a similar pattern, with poor estimation of the eta parameter. The Physics paper, however, showed more precision in the estimates of the eta parameter, with a standard deviation of the samples closer to 0.7. Further iterations had a small but negligible impact on the precision of estimation.

4.8.2.1.3 The three-parameter model

The three-parameter model took the longest to converge, and even then the evidence of convergence was the weakest of all the models. The sample trace of the beta parameter in Figure 4.8, for example, appears to shows an upward trend suggestive of non-convergence even after 6000 iterations. The pseudo-guessing parameter was also quite volatile (Figure 4.9).

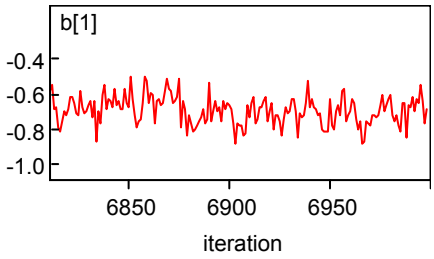


Figure 4.4: Sampling history for testlet parameters in Chemistry foundation tier: The beta parameter for item 1

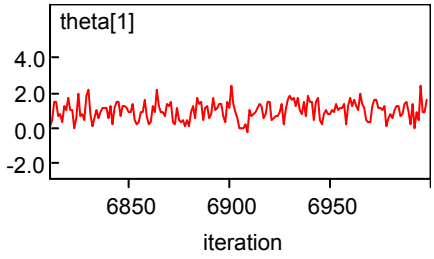


Figure 4.5: Sampling history for testlet parameters in Chemistry foundation tier: The theta parameter for examinee 1

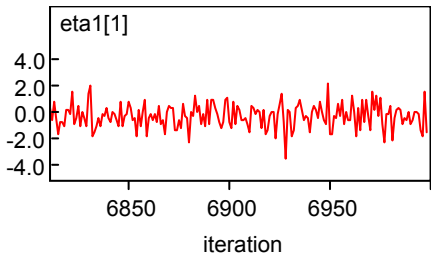


Figure 4.6: Sampling history for testlet parameters in Chemistry foundation tier: The eta parameter for examinee 1 on test section 7

Table 4.12: Means and standard deviations of the samples from Figures 4.3 to 4.5 over iterations 6001 to 7000

Node	mean	sd	2.50%	97.50%
b[1]	-0.71	0.08	-0.87	-0.57
eta1[1]	-0.31	0.91	-2.17	1.43
theta[1]	0.98	0.56	-0.02	2.10

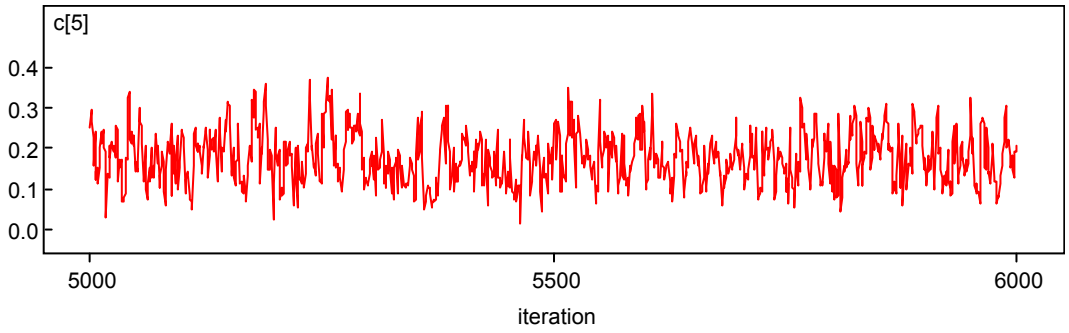


Figure 4.7: The pseudo-guessing parameter over iterations 5000 to 6000

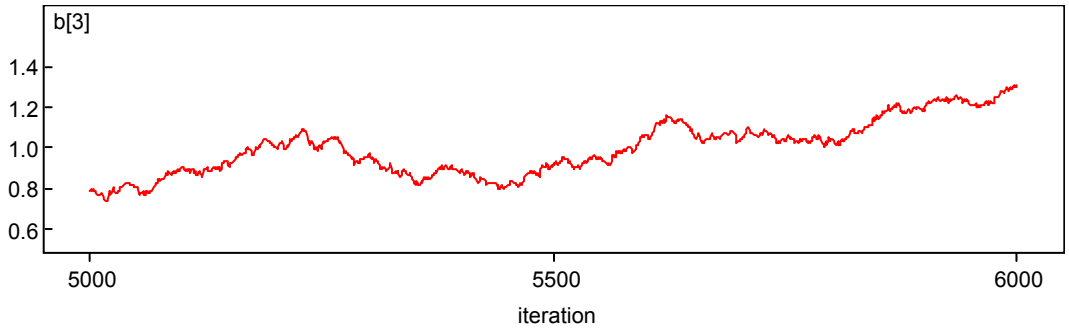


Figure 4.8: The beta parameter for item 3 over iterations 5000 to 6000

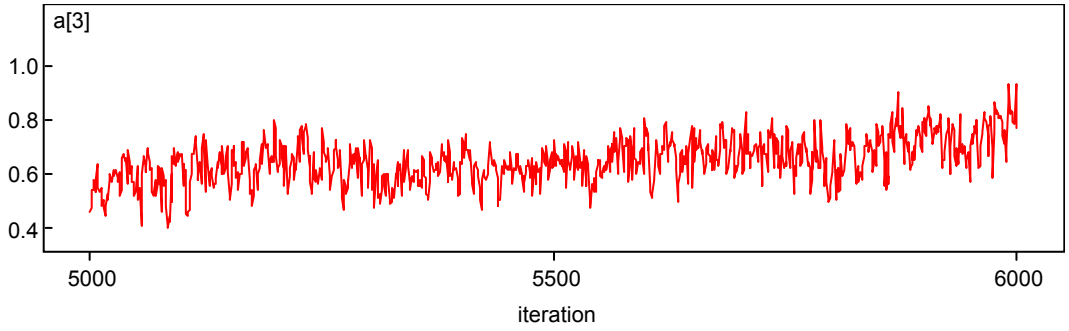


Figure 4.9: The alpha parameter for item 3 over iterations 5000 to 6000

4.8.2.2 The magnitude of testlet effects

Table 4.13 illustrates the magnitude of the testlet effects using the estimated values (posterior means) of $\sigma_\gamma^2$for each of the seven testlets for each of the tests. These values can be compared to $\sigma_\vartheta^2 = 1$, the variance of the ability parameter, to get an order of magnitude. Overall, the testlet effect would seem to be modest, suggesting that the items are conditionally independent. Testlet 1 in the Physics paper would seem to show most evidence of conditional dependence.

Table 4.13: Estimated values (posterior means) of σ_γ^2

Testlet	Physics	Chemistry	Biology
1	0.37	0.25	0.15
2	0.20	0.18	0.22
3	0.26	0.26	0.28
4	0.18	0.23	0.23
5	0.13	0.22	0.21
6	0.16	0.22	0.28
7	0.16	0.21	0.26

4.8.2.3 Deviance Information Criterion (DIC)

The minimum DIC estimates the model that will make the best short-term predictions, in the same spirit as Akaike's criterion (Spiegelhalter et al., 2003). The lower the value, the better the prediction; however, if the difference in DIC is small (less than 5) and the models make very different inferences, then it could be misleading just to report the model with the lowest DIC (Spiegelhalter et al., 2003). The DIC indices (Table 4.14) indicate that the 2 parameter testlet model will make the best short-term predictions for two out of the three tests, while the 2 parameter model will make the best short-term predictions for the other test. The biggest

difference appears to be for the Physics paper. This result is consistent with the more precise model estimations and the higher testlet effect observed. The worst model of all appears to be the 3 parameter model. This may be related to the poor estimation of the pseudo-guessing parameter and the lack of strong evidence of convergence noted above. It was noted in Chapter 3 that these multiple-choice items are relatively easy. The pseudo-guessing parameter is very difficult to estimate under these conditions (Bock & Moustaki, 2007).

Table 4.14: Deviance Information Criterion (DIC)

	1PL	2PL	2PLT	3PL
Biology	33383	33226	**32967**	35952
Chemistry	34638	**34221**	34267	38402
Physics	32132	31809	**31403**	35442

4.8.2.4 Posterior Predictive Model Checking (PPMC)

While the summary statistics such as the magnitude of the testlet effect and the DIC are of interest, and useful in guiding model selection, they are of limited diagnostic use. They do not reveal, for example, why the testlet model is most appropriate for Physics or which items display conditional dependence. For this more detailed diagnostic information PPMC methods are useful.

4.8.2.4.1 Observed score distributions

The analysis in Chapter 3 suggested that the observed score distributions for the GCSE Science tests were not accurately predicted under the Rasch model. In the frequentist framework the deviation can be measured using a chi-square statistic (Béguin & Glas, 2001); however, the model predicted deviations from the observed

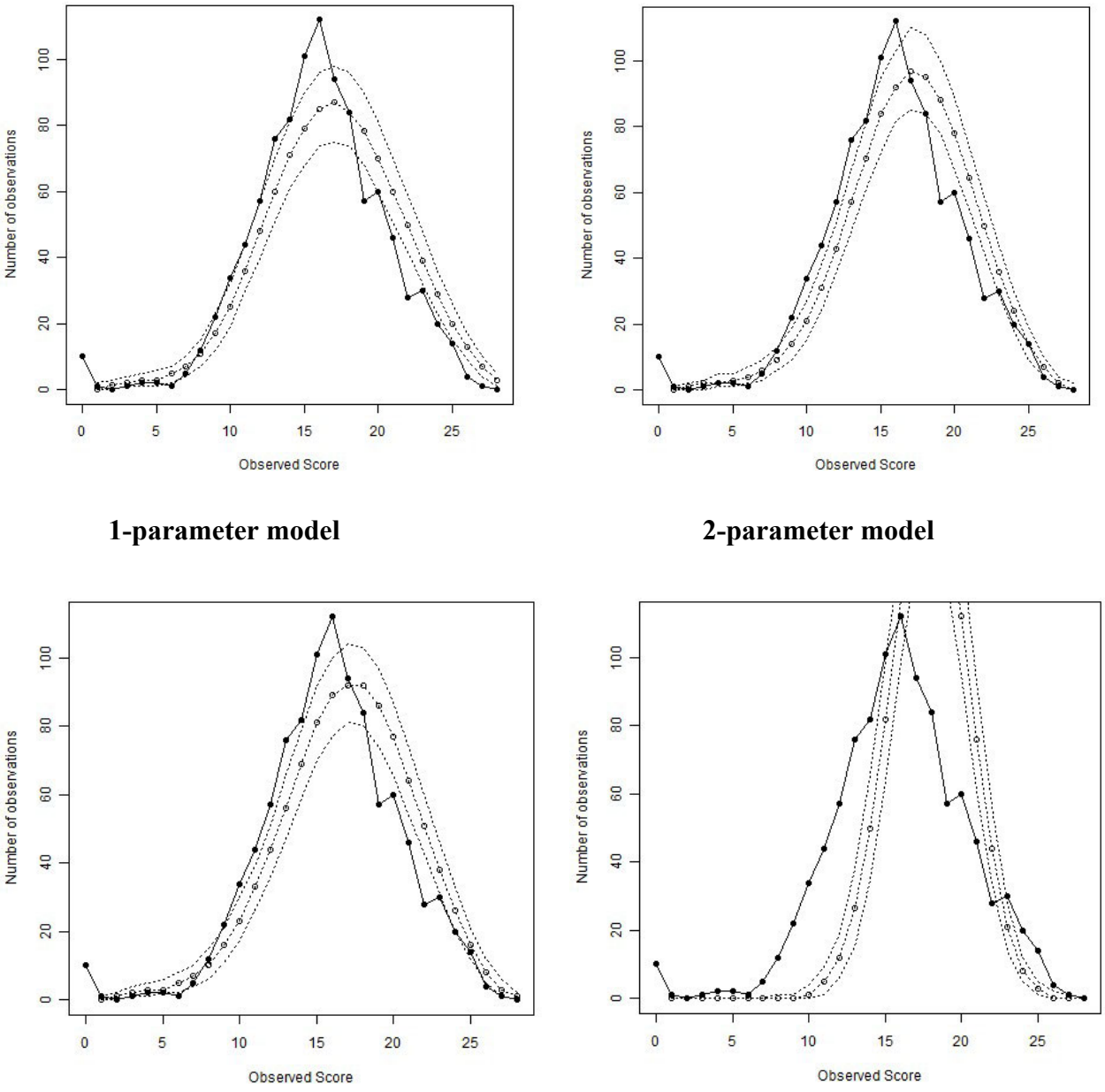
distributions may not follow a chi-square distribution (Sinharay et al., 2006). The replications produced in the Bayesian framework provide the relevant frequency distribution. Figure 4.10 illustrates the observed score distributions for the Physics test against the model replicated distributions. The observed score distribution is represented as the dark solid line.

The three-parameter model clearly provides the worst predictions for the observed score distribution. The presence of guessing to the extent predicted by the model creates an acute peak in the predicted distributions. At the other extreme, the one-parameter model under-predicts the density of the scores around the mean. The two-parameter models provide a better prediction of the density, but the location of the mean is misplaced. The difference between the score distributions predicted by the two-parameter models is interesting. The testlet model predicts a less acute peak than the standard two-parameter model. The testlet parameter appears to have a dampening effect on the discrimination parameter; this is not intuitive.

4.8.2.4.2 Point biserial correlations

The correlation of examinee scores with the binary outcomes on a particular item, the point biserial correlation coefficient, has been used to show the inadequacy of the Rasch models due to its lack of a discrimination parameter (Albert & Ghosh, 2000; Sinharay et al., 2006). Data simulated from the two-parameter model will have more extreme point biserial correlation coefficients than predicted by the Rasch model (Sinharay et al., 2006). Figure 4.11 highlights with a solid dot the observed correlations against 100 of the replicated datasets for the 1-pl model and the 2-pl model for the Physics test. For the 1-pl model the observed correlations are often at the extremes of the replicated correlations, and in several cases, beyond the extreme.

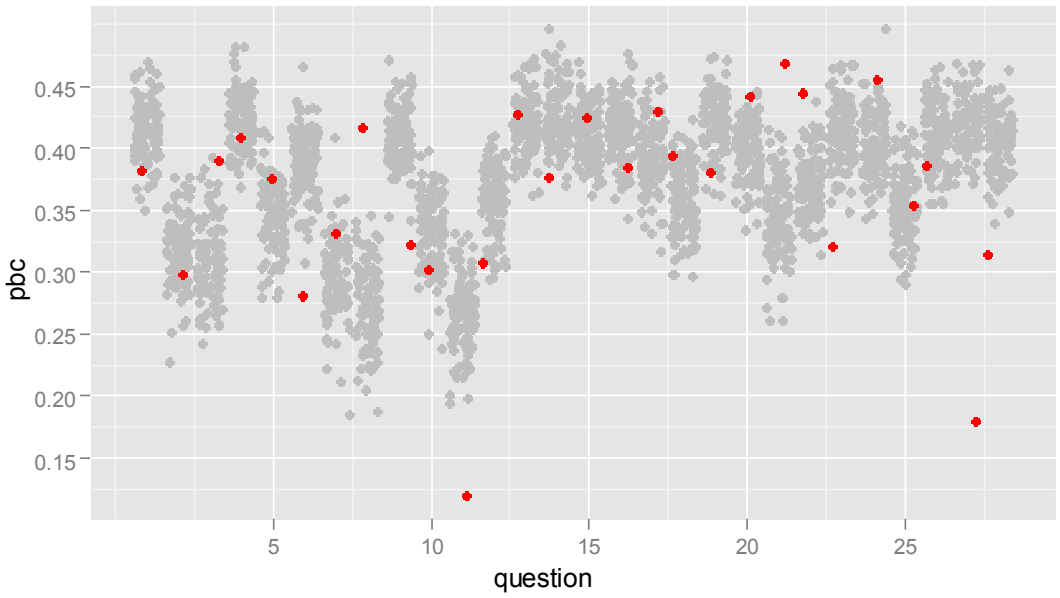
The 2-pl model accurately models the correlations. This pattern is consistent across the three tests.



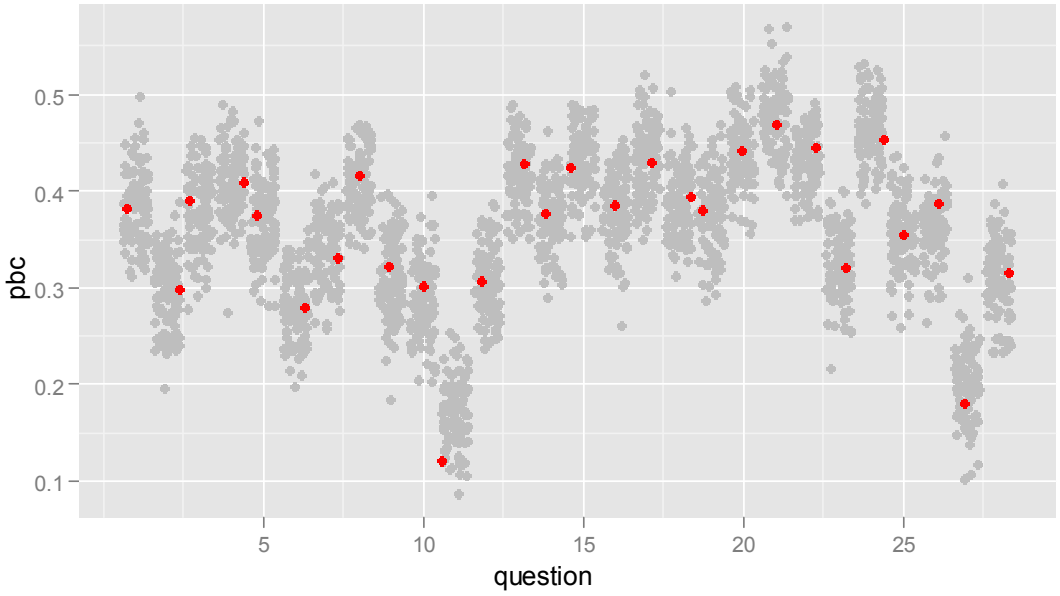
1-parameter model **2-parameter model**

2-parameter testlet model **3-parameter model**

Figure 4.10: Observed score distributions against model predicted distributions



The 1-pl model



The 2-pl model

Figure 4.11: Point-biserial correlations

4.8.2.4.3 Odds ratios (OR)

Unidimensional IRT models require local independence between items. The odds ratio (Agresti, 2002) measures the associations between item pairs in terms of the observed and expected ratios of: correct – correct, correct – incorrect, incorrect – correct and incorrect – incorrect. If local independence does not hold between two items then the observed OR will be larger or smaller than expected under unidimensional IRT models (Chen & Thissen, 1997). Chen and Thissen found the standardized log-OR not to have a $N(0,1)$ null distribution so once again the PPMC method is useful as it provides a relevant frequency distribution.

The testlet response model is designed to allow the relaxation of local independence between items in testlets. The first section within the Physics test showed the greatest magnitude of a testlet effect. The odds ratios are a useful way of understanding why this testlet effect occurs and how the testlet response model deals with it.

Tables 4.15 and 4.16 illustrate the odds ratios for the first two questions of the first Physics section under the two-parameter model and the two-parameter testlet model. The observed odds show the odds of repeating response patterns within the observed data for pairs of items (correct-correct, incorrect-incorrect, correct-incorrect, incorrect-correct). A high value reveals that the odds of repeating response patterns are high. So, the odds ratio of 8.42 between the question pair of question one and two shows that there is a high likelihood that candidates will have the same pattern of responses on both (0,0 or 0,1 or 1,0 or 1,1). Obviously, if both questions are particularly easy then the odds would be expected to be high (most candidates would get 1). In itself therefore, a high odds ratio is not indicative of loss

of local independence. The PPMC methods provide a relevant frequency distribution against which to judge whether local independence has been lost.

Under PPMC, therefore, replications are made under the relevant model. The first 10 replications are given in the table. It is clear, even from the first 10 replications, that the odds ratio between question one and two is unusually high compared to the odds ratio expected under a two-parameter model, which assumes local independence. In 1000 replications none were higher than the observed odds ratio. This leads to a PPP value of zero. A similar pattern is observed between questions one and three. The odds ratio between questions one and four is more suggestive of local independence as 59 per cent of the replications produced a higher odds ratio than that observed. This implies that the model is sampling around the observed value.

The two-parameter testlet model relaxes the requirement for local independence within testlets by allowing a testlet parameter to explain some of the variance of responses within each testlet. This leads to higher odds ratios in the replicated data between questions within testlets than observed under the 2-parameter IRT model. This can be observed in the final three columns of Table 4.15. For the question pair one and two, the replications under this model are now more successful in reproducing the odds ratios. 24 per cent of the replications of the replicated odds ratios are now higher than the observed odds ratios for this question pair. For the question pair one and three the model predictions are now sampling around the observed value, with a PPP value close to 0.5. For the question pair one and four, however, the model predictions are now higher than the actual odds ratios in the observed data. This is because the testlet parameter is assumed to be constant across all items in a section. While the first three items appear to share this testlet

parameter, the fourth item does not. The testlet parameter therefore degrades the model predictions for the relationship between the first three items and the fourth item. The second example, given in Table 4.16, shows how the testlet model leads to a similar degradation in the prediction of the odds ratios for the question pair two and four.

The general increase in odds ratios within testlets under the testlet model is shown in Figure 4.12. Under the two-parameter model the odds ratios are generally under predicted. This suggests the presence of conditional dependence. Under the two-parameter testlet model the odds ratios for those questions that did not display conditional dependence under the two-parameter model are generally over predicted. The two-parameter model produces better predictions for questions which are locally independent within testlets, while the two-parameter testlet model produces better predictions for questions which are conditionally dependent within the testlet.

Clearly these results are valuable to test designers. If the testlet design is deliberate and intended to enhance validity then loss of local independence may be sacrificed for the purpose of that validity. If a testlet model fits well that may be taken as evidence that each testlet is coherent. In this context, items that display local independence may be unrelated to the stimulus, which is indicative of poor test design (Pollitt, 1985). Candidates may seek to find some connection between the question and the stimulus which may lose them time or result in confusion. Of course the local independence may be due to a valid source of difficulty such as the introduction of a mathematical element or specific knowledge element to the test. Only inspection of the test items can reveal whether the local independence is justified.

Further, successful modelling under the testlet model is no guarantee of the quality of the items. Local independence may be violated in ways which are invalid. The most obvious example of this is known as cross information where the answer to one question is given in a subsequent question. Unexpected conditional dependence can therefore be as informative as unexpected local independence.

If the primary goal is to produce valid predictions from a set of given data then the fit of the testlet model could be improved by knowledge of which items are conditionally dependent. In this example, the testlet for the first section would be defined as the first three items, with the fourth item being modelled as independent.

Overall, the evidence suggests that the testlet model could produce some improvement in model predictions for tests with an explicit testlet design. This will impact on the predicted score distributions which could have practical implications. The PPMC reveals, however, that the greatest gain appears to be derived from moving from a one-parameter to a two-parameter model as the two-parameter model correctly predicts item point biserial correlations.

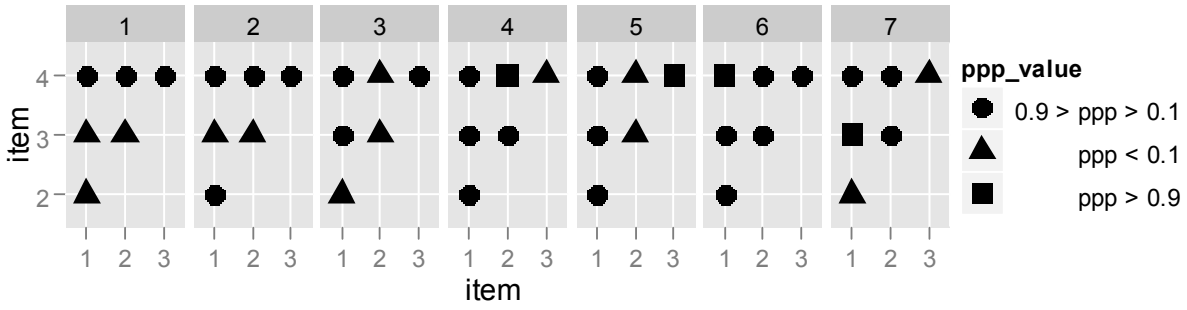
Table 4.15: PPP values for odds ratios under the 2 parameter and 2 parameter testlet models for Question 1 in Physics

		2 parameter Question			2 parameter testlet Question		
		2	3	4	2	3	4
Question 1	Observed odds	8.42	5.52	1.75	8.42	5.52	1.75
Replications	[1,]	3.07	2.74	2.31	5.41	4.44	1.89
	[2,]	2.62	1.66	2.16	6.41	7.18	2.08
	[3,]	3.13	2.61	1.68	6.21	5.26	1.82
	[4,]	3.24	2.41	2.47	8.92	4.36	2.66
	[5,]	2.51	2.05	1.68	7.21	5.99	2.04
	[6,]	2.93	2.46	1.51	7.70	4.44	2.75
	[7,]	2.64	1.99	1.80	5.60	6.29	1.76
	[8,]	2.41	1.81	1.70	5.64	7.16	2.18
	[9,]	1.83	1.93	1.70	10.28	5.70	1.84
	[10,]	1.94	1.77	1.30	5.70	7.18	2.03
		…	…	…	…	…	…
Proportion of replications above observed odds		0.00	0.00	0.59	0.24	0.47	0.80

Table 4.16: PPP values for odds ratios under the 2 parameter and 2 parameter testlet models for Question 2 in Physics

		2 parameter			2 parameter testlet		
		Question			Question		
		1	3	4	1	3	4
Question 2	Observed odds	8.42	7.02	1.55	8.42	7.02	1.55
Replications	[1,]	3.07	2.83	2.08	5.41	5.58	2.47
	[2,]	2.62	2.62	1.67	6.41	8.30	2.33
	[3,]	3.13	3.34	2.08	6.21	5.16	1.95
	[4,]	3.24	2.02	2.08	8.92	6.03	2.41
	[5,]	2.51	2.01	1.88	7.21	5.89	2.05
	[6,]	2.93	2.19	1.60	7.70	4.33	2.44
	[7,]	2.64	2.20	1.90	5.60	7.08	2.00
	[8,]	2.41	2.24	1.73	5.64	5.25	2.30
	[9,]	1.83	2.40	2.24	10.28	7.97	2.43
	[10,]	1.94	1.90	1.87	5.70	6.42	3.17
		…	…	…	…	…	…
Proportion of replications above observed		0	0	0.88	0.24	0.34	0.97

2-parameter model



2-parameter testlet model

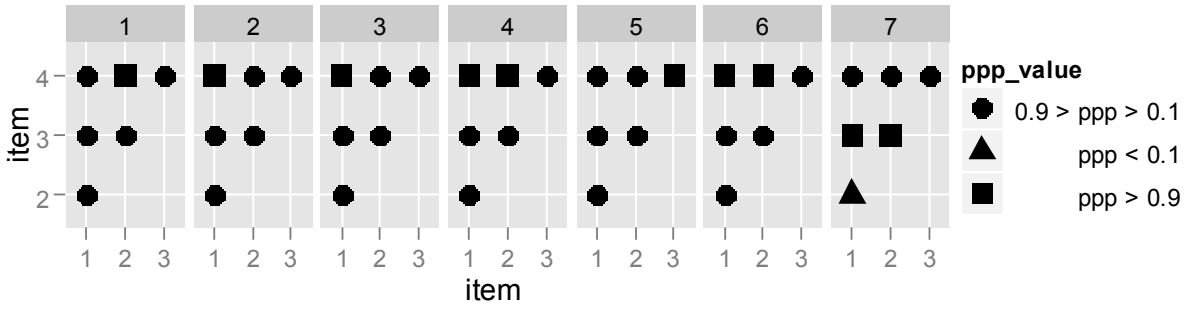


Figure 4.12 Odds ratio

4.9 Discussion

The purpose of this section of was to attempt to fit different models to the GCSE test data, to examine their fit, and review the implications suggested by these models. In order to increase understanding of the dimensionality that was highlighted by the analyses in Chapter 3, the Multi-Class Mixture Rasch Model (MMRM) for Test Speededness was fitted to explore whether some part of the dimensionality was due to time pressure affecting responses towards the end of certain tests. The two-parameter IRT model was fitted to improve the fit for tests where the item discrimination showed great variation. The three-parameter IRT model was fitted to

investigate guessing. Finally, the two-parameter Testlet Response Theory model was fitted to investigate weak local independence between responses. DIC and PPMC checks were then undertaken to examine how well the models performed in short term predictions.

Results from the MMRM suggested that one of the GCSE Mathematics tests is taken under time pressure. Estimations of the ability of candidates on the last six items of this test declined relative to estimations of the ability of candidates on the first 12 items for a large proportion of the cohort. If this is a general pattern for this test it represents construct irrelevant variance and should be dealt with. Candidates should be allowed more time or the number of items they are expected to complete should be reduced.

Even in those tests where the evidence for candidates working under time pressure was less compelling, the MMRM analysis revealed that the item parameters of certain items towards the end of a test are overestimated. The latent class approach of the MMRM allows an unspeeded cohort to be identified. This cohort can then be used to purify item parameters.

The MMRM analysis is however dependent on certain assumptions. It assumes that candidates take tests in a linear fashion when this is unlikely to be the case under time pressure. It also requires an arbitrary designation of which items are likely to be speeded. Nevertheless, it seems to offer a relatively intuitive summary of the pattern of responses that occur towards the end of a test. Unfortunately, the item parameters for the tests studied here are not available in any purified form so these results are difficult to validate. As some of the items are used on the higher tier forms of these tests, and appear towards the beginning of these forms, it may be

possible, in the next chapter, to make some judgement as to whether the parameters have indeed been purified through an MMRM approach.

DIC indices and PPMC checks suggested that the two-parameter IRT model would provide better short term predictions than a one-parameter model for the tests studied here. The two-parameter model preserves the item point biserial correlations observed in the data. The DIC indices and PPMC checks suggested that the three-parameter IRT model would produce degraded predictions, and greatly over-predict the peak in the distribution of scores around the mean. This may be due to poor convergence of the model or poor identification of the pseudo-guessing parameter.

The Testlet Response Theory model showed varying degrees of success on the three GCSE tests studied here. The odds ratios of some of the items showed patterns indicative of weak local independence within each testlet. The testlet model was able to reproduce these patterns, but in applying a testlet parameter across each testlet, odds ratios for those items within the testlet that were locally independent were over predicted.

The results from the Testlet Response Theory model are clearly useful to test designers and in any study of validity. Unexpected local independence or local dependence can both reveal elements of construct irrelevant difficulty. While coherent testlets, as indicated by successfully fitting a Testlet Response Theory model, can be used to defend the validity of a test, weak local independence may be caused by construct irrelevant easiness such as cross information. The model fitting can only support an inspection of the actual test, not replace it.

The PPMC methods showed great power in identifying aspects of the observed data that were preserved by the models. Unlike the results from statistical tests of fit they lend themselves to intuitive graphical interpretation. The simplicity

of their interpretation, however, belies the complexity of their calculation. The results reported here required the use of three different programming languages: R, WinBUGS and C++. For this reason, despite their power and appeal, they are unlikely to be used by the casual researcher. The more complex models under the MCMC framework are relatively easy to estimate; however they can be extremely time-consuming. This in itself is a recommendation for the simplest possible model that successfully produces the prediction of interest.

While the PPMC methods revealed some practical aspects of misfit, for example, on predicted summed score frequency distributions, the full impact of the use of different models which display different levels of fit and have different levels of success in their various predictions can only be assessed in context of particular applications of those models. The next two chapters will therefore attempt to assess whether the fit of models analysed in these last two chapters is good enough for IRT methods of test equating to make a contribution to the maintenance of standards in public examinations in England.

5. Vertical Test Equating

5.1 Overview

Two findings seem worth pursuing from the previous chapters. Analysis from these chapters suggests that modelling the discrimination of items discretely can improve the fit of IRT models and produce better short term predictions. Results from a mixed Rasch model, the MMRM, suggested that item parameters towards the end of certain GCSE tests were overestimated as candidates were under time pressure. The MMRM can be used to purify the item parameters from this nuisance factor. The analyses so far, however, have not been able to estimate the practical consequences of using one model or another or item parameters that have been poorly estimated. The question could be re-phrased as follows: how much difference does it make in practice if a Rasch model is used as the basis for test equating with little care paid to the quality of the item parameters? Obviously, the answer will depend on specific features of the tests being modelled; but should the differences appear generally small then it would seem preferable to favour the simplest model that gives about the right result in most cases. As equating would never be done in isolation from other statistical indicators, these indicators could be relied upon to highlight potential problems, and the need for more complex modelling.

The practical scenario in which this question will be asked is in equating the tiers of GCSEs. The foundation and higher tiers of tiered GCSEs share common grades. These grades are intended to be equivalent, so a grade C on foundation tier Mathematics is intended to have the same currency as a grade C on higher tier Mathematics. No distinction is made when these grades are reported. Grade C in the

GCSE still carries the important connotation that it represents a pass. It is therefore critical that this equivalence is maintained. Technically, it is extremely difficult to establish or maintain this equivalence, however. This is an area, therefore, in which test equating potentially has a great deal to offer.

5.2 What is tiering?

A tiered GCSE is a GCSE that is available at more than one level, or tier, of difficulty. Currently, only two tiers are available for GCSE: foundation tier, which is of lower demand; and higher tier, which is of higher demand. Candidates who take the foundation tier have access to lower grades, while candidates who take the higher tier have access to higher grades (Figure 5.1). Tiering was introduced primarily to ensure that GCSEs could discriminate across a wide range of ability, and can in this regard be compared to multi-stage tests (Wheadon & Béguin, 2010).

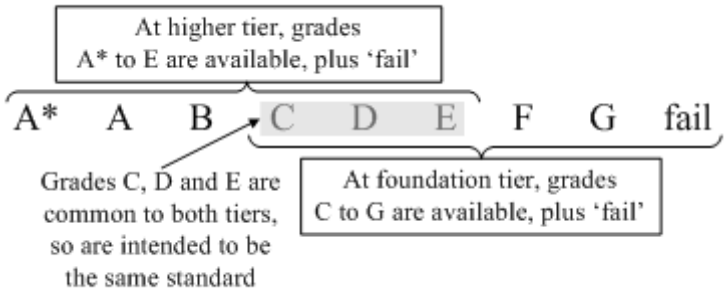


Figure 5.1: The GCSE grades available to different tiers of entry

According to the parallel drawn with multi-stage tests, the first stage of a test is a teacher's judgement on how well a candidate is likely to perform while the second is the test itself. The teacher judgement routes a candidate to a level (Wheadon & Béguin, 2010). Where this metaphor breaks down, however, is that in a

multi-stage test all levels are accessible to a candidate taking a test on the day; that is not necessarily the case with tiering. While candidates can choose on the day which tier they will take, it is possible that they won't have been prepared for some of the material on one or other tier.

Tiering is designed to ensure that the level of demand on candidates is appropriate. As Pollitt, Ahmed, and Crisp (2007) have clarified, however, it is not always clear in the context of assessment whether differing demand relates to the level of difficulty of items or to the cognitive demands of the syllabus. This distinction is pertinent to the technical question of tiering: if candidates on different tiers have studied syllabuses of different demand or are expected to progress from one tier to the next then vertical scaling may be a more appropriate framework for tiering than multi-stage testing. If the candidates have followed the same syllabus then the comparison with multi-stage testing, where items differ in difficulty alone, holds. Whether tiers are differentiated by syllabus content or item difficulty affects the interpretations of the outcome from any comparability study, whether it uses incumbent or equating procedures.

5.3 Tiers with different syllabus content

As is often the case in assessment in the UK, the answer to the question of whether different demand relates to different content or different item difficulty varies by subject. The reasons for the differences are to be found in the educational and assessment history leading up to the inception of the GCSE in 1988. The original GCSE criteria stated that:

> All examinations must be designed in such a way as to ensure proper discrimination so that candidates across the ability range are given opportunities to demonstrate their knowledge, abilities, and achievements – that is, to show what they know, understand and can do. Differentiated papers or differentiated questions within papers will be required accordingly in all subjects (cited in Good & Cresswell, 1988a, p. 2).

The statement does not clarify whether the papers should be differentiated by content or by difficulty or both.

In mathematics, the interpretation was placed quite squarely with the need for a differentiated syllabus. The Cockcroft report (1982) was an influential state of the nation report that argued the case for reform in mathematics assessment and laid some part of the foundation for the GCSE. On differentiation Cockcroft reported on the failures of the incumbent system to differentiated by content, a failure that had dire educational consequences,

> Examiners have a duty to set papers which cover as much of the syllabus as possible. Because they are aware that many low-attaining candidates will attempt the papers, they feel obliged to include within them a number of trivial questions on those topics in the syllabus which are conceptually difficult so that low-attaining candidates may find some questions which they are able to attempt. Teachers in their turn feel obliged to cover as much of the syllabus as possible so that their low-attaining pupils may be able to answer such questions, even

> though some of the topics which are included are conceptually too difficult for these pupils. This leads to teaching of a kind which, instead of developing understanding, concentrates on the drilling of routines in order to answer examination questions. We therefore have a 'vicious circle' which is difficult to break (Cockcroft, 1982, para. 445).

This sentiment was echoed in the white paper which signalled the intention of the government to introduce a single system of assessment. It stated that in some subjects such as mathematics or foreign languages,

> certain concepts are within the grasp of some candidates but beyond the reach of others (cited in Cockcroft, 1982, para. 520)

Mathematics is a clear example of a subject that has, since the introduction of a single system of assessment in 1988, differentiated by syllabus content. In the AQA GCSE Specification this is made quite clear,

> the subject content unique to the Foundation tier is based on the Foundation Programme of Study; the subject content common to both tiers and of the Higher tier only is based on the Higher Programme of Study; in general, the Higher tier content of the specification subsumes the Foundation tier content. (Assessment and Qualifications Alliance, 2008, p. 18)

As the higher tier subsumes the foundation tier, the structure would appear similar to that of progression through stages of learning over time. The other subjects that have quite clear differentiation by programme of study are modern foreign languages.

5.4 Tiers with the same syllabus content

At the other end of the syllabus spectrum from mathematics in terms of differentiation by content is english language. For english language it was felt that the same questions could be asked of candidates of all abilities; their answers would differentiate them, not the task. This has become known as differentiation by outcome. The current AQA syllabus makes no distinction between the programme of study for the two tiers and both tiers use a common mark scheme (Wheadon, Spalding & Tremain, 2008). Most other subjects lie in between these extremes, but with the exception of mathematics and modern foreign languages, there is no differentiated programme of study. In AQA Science, for example, there is no distinction made in the programme of study, but there is reference to higher level skills required to respond to items on the higher tier question papers. In AQA Geography there is no distinction noted at all, simply a reference to the different grades available. With the exception of mathematics and modern foreign languages, therefore, the differentiation is intended to be on item difficulty rather than subject content.

5.5 Current approaches to maintaining standards across tiers

The current approach to maintaining standards across tiers is based on the weak criterion referencing approach that was described in Chapter 1. Using a combination of judgement and statistics examiners arrive at a recommendation for the grade C boundaries on the higher tier and the foundation tier separately. Some attempt is made to focus the examiners' minds on a common grade C standard by determining the order in which the grades are considered. According to The Code of Practice (Qualifications and Curriculum Authority, 2009) examiners are required to consider grade C on the higher tier directly after considering grade C on the foundation tier. The standard is therefore presumed to be fresh in their minds. Some statistical guidance may also be available on the item facilities gained by candidates on common items on the two tiers. Reference can also be made of course to performance descriptors.

There are substantial problems with this judgemental approach. Firstly, the examiners are making holistic judgements on two papers in which the items differ. According to the Good and Cresswell effect (Good & Cresswell, 1987) examiners will find it easier to reward performance on the relatively easier questions posed on the foundation tier. Secondly, no direct scrutiny or comparison is made of the performance on the common items. To isolate this comparison from the Good and Cresswell effect the comparison of the common items would have to be undertaken in isolation from the other items in the question paper. Apart from the practical difficulties of doing this, the lack of scrutiny of the remaining items in the question paper directly contradicts the underlying purpose of their scrutiny which is to make holistic judgements. The statistical guidance is of little help as there is no way in

which performance on the common items can be extrapolated to performance on the papers as a whole.

Black & Bramley (2008) suggest that a rank-ordering method could be used to solve the between-tier comparability issue. According to this method subject experts put sets of (usually 10) scripts into rank order of perceived quality. The sets of paired comparisons that result can then be analysed by fitting a Rasch model. If this were done on the entire scripts then it would be liable to the Good and Cresswell effect. If it were done on only the common items then the information in the rest of the scripts would be wasted.

Turning to the statistical guidance, for tiering there is the particular issue that statistical models that draw on prior achievement assume that the allocation of candidates to tiers has no effect on their progress towards their outcomes. There is some evidence that students of similar ability achieve higher GCSE grades when they are placed in higher sets (Ireson, Hallam & Hurley, 2005). If this were generally true then candidates with the same level of prior achievement would be expected to achieve different outcomes depending on which set they were entered into. The assumption of equal value-added required by a regression model therefore no longer holds. So called school compositional effects are, however, notoriously hard to pin down (Harker & Tymms, 2004).

It is of course possible to predict outcomes for a tier based on prior results for that tier only, and this is often done in practice. Issues arise, however, when the relative entry between tiers changes. Further, this approach does not solve the problem of how to set the relative standards between tiers in the first place or how to evaluate whether the relative standard is correct.

Given the judgemental and statistical limitations inherent in current approaches to comparability between tiers it is hardly surprising that a thorough review of tiering (Baird et al., 2001) found a number of areas of concern. On the issue of comparability the study suggested that potentially large differences in standards existed in some subjects at the overlapping grade C at GCSE. This study used general linear models and predictive methodologies that are, however, constrained for the reasons noted above. A subsequent study using OPLM methods of test equating (Wheadon & Béguin, 2010) found that differences in standards do exist, but these are not as large as suggested by the previous study.

5.6 Potential IRT test equating approaches to tiering

5.6.1 Common item non-equivalent groups design

The ability of the populations on the different tiers obviously differs. This requires the use of a non-equivalent groups design. The performance is linked by means of the common items that are taken by each tier.

5.6.2 Scaling or equating

The problem of maintaining standards across tiers can be conceived of as either a vertical scaling or a vertical equating issue (Kolen & Brennan, 2004). The equating methodology is identical, but the inferences drawn are different. Under vertical scaling equated scores are not considered equivalent. So, according to the mathematics example given above where the syllabuses for the tiers differ, the same equated score is no guarantee that candidates know and can do the same level of mathematics. Under vertical equating the same equated score is intended to represent

the same level of performance. The scores from vertical equated test forms are therefore considered to be equivalent.

5.6.3 Separate or concurrent estimation

Separate estimation sets the ability (theta) scale for the foundation tier as the base scale, and the common items are used to place item parameter estimates, examinee ability estimates, and estimated ability distributions on the base scale using linking methods (such as the Rasch shift constant method described in Chapter 2). All the item parameters and estimated ability distributions are then on the base (foundation) scale. The mean and standard deviation of the estimated ability distributions thus transformed can be used to compare the difference in mean ability and variability on the different tiers.

The alternative to separate estimation is concurrent estimation. Under concurrent estimation items that are not common between tests are effectively coded as missing or not reached. Under MML estimation it is critical that the estimation program allows for multiple groups so that separate ability distributions can be specified for each tier of entry (Kolen & Brennan, 2004). OPLM (Verhelst & Glas, 1995) allows for this distinction to be made.

Kolen and Brennan (2004) suggest that, in theory, concurrent estimation uses all available information for parameter estimation and is therefore expected to be more stable than separate estimation. In practice, however, they recommend separate estimation, as it allows the item parameter estimates from the separate estimations to be compared. This comparison can reveal items that are performing differently across levels and are not suitable as linking items. Where a population parameter is included in the concurrent estimation, as it is under OPLM, this comparison is also

readily made. The authors also note that multi-dimensionality is likely to be more of an issue if different levels of achievement are calibrated concurrently. Separate calibration is therefore characterised as the safer option.

5.7 Potential issues in equating across tiers

5.7.1 Groups of different ability

Although vertical test equating is designed to measure the difference in the ability between groups, when the difference in the ability is too large, the results from different equating procedures can differ. This has been taken to indicate that the equating is no longer robust (Dorans, Pommerich & Holland, 2007; Kolen & Brennan, 2004). Kolen and Brennan's recommendations on the points at which the methods differ are reproduced in Figure 5.2. In addition, it is noted that ratios of group standard deviations on the common items of less than 0.8 or greater than 1.2 tend to be associated with substantial differences among methods (Kolen & Brennan). It is acknowledged, however, that in these situations IRT methods might function more adequately than other methods. Why the results differ is, however, unclear. Cook and Paterson (1987) suggest that poor correlations between item difficulties for two separate groups under these conditions can indicate that the constructs being assessed are different.

Mean group differences (standard deviation unit on the common items)	*Diagnosis*
.1 or less	Few problems for any equating method
.3	Substantial differences between methods
.5	Especially troublesome

Figure 5.2 Diagnosing differences in group means on common items

As there is often no alternative but to equate groups that differ quite markedly in ability, particularly for tests that are administered at different points in the year, some effort has been expended in attempting to produce more stable results between procedures. One approach that has been attempted is to use matching procedures to make otherwise more disparate groups more similar (Dorans, 1990; Eignor, Stocking & Cook, 1990; Kolen, 1990; Lawrence & Dorans, 1990; Livingston, Dorans & Wright, 1990; Schmitt, Cook, Dorans & Eignor, 1990; Skaggs, 1990; Wise, Plake & Mitchell, 1990). While results are generally inconclusive, one study found that matching groups based on their results on the common items provided greater agreement among the results of the various equating procedures studied than were obtained under representative sampling (Schmitt et al.). This rather weak finding has led to matching largely being discontinued.

5.7.2 Disordered thresholds

A second issue for equating in the context of examinations in England is the presence of disordered thresholds which can occur when certain categories on a polytomous item attract few responses. High-stakes tests in England tend to use open response formats. These can be relatively short scales with clearly specified mark

schemes or they can be longer, with impressionistic mark schemes. Even among the shorter scales the proportion of responses observed in some response categories can be very low.

Chapter 3 illustrated that in Mathematics a mark may be available for remembering to add the correct units to an answer. As very few candidates forget to do this very few candidates score one mark less than the maximum. This response pattern can cause difficulty in estimating IRT parameters. If there are no observations in a particular category then the estimation may fail. If there are few observations then the category will never be modal, i.e. it will never be the most likely outcome at any point along the ability scale. Disordered thresholds are not necessarily a problem for measurement; they simply correspond to a very narrow discrimination of ability (Linacre, 2004b). For equating, however, disordered thresholds can be unstable between estimations. For this reason they are typically collapsed for equating (see, for example, Lundgren-Nilsson, Tennant, Grimby & Sunnerhagen, 2006) or item calibrations are used instead (Linacre, 2004a). Using item calibrations in equating reduces the information available for the equating, however, as the discrimination amongst categories that are not disordered is also lost (Linacre, 2004b). The potential improvements or practical differences in test equating derived from collapsing categories has not, however, been reported.

5.7.3 Evaluating test equating quality

Different quality measures are appropriate for separate and concurrent estimation, so they will be addressed here in turn.

5.7.4 Quality measures for separate estimation

Some quality measures suggested for separate estimation are as follows.

5.7.4.1 Gradient of line of best fit

For separate estimation to work in equating the gradient of the line of best fit drawn between item difficulty parameters should be close to 1. If it differs from 1 then a different result would occur dependent on which test was used as the base of the equating. This violates one of the essential conditions of equating, the symmetry requirement. The equating transformation for mapping the scores of Y to those of X should be the inverse of the equating transformation for mapping the scores of X to those of Y (Kolen & Brennan, 2004).

5.7.4.2 Item between-link fit and item within-link fit

Wright and Stone (1999) suggest two Rasch statistics that can be used to evaluate the quality of an equating link. These are measures of between-link fit and within-link fit. Item-within-link fit analysis focuses on the extent to which linking items exhibit adequate fit within the various forms on which they appear. The statistic is:

$$MSF_{IWL} = \frac{\sum_{i=1}^{L}\left(MSF_{ij} + MSF_{ik}\right)}{2L} \tag{5.1}$$

where L is the number of items within a link, MSF_{ij} is the infit mean square for item *i* on form *j*, and MSF_{ik} is the infit mean square for item *i* on form *k*. Under the null hypothesis that items exhibit perfect fit within the link, MSF_{IWL} has an expected value of 1.

Item-between-link-fit focuses on the stability of the item calibrations between forms. It depicts the difference between all linking item parameter estimates from

two of the preliminary forms once those estimates have been placed on the same scale. The statistic is:

$$\chi^2_{IBL} = \frac{\sum_i^n \left(d'_{ik} - d'_{ij}\right)^2}{\sum_i^n w_{ijk}} \qquad (5.2)$$

where $w_{ijk} = \left(se_{ik}^{\ 2} + se_{ij}^{\ 2}\right)$ and the within form item difficulties d_{ik} have been translated to their equated values d'_{ik}. Values substantially greater than 1 indicate that items are performing substantially differently on one form to another. Both of these statistics can be standardised to take account of their expected variance.

5.7.4.3 Concurrent estimation

For concurrent estimation the fit of items as well as global measures of fit can be examined in order to evaluate the quality of the equating. The most useful tool for doing this is an inspection of the Item Response Functions of the common items. These will reveal whether the model fit across responses from both test forms is adequate. Under OPLM poor fit can then be addressed by adjustment of the discrimination parameter.

5.8 Method

5.8.1 Design

This part of the study aimed:

i. to evaluate the quality and compare the results derived from equating a tiered GCSE under the Rasch method and OPLM;

ii. to examine whether the use of a mixture Rasch model, as suggested by the previous chapter, has a substantial impact on results derived from (i).

5.8.1.1 Equating methods

The Rasch method: the item parameters and person parameters were estimated for each form separately under CML and MML respectively using the eRm package (Mair & Hatzinger, 2007). Software written in R by the author then placed the item parameters on the same scale using the shift constant method and calculated expected scores for each group on the separate test forms. These expected scores were then used as the basis for equipercentile equating.

The OPLM method: the OPLM model was used to produce item difficulty and item discrimination parameters concurrently under CML. These parameters were then fixed and passed to the MML estimation of the ability parameters. The number-correct distribution for the test form the candidates did not take can then be estimated from the person and item parameters and used in equipercentile equating of the two test forms.

5.8.1.2 Evaluating the equating

At each stage the quality and outcomes were evaluated under the two different equating methods. To evaluate the quality of the equating, the measures described in section 5.7.4 were used. The outcomes cannot be verified as there are no objective verifications available as to what they should be. Where candidates take a common coursework element, however, linear regression can be used to derive the foundation tier coursework mark at grade C using the relationship between the foundation tier

paper and coursework. The grade C coursework mark that results can then be used to predict the higher tier grade C mark using the relationship between coursework and the higher tier paper. The approach is obviously imperfect as it assumes that there is no interaction between tier of entry and the relationship between coursework and the outcome on the written tests. It also assumes that there is no error in the coursework marks. The results from the equating can also be compared against the marks that were actually awarded. This provides no objective evaluation, but it gives some indication of the degree of how likely any outcome can be defended in practice.

After initial equating any items that displayed obvious differential functioning between test forms were examined. They were checked to ensure that no clerical errors had occurred in determining the relative position of the item in the data sets, and to ensure that in both presentation and the mark scheme they were identical. If this was not the case they were then removed from further analysis.

5.8.1.3 Collapsing categories

Two possible approaches were considered when developing the program to collapse categories with disordered thresholds. One was to collapse categories with a low proportion of observations. The other was to fit the IRT model and then collapse all categories that displayed disordered thresholds. The first approach was chosen for two reasons. Firstly, if there are no observations in a category then the estimation of the IRT model may fail. Secondly, if categories with disordered thresholds are automatically collapsed then that may obscure issues with the quality of the items. This is particularly an issue for linking items where the quality is of key concern. The disadvantage of collapsing purely based on proportion of observations is that categories with a low proportion of observations that are not disordered may be

collapsed. This reduces the information available to the equating. In practice it is probably advisable to specify a conservative proportion of observations below which categories will be collapsed (a minimum number of 10 observations has been advised (Linacre, 1994)), and then to inspect any disordered thresholds that remain for item quality issues.

There are also different approaches to how categories should be collapsed. The approach chosen is detailed in Figure 5.3. First of all the user specifies a minimum proportion of observations for each category. The algorithm then considers all the mid-categories from the highest to the lowest – by considering only mid-categories it is ensured that more than one category remains for each item. If the proportion of responses in that category is lower than required then all the categories above are reduced by one. An alternative approach could have been taken, of course, to collapse from the bottom upwards which could have a different impact on the category distributions that remain.

If the item is a linking item then the same recoding is undertaken for the linking item. It was found that not collapsing linking items in tandem produced anomalous results. This could have been predicted as the category thresholds for a three mark item cannot be compared to the category thresholds for a two mark item; the more steps there are along the latent variable range covered by the item, the closer those steps will be together.

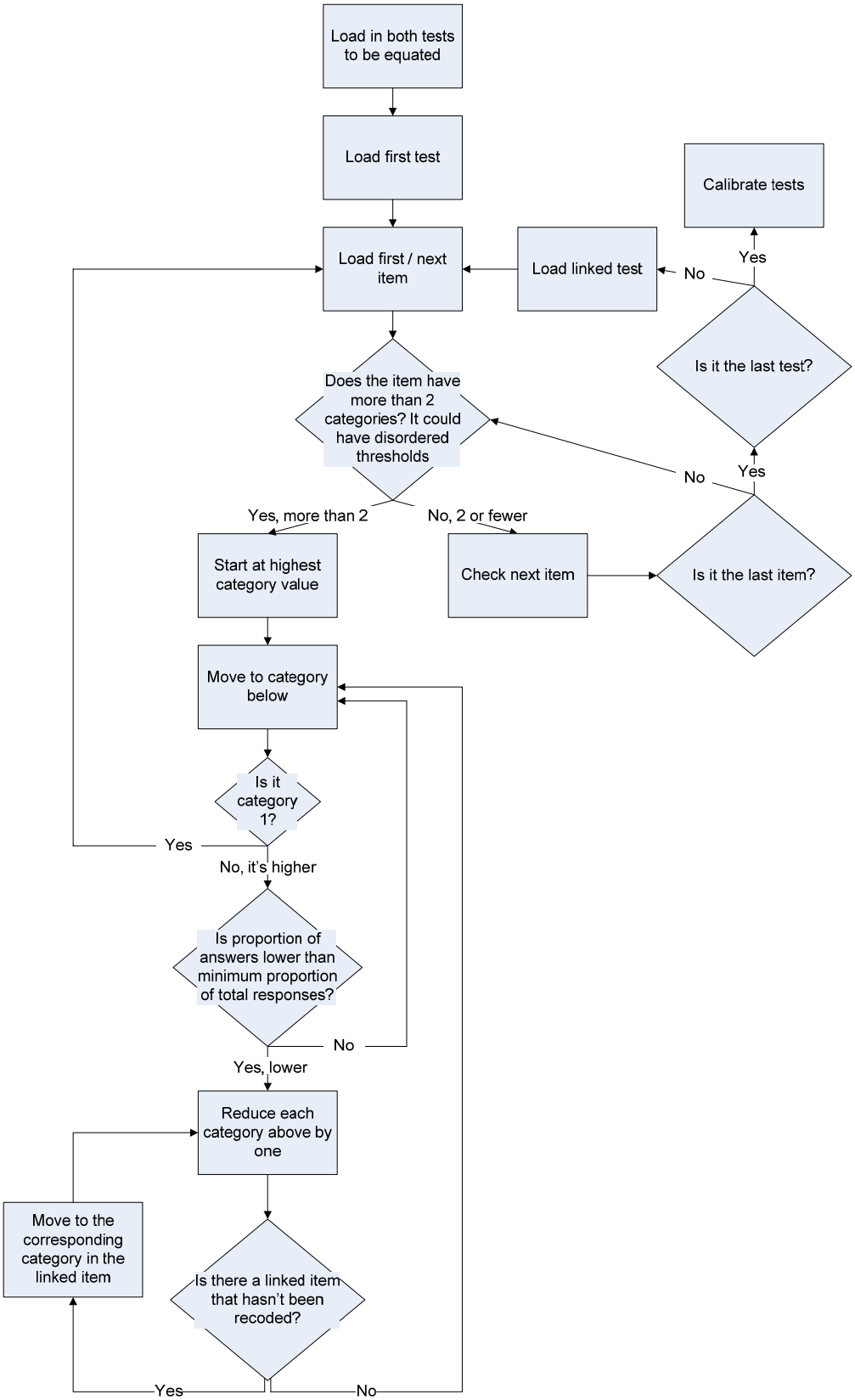


Figure 5.3: Collapsing categories

5.8.1.4 MMRM samples

Random samples of 1000 candidates were taken for the foundation tier. As the MMRM is a dichotomous model the answers then had to be dichotomised. This was done using an algorithm which split the responses according to the median value. The full polytomous response was retained, however, for the subsequent equating. The MMRM model was then implemented under MCMC in WinBUGS (Spiegelhalter et al., 2003). 5000 iterations were used of which 4000 were burnt in. Only the first 12 items and the last 6 items were modelled. Once the latent classes had been estimated, candidates whose ability profile declined towards the end of the test were removed from the sample. This purified sample was then used for the equating. Multiple samples were used to check the stability of the classifications.

5.8.1.5 Equating with matched samples and collapsed categories

Using samples whose mark distribution is not randomly equivalent to the population that they had been taken from and collapsed categories complicates the equating procedure as neither the marks nor the cumulative percentages in the sampled values relate to the corresponding values in the population. The following process was therefore developed to derive any equivalent score from one form to another:

1. Calculate the cumulative percentages for the full cohort on both test forms using the uncollapsed items scores.
2. Calculate the cumulative percentages for the full cohort on both test forms using the collapsed test scores.
3. Find the equivalent collapsed scores between the unrepresentative samples of candidates using the IRT estimated summed scores for those samples.

4. Match these collapsed scores from the sample against the collapsed scores on the full cohort.
5. Use the cumulative percentages on the full cohort to find the equivalent mark on the uncollapsed scores.

The number of matches required obviously creates a level of inaccuracy in the derivations; this is exacerbated by tests that do not discriminate well among the candidates.

5.8.2 Components

The GCSE Mathematics tests analysed in Chapters 3 and 4 were used for the analysis. They showed poor fit to the Rasch model and evidence of candidates working under pressure towards the end of their tests. The equating design has the majority of the anchor items at the end of one tier (Figure 5.4) but at the beginning of the other. The impact of the time pressure on the equating is therefore likely to be an issue.

Mathematics Paper 1

Foundation						
1a	1b	1c	1d	2a	2b	3a
3b	4a	4b	4c	5a	5b	5c
6a	6b	6c	7a	7b	8a	8b
8c	9ai	9aii	9b	10a	10b	10c
11	12a	12b	13a	13bi	13bii	13c
14	15a	15b	16a	16b	16c	17a
17b	17ci	17cii	18a	18b	18c	19
20	21	22a	22b	22c	23	24a
24b	25a	25b	26	27a	27b	27c

Higher						
1a	1b	1c	2	3a	3b	4a
4b	5	6	7a	7b	8a	8b
8c	9	10a	10b	10c	11a	11b
11c	11d	12a	12b	13a	13b	14a
14b	15a	15b	16	17	18a	18b
19a	19b	20a	20b	21	22a	22b
23	24a	24b	25a	25b		

Mathematics Paper 2

Foundation						
1	2a	2b	3a	3b	4a	4b
5a	5b	5c	6a	6b	6c	7a
7b	8a	8b	9a	9b	10a	10b
10c	10d	11	12	13a	13bi	13bii
13biii	14ai	14aii	14b	15	16a	16b
17a	17b	17c	18a	18b	19	20a
20b	20c	21	22a	22b	23a	23bi
23bii	24	25	26	27ai	27aii	27b

Higher						
1a	1b	1c	2	3a	3b	4a
4b	5a	5b	6a	6bi	6bii	7a
7b	8a	8b	8c	9a	9b	10
11	12	13a	13bi	13bii	13ci	13cii
14a	14b	14c	14d	15a	15b	15c
16	17	18	19	20	21	22a
22b	23	24a	24b	25a	25b	26a
26b						

Figure 5.4: Equating design for Mathematics. The linking items are shaded

5.9 Results

5.9.1 MMRM Samples

Before the equating was done the MMRM samples were taken from the foundation tier. The latent classes are illustrated in Table 5.1. The non-speeded group was large enough in each case to be taken as the purified sample for the equating. As expected this was a more able group than the entire cohort (Table 5.2). The mean and standard deviation of the sample against which it would be compared was considered close enough to the entire cohort for it to be considered representative.

Table 5.1: Latent classes by time pressure

	Speeded Group (1 - Unspeeded, 7 - Speeded)						
	1	2	3	4	5	6	7
Paper 1	558	17	36	84	180	125	0
Paper 2	457	22	30	24	13	434	20

Table 5.2: Descriptive statistics for the samples

	Paper 1			Paper 2		
	N	Mean	SD	N	Mean	SD
Entire cohort	30701	53.09	19.69	30544	52.35	19.56
Representative sample	1000	53.21	19.36	1000	52.68	19.55
Purified sample	558	64.46	13.97	457	65.80	13.31

5.9.2 Collapsing categories

Table 5.3 illustrates that the majority of items contained categories with few observations. This could lead to poor and unstable parameter estimation, and, for the smaller, purified sample, failure of the estimation. A minimum value of 50 observations was set for each category, resulting in the collapse of categories illustrated in Table 5.3. This reduced the number of disordered thresholds in the equating from 63 to 29 in Paper 1 and from 70 to 33 in Paper 2 (Table 5.4). The disordered thresholds in the categories used in the common item links were similarly reduced from 7 to 2 and from 6 to 1. It is possible that some information was lost, however, as the linking categories were reduced from 22 to 16 and 23 to 17, which suggests in each case that one category that was collapsed was not disordered.

Table 5.3: Category frequency distributions before and after collapse for foundation tier Paper 1

Before Collapse			After collapse		
0	1	2	0	1	2
0.02	0.02	0.96	0.02	0.98	
0.15	0.21	0.64	0.15	0.21	0.64
0.05	0.06	0.89	0.05	0.06	0.89
0.49	0.01	0.50	0.49	0.51	
0.05	0.04	0.91	0.05	0.95	
0.14	0.04	0.82	0.14	0.86	
0.16	0.22	0.62	0.16	0.22	0.62
0.37	0.05	0.58	0.37	0.63	
0.45	0.01	0.54	0.45	0.55	
0.31	0.07	0.62	0.31	0.07	0.62
0.37	0.05	0.58	0.37	0.63	
0.60	0.10	0.30	0.60	0.10	0.30
0.52	0.05	0.44	0.52	0.48	
0.17	0.24	0.59	0.17	0.24	0.59
0.73	0.03	0.24	0.73	0.27	
0.31	0.02	0.67	0.31	0.69	
0.17	0.13	0.70	0.17	0.13	0.70
0.73	0.22	0.05	0.73	0.22	0.05
0.58	0.35	0.07	0.58	0.35	0.07
0.74	0.01	0.25	0.74	0.26	
0.78	0.04	0.18	0.78	0.22	
0.57	0.35	0.08	0.57	0.35	0.08
0.51	0.46	0.03	0.51	0.46	0.03

Before Collapse			
0	1	2	3
0.87	0.03	0.01	0.09
0.72	0.05	0.06	0.18
0.75	0.05	0.04	0.16
0.38	0.25	0.01	0.36
0.79	0.17	0.02	0.02

After collapse		
0	1	2
0.87	0.13	
0.72	0.1	0.18
0.75	0.05	0.2
0.38	0.25	0.37
0.79	0.17	0.05

Before collapse				
0	1	2	3	4
0.27	0.16	0.03	0.08	0.46

After collapse			
0	1	2	3
0.27	0.16	0.11	0.46

Table 5.4: Categories before and after collapsing

	Paper 1		Paper 2	
	Before	After	Before	After
Categories	200	159	200	156
Disordered	63	29	70	33
Linking categories	22	16	23	17
Link categories disordered	7	2	6	1

5.9.3 Descriptive measures

The reliability coefficients for the tests (Table 5.5) appeared sufficient and equal, which is a requisite for equating. The G6 value refers to Guttman's Lambda 6, which considers the amount of variance in each item that can be accounted for by the linear regression of all of the other items. The average r is the average inter-item correlation. The values for the purified sample were similar.

Table 5.5: Reliability of the representative sample after collapsing

		alpha	G6	average r
Paper 1	Foundation	0.94	0.95	0.21
	Higher	0.94	0.95	0.26
Paper 2	Foundation	0.93	0.95	0.20
	Higher	0.92	0.94	0.20

5.9.4 Initial inspection for DIF

One item appeared to be common, but was relatively easier in the context of the foundation tier paper. On close inspection of the mark schemes it became apparent that the foundation tier candidates were not being penalised a mark for inaccuracy in the same way as the higher tier candidates. This question was removed as a common question.

5.9.5 Rasch equating quality measures

Using the Rasch equating method the between-link and within-fit measures appeared adequate, with the item-within-link fit slightly improved by use of the purified sample. The gradient, however, was poor in both cases. This means that a different result will be obtained depending on whether the foundation tier or the higher tier is used as the basis of the equating, which violates one of the conditions of equating.

Table 5.6: Equating quality measures under the Rasch method.

	Paper 1		Paper 2	
	Representative Sample	Purified Sample	Representative Sample	Purified Sample
N	1000	558	1000	457
Item between				
link fit	19.96	15.02	15.74	9.27
df	15	15	16	16
w	0.14	0.16	0.13	0.14
Item within link				
fit	0.91	0.96	0.93	1.02
R	0.81	0.84	0.90	0.93
Slope	0.67	0.70	0.80	0.79
Constant	-2.28	-2.20	-2.35	-2.23

5.9.6 OPLM equating quality measures

The OPLM equating quality measures (Table 5.7) showed statistically significant and substantial levels of misfit according to both the R0 and R1M statistics, with the exception of the purified sample for Paper 2. The R0 results suggest that the MML estimation is poor, which could be due to incorrect assumptions regarding the ability distribution or misfit in the items; and the R1m results suggest that the CML estimation of the item parameters is also degraded due to incorrect specification of the discrimination parameters or differential item functioning across subgroups.

It might be expected that a purified sample would show better fit to OPLM as the assumptions of the model are better observed. However, Table 5.7 shows that for Paper 1 the fit was worse. The R1m and the R10 statistics both increased in absolute value and in terms of their effect size. For Paper 2 the fit did seem to improve but substantial levels of misfit remain.

Table 5.7: OPLM equating quality measures

	Paper 1		Paper 2	
	Representative Sample	Purified Sample	Representative Sample	Purified Sample
N	1000	558	1000	457
R0	588.25	713.48	1665.56	429.74
Df	440	444	429	419
w	0.77	1.13	1.29	0.97*
R1				
M	2003.48	1974.07	4979.49	1339.41
Df	1295	1305	1253	1091
w	1.42	1.88	2.23	1.71

*$p > 0.05$

5.9.7 Inspection of Item Response Functions

Despite the results from the R0 and R1m statistics visual inspections of the Item Response Functions under OPLM generally revealed them to be acceptable. The Item Response Functions for the link items for the representative sample were of particular interest; it may be hypothesised that if some of the linking items were speeded on the foundation tier and some were not, this would be indicated by differential item functioning. Figure 5.5 illustrates the Item Response Functions under the Rasch model and OPLM for the last item in the foundation tier. In the diagram on the top left, for the foundation tier under the Rasch model, the observed item responses for the foundation tier fall below the curve. In the diagram on the bottom left, for the higher tier under the Rasch model, the observed responses rose above the curve. The slope is too shallow for both tiers. By increasing the discrimination under the OPLM model the item can be successfully modelled. The diagrams on the right illustrate the better fit achieved through this adjustment.

Increased discrimination, or poor fit under the Rasch model, can be interpreted as symptomatic of an item taken under time pressure. The item discrimination parameters for the representative population generally seemed raised towards the end of the foundation tier test for the representative sample under OPLM (Table 5.8).

5.9.8 Equating results

The outcomes from the equating are detailed in Table 5.9. According to the award, 23 per cent of foundation tier candidates were worthy of a grade C on Paper 1 and 22 per cent on Paper 2. On the equivalent higher tier papers 84 per cent and 86 per cent were judged worthy of this grade. The Rasch method on the representative sample suggests more generous grade boundaries, 12 marks lower for Paper 1 and 6 marks lower for Paper 2. The OPLM method on the representative sample agrees on more generous boundaries than the award but not to the same degree. It suggests 8 marks lower on Paper 1 and 5 marks lower on Paper 2. The purified sample still suggests more generous boundaries than the award, but they are closer. The Rasch method suggests 8 marks lower on Paper 1 and 3 marks on Paper 2. The OPLM method suggests 6 marks lower on Paper 1 and 5 marks lower on Paper 2.

On the basis of this analysis it appears that the estimations from the Rasch model are affected to a greater degree by the placement of the linking items. If there is a speeded dimension to the test, as it appears there is, this is a clear violation of the unidimensionality required by both measurement models, Rasch and OPLM. On the evidence of the Item Response Functions, however, such as the one illustrated in Figure 5.5, it would seem that OPLM can adapt to this violation more readily by allowing item discrimination parameters to vary.

All the analysis from the preceding two chapters has suggested that modelling the discrimination of items discretely will lead to better fit and better short term predictions. The results from the OPLM method are likely therefore to be more accurate. Similarly, the analysis from the preceding chapter suggests that the purified sample will lead to better predictions. Overall, therefore, it would seem sensible to evaluate the award against the results from the OPLM method on the purified sample. On this basis, the standards for Paper 2 would seem to be better aligned than for Paper 1.

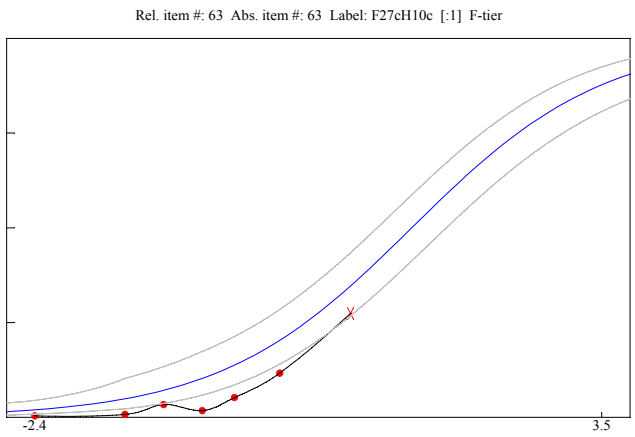
The practical differences between using a model with better fit are now apparent. The difference between the OPLM method on the purified sample and the Rasch method on the representative sample is 6 marks or nearly 6 per cent. The statistical tolerance for predictions carrying forward the standard for a cohort of this size is likely to be 2 per cent. Choice of model and care with equating design and estimation method are clearly important.

Of course, it is impossible to validate the OPLM method in itself. Predictions taken from a common coursework test proved to be wildly inaccurate. This is unsurprising given the different assessment mode of this test. The equating results do seem plausible and are fairly consistent given that the same candidates take both papers. In the absence of any other objective basis for comparison of the performance standards between the tiers there may be no alternative but to accept this evidence. That does not mean, however, that the quality of the design could not be improved to overcome some of the difficulties experienced with this equating. That there are objective measures of the quality of the equating is useful; such measures can be used, with experience, to evaluate the weight the equating evidence should be given.

Two obvious improvements could be made to the design. The first would be to review the mark schemes to reduce the incidence of missing categories. The second would be to place the linking items earlier in the foundation tier tests. This latter suggestion may meet some objection, however, as there is a notion that question papers should be designed with an incline in difficulty. The linking items are clearly considered to be the most difficult and are therefore located at the end of the tests. Figure 5.6 plots the category thresholds for the purified sample (those not under time pressure) for Paper 2 of the foundation tier. It would be very hard to interpret the line connecting the thresholds as an incline. The compound between difficulty and lack of time may have helped perpetuate the myth of the incline of difficulty.

A third possible improvement is to ensure that all candidates can complete the foundation tier papers. This analysis cannot however reveal whether candidates are running out of time or whether they are running out of energy or motivation. This should be looked into before more time is allowed.

Item response functions under the Rasch model



Item response functions under OPLM

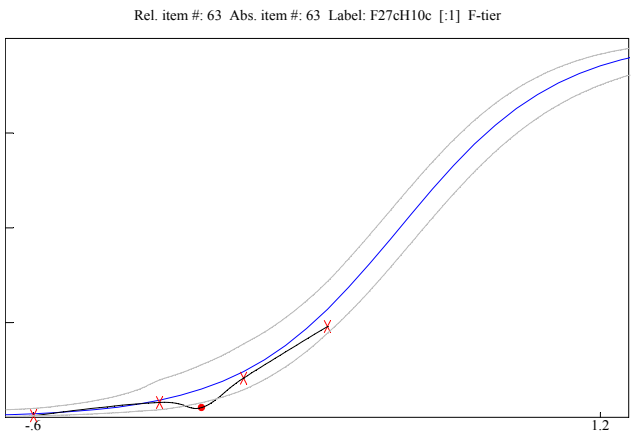


Figure 5.5: Item response functions under the Rasch model and under OPLM

Table 5.8: Item discrimination (A) and difficulty (B) parameters for the representative sample of foundation tier Paper 1

Item	Label	A	B
1	1	1	-1.294
[2]			-1.505
[3]			-1.501
2	2a	2	-1.31
[2]			-0.649
3	2b	3	-0.774
4	3a	2	-0.3
5	3b	2	-0.354
[2]			-0.295
6	4a	2	-1.631
7	4b	2	-1.251
8	5a	2	-1.691
9	5b	2	-0.762
[2]			-1.16
10	5c	3	0.054
11	6a	2	0.13
12	6b	1	1.232
[2]			-1.096
13	6c	1	1.368
[2]			-1.307
14	7a	2	-0.908
15	7b	2	0.274
16	8a	2	-1.92
17	8b	2	-0.541
18	9a	3	-1.206
19	9b	4	-0.529
20	10a	3	-1.972
21	10b	4	-1.708
22	10c	2	-1.631
23	10d	3	-1.844
24	11	2	-0.8
[2]			-0.847
25	12	2	-0.131
[2]			-0.292
[3]			-1.146
26	13a	4	-0.553
27	13bi	4	0.059
28	13bii	6	-0.326
29	13biii	2	0.706
30	14ai	6	-0.449
31	14aii	3	0.542
[2]			-0.597
32	14b	2	-0.654
[2]			-0.638
33	15	5	-0.15
34	16a	3	-1.055
35	16b	2	-0.23
[2]			-0.699
36	17a	2	0.375
[2]			-0.573
[3]			-1.036
37	17b	4	-0.166
38	17c	2	0.168
[2]			1.018
39	18a	3	-0.867
40	18b	3	-0.434
41	19	3	-0.049
[2]			-0.26
42	20aH1a	2	-1.828
43	20bH1b	3	-1.063
44	20cH1c	2	-0.232
45	21H2	2	0.386
[2]			-0.713
46	22aH5a	2	-0.115
[2]			0.095
[3]			0.542
47	22bH5b	5	0.322
48	23aH6a	1	-0.874
49	23biH6bi	3	-0.171
50	23biiH6b	2	0.509
51	24H9b	3	0.232
[2]			0.248
52	25H12	6	0.039
53	26	5	-0.113
54	27aiH13b	3	-0.38
55	27aiiH13	3	0.808
56	27bH13ci	2	-0.014
[2]			-0.459

Table 5.9: Equating results

	Foundation		Higher Tier									
			Representative Sample						Purified Sample			
	Award		Award		Rasch		OPLM		Rasch		OPLM	
	Mark	Cum %	Mark	Cum %	Mark	Cum %	Mark	Cum %	Mark	Cum %	Mark	Cum %
Maths Paper 1	70	23.01	28	84.36	16	95.60	20	91.71	20	92.20	22	89.82
Maths Paper 2	69	22.21	30	85.79	24	93.20	25	92.19	27	89.80	25	91.49

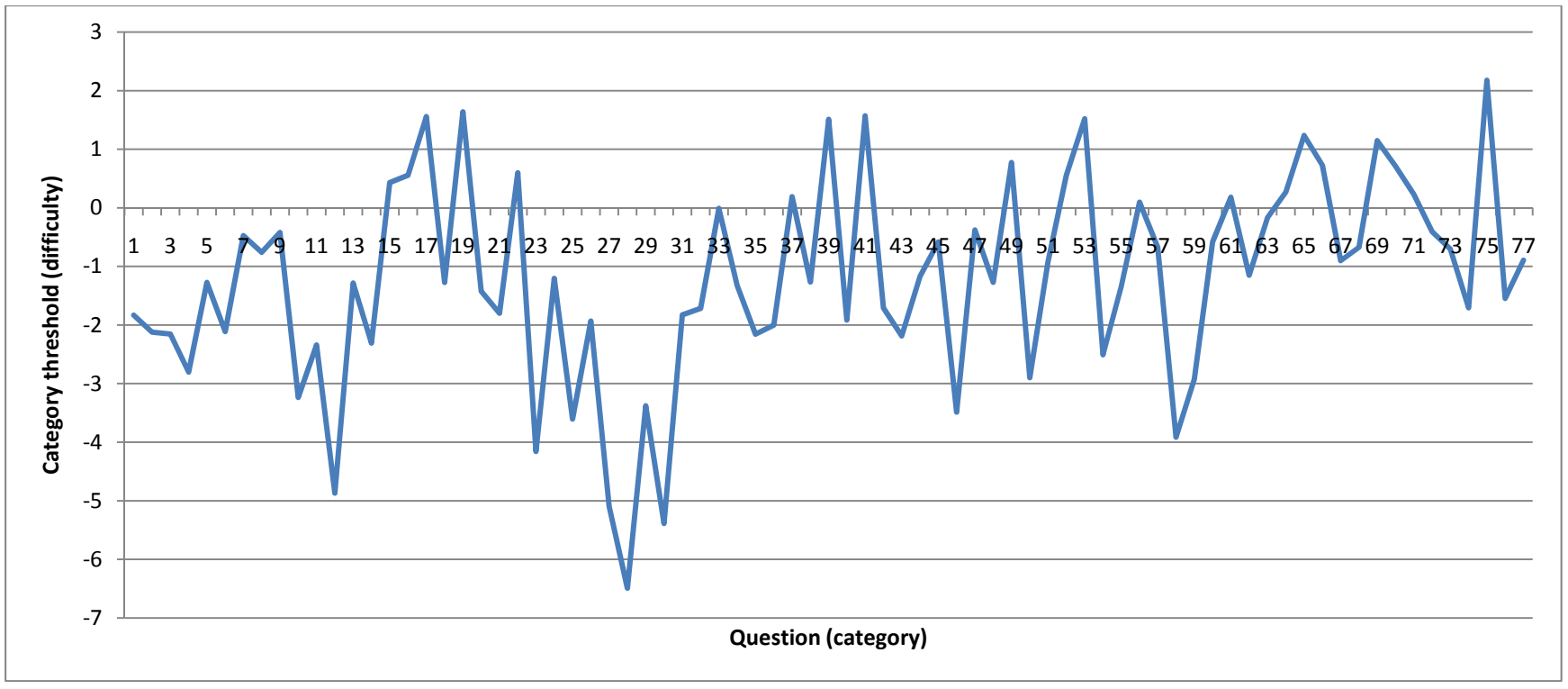


Figure 5.6: Category thresholds for Mathematics foundation Paper 2 for the purified sample in item order

5.10 Further context effects

Following the detailed study of equating across tiers in Mathematics a number of other tiered GCSE tests were analysed for context effects. Under separate calibration in the Rasch model the difficulty parameters for the common items can be quickly and easily compared to see if they are performing differentially between one test form and another.

5.10.1 Differences in mark allocations or mark schemes

A number of items that appeared on first sight to be 'common' in fact had different maximum marks or different mark schemes. The problem with a different maximum mark for an item is that, unless the mark scheme is very specific, the category thresholds or steps are likely to be different distances apart. Figure 5.7 illustrates the category thresholds for an identical item which is allocated 6 marks in the foundation tier and 9 marks in the higher tier. The first category threshold is lower for the foundation tier than the higher, while the 6 mark threshold is higher for the foundation tier than the higher. The same range of latent ability is covered by 6 marks in the foundation tier as is covered by 9 marks in the higher tier. It is therefore not sensible to equate on these thresholds.

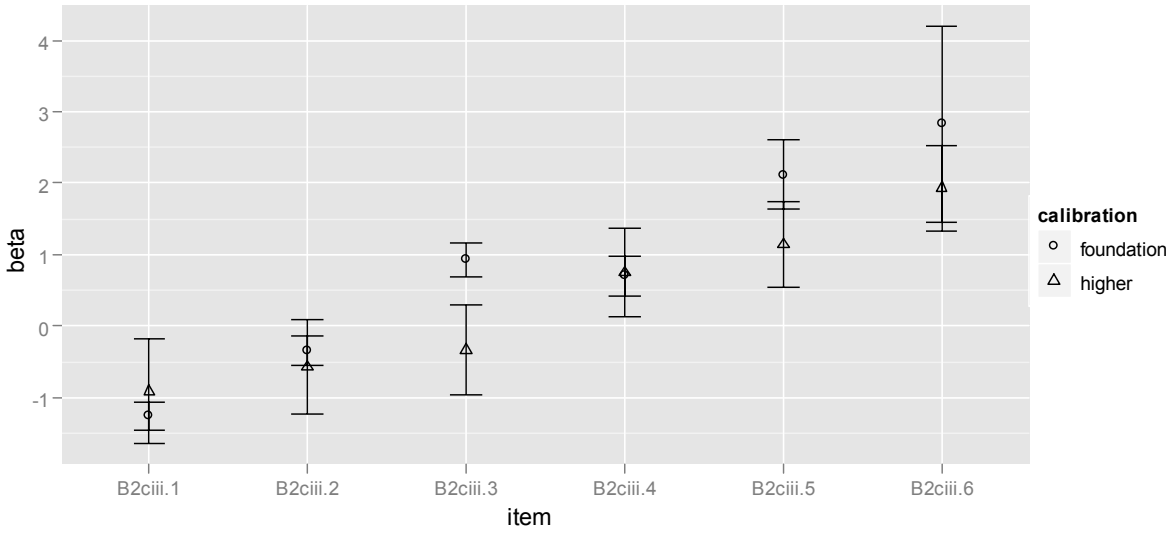


Figure 5.7: Relative difficulties of category thresholds for an item with a different maximum mark on different forms

5.10.2 Cognitive clues

The Geography GCSE papers attempt to minimise context effects by ensuring that common items are placed in similar locations in each paper. They do not completely eliminate them, however, as common items may be preceded by different items in the same content block. In one paper, for example, there is a section on urban redevelopment. The foundation tier candidates are given two very structured items to begin the section with. They are asked to study a sketch of a town undergoing redevelopment and note two changes that have happened. They are then asked, for each change, to explain why it has happened. Finally they are asked to explain why changes in such areas may bring disadvantages to some of the residents. In the higher tier this same final item is preceded by an open-ended item that asks how the changes shown in the sketch may improve the environment and the lives of the local people. The category thresholds depicted in Figure 5.8 show that the step to the second category is much smaller for the higher tier candidates. It could be speculated that, for the higher tier candidates, having thought of the advantages to some of the

residents of the changes it is easier then to consider the disadvantages. This logical step is missing for the foundation tier candidates. A similar structural difference caused the opposite effect in another paper: in this case the structure preceding the open-ended item made the subsequent open-ended item slightly easier for the foundation tier candidates.

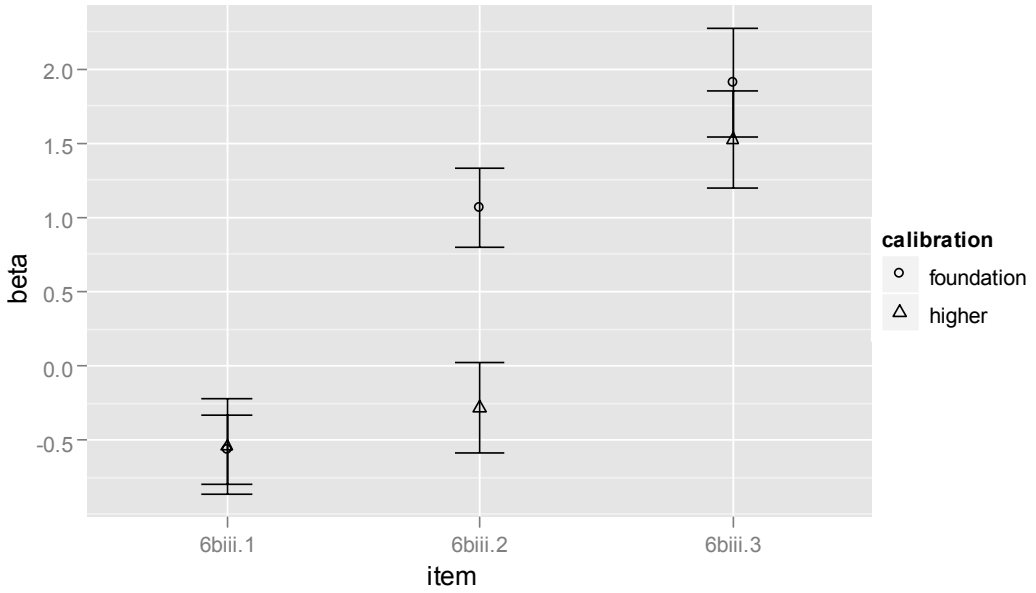


Figure 5.8: Impact of context on relative difficulties of category thresholds for an item

One way around this problem would be to design an entire content section as common between tiers. This would ensure that the thought processes stimulated by the items is as similar as possible. The disadvantage of this approach is that fewer content areas and fewer skill areas are sampled by the common items so they are no longer representative of the subject as a whole.

5.10.3 Interactions between ability and item correlations

The gradient of the line of best fit for a Chemistry paper was relatively steep at 1.33, although the correlation between the item difficulties was relatively high at 0.91

(Figure 5.9). Item 8A is clearly an outlier. Examination of the item fit showed a higher outfit for the item (1.45) on the higher tier than on the foundation tier (1.18). The correlations between the responses and the options on the higher tier revealed that an incorrect option had distracted many of the high-scoring candidates. A slight ambiguity in the item which could point to that option then became apparent; obviously this ambiguity was also apparent to some of the highest scoring candidates. The fit statistics also revealed that item 9D over-fitted the higher tier test (0.72). Over-fit is not normally considered an issue, as it is usually taken to represent the presence of items targeting a similar area, but over-fit can lead to the difficulty parameter being poorly estimated. Removal of these items leads to an improved gradient of 1.16 and a correlation of 0.99.

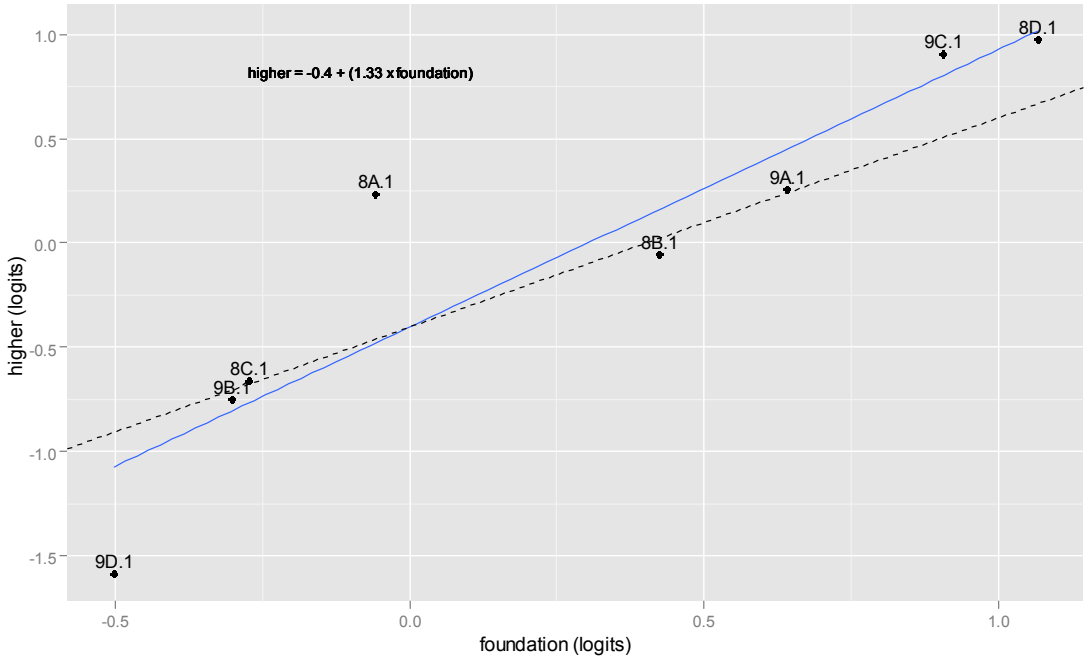


Figure 5.9: The line of best fit through category thresholds for a GCSE Chemistry paper. The solid line is the line of best fit, while the dotted line illustrates a gradient of 1.

5.11 Discussion

The purpose of this chapter was to examine the consequences or poor model fit in a practical equating scenario within the public examination system in England. Specifically, it was to consider the improvements that could be gained from the use of an item discrimination parameter in the IRT models and consider the impact on parameter estimation of a subset of candidates that could undermine the invariance of the estimated parameters. The impact of modelling item discrimination discretely was studied by comparing the quality and outcomes from the Rasch method of equating to the OPLM method on a GCSE test. The impact of the subset of candidates on parameter estimation was studied through a mixture Rasch model that isolated that subset and allowed comparisons between a purified and a representative sample. The subset of candidates that was isolated were those whose performance appeared to decline towards the end of a test. The mixture Rasch model used was the MMRM. The study was extended to consider context effects that may affect equating across a number of GCSE tests.

There were two major practical difficulties in undertaking this study. The first was due to the number of categories with relatively few observations within the polytomous items studied. This is due to the nature of the mark schemes that are used for GCSE Mathematics. These mark schemes allow for credit that can be given to very small subsets of candidates who make clerical errors. Few observations within a category can lead to unstable estimation of item parameters and disordered category thresholds. This difficulty was overcome through an algorithm that considered the proportion of observations on each category on each test form and collapsed them in tandem. The algorithm appeared relatively successful in reducing

the incidence of disordered thresholds without overly reducing the information available to the test equating.

The second practical difficulty came from equating using non-representative samples. Again algorithms had to be prepared which matched the performance from the non-representative sample against the performance of the representative sample. Both of these practical difficulties illustrate the technical challenge in implementing equating; a challenge, however, which is not insurmountable.

The results from the study suggest that the improved fit of OPLM and the additional information available under the concurrent estimation of item parameters under the OPLM method of equating can prove useful in sub-optimal equating conditions. The speeded element in these tests introduced multi-dimensionality that violates the assumptions of both OPLM and the Rasch model. OPLM, however, appeared to produce a more robust model of the data when speeded and unspeeded samples were compared. The Rasch model is constrained by the need to model all items with the same discrimination. OPLM allows the speeded items to be modelled as more discriminating than other items in the test.

Rather than attempt to cross-validate these results, which would probably prove a thankless task, it is suggested that energy should be devoted to improving the conditions for test equating under an IRT model. The following recommendations are suggested by the study:

i. mark schemes and item presentation should be identical across test forms
ii. linking items should be located early in the test forms
iii. linking items should consist, where possible, of coherent blocks of items, but should remain representative of the subject as a whole

iv. redundant linking items should be included in case they perform differently across test forms (the rule of thumb is for 20 per cent of items to be common between forms to be equated)

As the equating design improves then validation of results is more easily sought in the many simulation studies that prove the efficacy of IRT methods of test equating.

This study has shown that IRT methods of test equating can provide an estimate of relative performance standards between different test forms. It is, granted, hard to validate this estimation, but it does provide objective evidence supplemented by quality criteria that can be interrogated. Once experience has been gained in these quality criteria then the evidence it provides can be appropriately weighted. The importance of this finding should not be underestimated. The comparability of the key grade C in the GCSE is not routinely evaluated either by judgement or statistics. Judgemental approaches have been shown to be flawed where the difficulty of test forms differs. Statistical approaches based on prior achievement are effective in carrying forward the respective standards of tiers but cannot make an estimation of their relative standard as value-added may differ by tier.

Further, the analyses of this and the preceding two chapters have shown, contrary to the fears of some commentators, that Rasch and IRT models rather than constraining the validity of tests prove a useful framework in which the validity can be measured. The performance of a subgroup of candidates highlighted by the mixture Rasch model suggests the presence of construct irrelevant difficulty introduced by time pressure. Information from odds ratios can support or challenge suggestions of cross information that are a source of construct irrelevant easiness.

IRT test equating can only be done if the quality of the tests supports the fitting of IRT models. This can be seen as a challenge to test quality rather than a threat.

Operating within the framework of reliability and validity IRT methods of test equating have a key advantage over other statistical methods for maintaining standards. It is very hard to conduct a test equating study without consideration of the fabric of those tests. This contrasts with the use of predictions based on prior achievement models. These predictions can be prepared and met as long as the minimal requirement that tests discriminate to a certain extent is satisfied. Those preparing the predictions and those consuming the predictions need never concern themselves with the content or quality of the tests, and can be quite ignorant of the performance standard that those predictions represent. Rather than actively working to improve the quality of tests predictions can be used to compensate for flaws in those tests. When IRT is applied carefully and rigorously flaws in those tests it seeks to model are apparent.

This study, however, benefited from the advantage that the test equating was being done in the absence of any other objective measure of a relative performance standard. It is less obvious, however, how IRT methods of test equating can contribute to the maintenance of standards over time within the constraints of the English public examination system. The next two chapters will consider how IRT methods of test equating can make a practical contribution to the maintenance of standards over time.

6. Horizontal Test Equating

6.1 Overview

The Code of Practice positions the maintenance of standards over time at the heart of the purpose of awarding in the English public examination system. This purpose could be disputed, for it could be argued that all that is needed is the correct rank ordering of candidates year on year. This will allow universities and companies recruiting recent graduates to make relatively accurate judgements on the calibre of their applicants. In the longer term it could be argued that the maintenance of standards over time when syllabuses and educational priorities change is a philosophical question that should not trouble those in the business of awarding grades. In the shorter term, however, the advent of the modular system means that unit standards within the two-year period of a GCSE course need to be maintained for the system to be fair. Candidates entering early should not be unduly advantaged or disadvantaged by the relative easiness or severity of unit grading. Shifts in the entry policies of schools should be determined by educational reasons not by tactical decisions based on perceptions of the relative easiness or difficulty of different examination series within a year.

The purpose of this chapter is to consider the role of IRT test equating in maintaining standards within the modular system of public examinations in England. It seeks to test the assumptions of IRT models in a practical experiment designed to compare the relative standards of two modular sessions. Most of the issues, however, could be generalised to the year to year priorities of maintaining subject standards in the more traditional linear system.

6.2 What are modular examinations?

Up until the end of the 1990s most public qualifications in England were linear; all assessments for a qualification had to be entered in the same examination series. A candidate taking A-level Chemistry, for example, took all their assessments for that qualification in June of their final year of school. In 2001 there was a national move towards modular structures for A-levels in which assessments could be entered at different points during a candidate's course of study. From 2008, a candidate following a two-year course leading to an A-level could enter one unit in January of Year 1, another in June of Year 1, another in January of Year 2 and the final unit in June of Year 2 (subject to the availability of the units at these times). They can also re-take units if they wish.

A similar move towards what has been termed modularisation has occurred with GCSEs, but in a piecemeal fashion. For example, a candidate following a two-year course leading to a GCSE qualification may enter one unit in June of Year 1, another in January of Year 2 and another in June of Year 2 (subject to the availability of the units at these times). Again, they can re-take units. Until recently there was no limit on the number of units that a GCSE could consist of. From 2009 the vast majority of all new GCSEs will be modular.

These changes in structure pose the difficult question of whether candidates at earlier stages in their programme of study should be expected to perform as well as candidates at later stages in their study. The former may benefit from recency of instruction, while the latter may benefit from maturity and a synoptic overview of the subject.

One investigation compared the relative progress of 15-year-olds with the relative progress of 16-year-olds from a baseline measure of prior achievement (Pinot de Moira, 2009a). The author concluded,

- Within the limitations of the data available, 15-year-old candidates do appear to perform differently from their 16 year counterparts.
- The difference is most extreme at the mid-points of the grade distribution.
- The difference in performance between 15 and 16-year-old candidates varies between subjects. (Pinot de Moira, 2009a, p. 9)

On the basis of this evidence she advised that generic advice regarding the effect of maturity is unlikely to be possible and that it is difficult to generalise from these results as candidates entering for different subject and different sessions are self-selecting rather than random samples.

6.3 Current approaches to maintaining unit standards in a modular system

The following account refers largely to the procedures used by AQA as little is known of the detail of how other awarding bodies maintain unit standards over time.

6.3.1 A-levels

While most A-level units are taken in June, each A-level specification will generally make one or two units available for January entry. This allows candidates to enter early or to re-take the unit from the following June. For the second January award of

the new four units structure A-levels in 2010 statistical predictions were prepared for each unit based on the unit outcomes the previous June (Pinot de Moira, 2009b). While the predictions were based on 17-year-olds only, no compensation was made for the fact that the 17-year-olds were entering the unit five months earlier than the 17-year-olds on which the prediction was based. It would have been misleading if not impossible to produce any generic statistical adjustment based on expectations of performance standards as the impact of maturation on performance is likely to vary from subject to subject. Jones (2008), for example, considers that the impact of maturation is likely to vary dependent on the degree to which subject content is discrete or cumulative in nature. The concern instead was to attempt to predict the unit outcomes that would be required to maintain subject standards once all the unit outcomes from January and June (the series yet to take place) had been aggregated (Jones, 2009b).

Once the entry to a January unit has stabilised, predictions based on January performance can then be used to maintain standards for that unit from January to January. Apart from the obvious assumption built in to the original January standard, this creates a second issue. As the January unit uses a separate set of predictions to the June unit the performance standards can become decoupled (Eason, 2007, 2008; Jones, 2005, 2008). When this discrepancy becomes extreme the predictions for January can be re-calibrated against the predictions that would have been derived from the preceding June. While it is the subject standard, the standard represented by the aggregation of all unit outcomes, that has currency and is therefore the priority, some measurement of the relative performance standard could inform the way in which the relative outcomes of units are balanced in order to achieve the maintenance of this standard.

6.3.2 GCSEs

The situation for GCSEs can be more complex than for A-levels, as units can be available up to 3 times a year in November, January, March, and June. The impact of the unit outcomes in the earlier series (November, January and March) on the subject outcomes is much harder to gauge (Eason, 2009; Whitehouse & Eason, 2007). As a result the unit standards for the series are usually based on the June unit outcome. In such cases a measure of the performance standard would be even more useful as it is less likely to conflict with the need to maintain the aggregated subject outcome.

6.4 An IRT test equating approach

An IRT test equating approach would seem an appropriate framework within which to approach the modular problem as the modules to be equated are designed to the same specification. The major theoretical issue, however, is that all test equatings are population dependent,

> If the assumptions of IRT hold, then IRT true-score equating is invariant over all subpopulations, which seems to make the task of examining invariance irrelevant. However, in general, the population invariance of IRT true-score equating does not hold when equating functions are used with observed scores. (Brennan, 2008, p. 109)

Some recent studies are encouraging. Five studies reported little sensitivity of equating results for subgroups formed on the basis of characteristics such as gender, race/ethnicity, and the geographic location (Petersen, 2008). Population invariance is likely to depend on the construct similarity of the tests being equated and whether

the selection variable for constructing the subgroups is related to the construct being equated (Petersen). Modular public examinations in England would seem to satisfy this first requirement. The second, however, is more problematic. Yi, Harcourt Assessment, Harris, and Gao (2008), for example, examined the population dependence of the equating of three different equivalent forms of a Science achievement test. They found that candidates studying physics had a different equating function to candidates not studying physics. The equated scores for the physics subgroup did not match the equated scores for the total group.

A particularly relevant study is that of Cook and Paterson (1987) which found that when relatively parallel forms of a Biology achievement test were equated using groups of students who took the tests at different times of the year (May and December) they got very disparate equating results. After close examination of the administration groups they found that students who took the test in May were primarily sophomores completing a course in biology, whereas students who took the test in December were primarily seniors who had not taken biology since their sophomore year. They concluded that the disparate equating results were due to an interaction between recency of instruction and test content. Their advice is that careful thought should be given to the selection of the group to be used for equating.

The major practical issue with the test equating designs needed to maintain unit standards over time is security. If anchor items are presented to a cohort in November and then repeated in March, the March cohort may contain re-take candidates from November. These candidates may then gain an unfair advantage. With large item-banks these problems can be ameliorated in a variety of different ways , but item-banks are expensive and time-consuming to develop.

A pragmatic solution is the use of a Post-Equating Non-Equivalent Groups design (PENG) which is used in the Netherlands (Alberts, 2001). This design requires learners who are not participating in the live examinations to take some of the old and some of the new items after the administration of the live tests. 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). The difficulty lies in finding a cohort who is adequately prepared and motivated, and persuading enough schools to take part. This is solved in the Netherlands to some extent as candidates from the vocational stream, who would not normally take these examinations, are given credit for their participation in the trials. The low number of schools involved, however, can lead to substantial school level effects.

Clearly, the literature can offer only generic guidelines as to how likely an equating design is to prove robust. Potential population dependence and the severity of that dependence is very hard to predict in advance of the data collection. This study will therefore seek to examine how certain assumptions of the IRT model are tested in a practical equating situation.

6.5 Method

6.5.1 Participants

The key difficulty with the PENG design lies in finding a suitable cohort. They need to be familiar with the curriculum but not participating in the live examinations. For this experiment a suitable cohort appeared to be GCSE Science pupils in their second year of study at GCSE who had already completed the initial set of modules in their first year. It was hoped that they would have retained a good knowledge of the

curriculum and be motivated to further probe their strengths and weaknesses. 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.

6.5.2 Components

The modules taken by these pupils in their first year were from a specification that consisted of six separate multiple choice tests and a test of practical skills. To achieve a GCSE in Science candidates have to take two tests (A and B), in each of Biology, Chemistry and Physics, as well as the practical work. Each test is available three times a year, in November, March and June. Ideally a separate linking test would be constructed between each session of each test (A and B) in each subject. If this were to be done separately for each tier this would require twelve separate anchor tests to link June and November; another twelve to link November with March; and another twelve to link March and the subsequent June. In total this would represent thirty-six anchor tests. Even if the experiment were successful the logistics would be prohibitive.

A pragmatic solution to the tiers is to equate them using the existing common item structure as demonstrated in the previous chapter. This would reduce the number to eighteen. A further compromise would be to combine items from the A and B tests for each subject. This would reduce the number of tests to nine. A more radical solution is to combine Physics, Chemistry and Biology items into a single test. This reduces the number of tests to six per year. A test which combined A and B items from Physics, Chemistry and Biology could reduce the number to three. The solution chosen, following consultation with the examiners, was to combine Physics

A, Chemistry A and Biology A items into a single test. They felt that candidates would be comfortable switching between the different subjects. Only a single link, between June and the following November was attempted.

6.5.3 Design

Schools generally teach Science in periods of 45 minutes, which limits the length of the anchor test. It was calculated that candidates should be able to complete 30 items in this time, which meant that each subject would be allocated 10 items; 5 from June and 5 from November. This limitation on the number of items presented a further problem. The Science tests for this specification were based on groups of four items which follow a single stimulus. The stimulus can vary from one sentence to a paragraph with accompanying figures and tables. Although it is intended that the stimulus represents a thematic grouping for the following four items and does not reduce the independence of the items, inevitably the coverage of a single group is limited. Again, following consultation with the examiners, they felt that a better representation of the candidates' ability would be gained by selecting individual items from different contexts. This meant they could choose the best discriminating items that represented key skills and knowledge. Had the analysis from Chapter 4, on the weak local independence of items within these tests, been available at the time, this decision might have been different.

6.5.4 Sample size

A sample of convenience was chosen. A large number of centres were approached but only five centres (with 176 candidates between them) agreed to participate. The

low participation rate would need to be addressed if this method was to be pursued in the future.

6.5.5 Plan

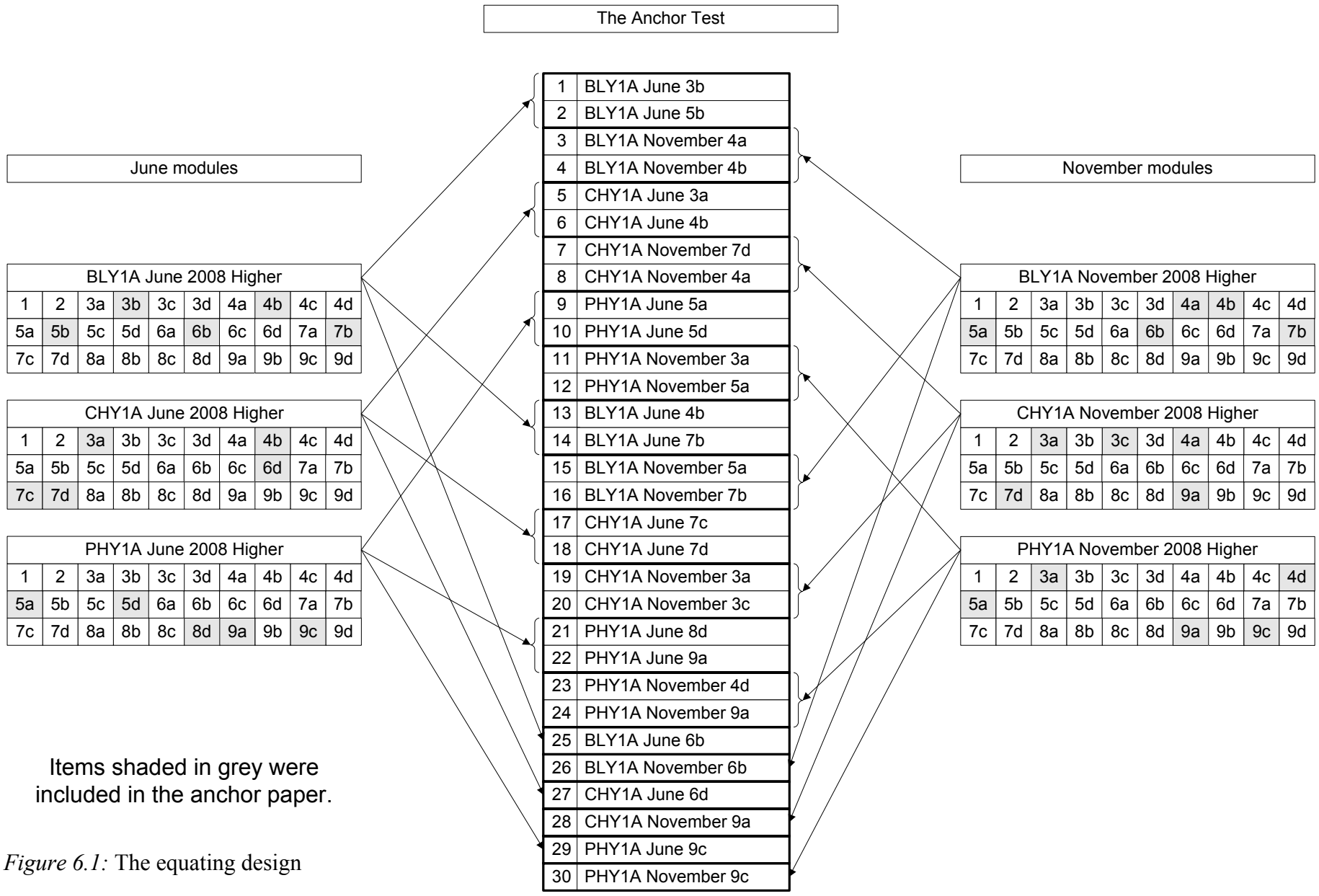
In order to construct the anchor tests the examiners were invited to a test-construction session. They were supplied with the items they had used the previous June, and items designated for the forthcoming November. Statistical information, means, standard deviations, discrimination indices and item-test correlations were provided on the items that had been used in June to aid their selection. The examiners were asked to work individually at first on their own sub-test, and then to collaborate, to create a comprehensive test of Biology, Chemistry and Physics. The collaboration was intended to avoid over-representation of certain generic scientific skills such as the interpretation of data. On completion they were satisfied that the test (illustrated in Figure 1) was indeed representative.

The test was then sent out to participating schools under secure conditions with instructions that it was to be taken on the day of the live test. Accompanying the test was a questionnaire which attempted to gain an insight into the motivation of the candidates and how well prepared they felt for the test.

Finally, once the tests had been returned and marked and the analysis completed the results were presented to the examiners. This allowed them to review how the items had performed on the different test versions and consider why they had behaved as they did.

6.5.6 Analysis

The OPLM method was used for the test equating. Analyses from the previous two chapters suggest that this will lead to better model fit and better short term predictions. Moreover, the sparseness of the equating design means that the use of concurrent estimation, which estimates item parameters based on information in both test forms, is more important.



Items shaded in grey were included in the anchor paper.

Figure 6.1: The equating design

6.6 Results

6.6.1 Descriptive statistics

176 candidates from 5 centres took part in the trial. Table 6.1 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 test as well as 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 were up to a year younger than the other participants in the trial and were likely to have poor motivation as they had just undertaken a proportion of the same questions in the live environment. All such live candidates (who came from a single 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.

Table 6.1: 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 15-year-olds were taken from the total entries, summarised in Table 6.2. A sample was required due to restrictions in the software. It may seem an odd decision to sample only 15-year-olds, as 16-year-olds took the anchor test, but the 16-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 6.2: Entries for the Science tests. Figures are not given for Physics as the equating was not successful

	June			November		
	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	76.10%
Chemistry	15,391	23,993	64.15%	55,937	73,049	77.31%

6.6.2 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 0.65, which is reasonable given the reliability of the trial test (Coefficient alpha = 0.63) and the live tests (Coefficient alpha = 0.72, 0.74 and 0.77 in June 08, for example). This provides some reassurance that the anchor test was testing the same construct as the other module tests. Disattenuated correlations were all over 0.9.

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 the topic of '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.

6.6.3 Context effects

It is apparent that whereas the majority of the Chemistry and Biology anchor items showed good fit to the model, the Physics items in the trial performed differentially in the live tests. Table 6.3 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 6.2. The final column in Table 6.3 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 multiple-choice question.

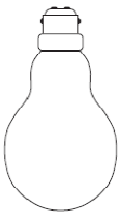
Table 6.3: Observed scores and expected scores derived from OPLM for the question in Figure 6.2

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

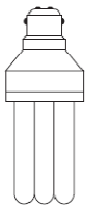
QUESTION FIVE

The diagram shows information about four types of electric lamp.

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

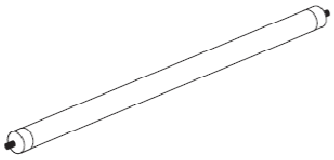


100 watt filament lamp
Average life = 1000 hours
Cost = £0.50



20 watt energy-saving lamp
Average life = 10 000 hours
Cost = £3.00

10 watt LED spotlight
Average life = 60 000 hours
Cost = £30.00



15 watt fluorescent tube
Average life = 5000 hours
Cost = £5.00

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?

Figure 6.2: An anchor item from Physics. Question 5A was presented in the trial without the subsequent questions 5B to 5D.

As Chapter 2 clearly highlighted, there is weak local dependence between these items which visual inspection does not necessarily reveal. Figure 6.3 shows the PPP values for the odds ratios for the 2-parameter model for the November Physics test, with the extreme values highlighted with either a triangle or a circle. The prevalence of low values confirms this weak local dependence. There is a higher proportion of candidates with the same response pattern to both items than would be predicted under a model that assumes local independence.

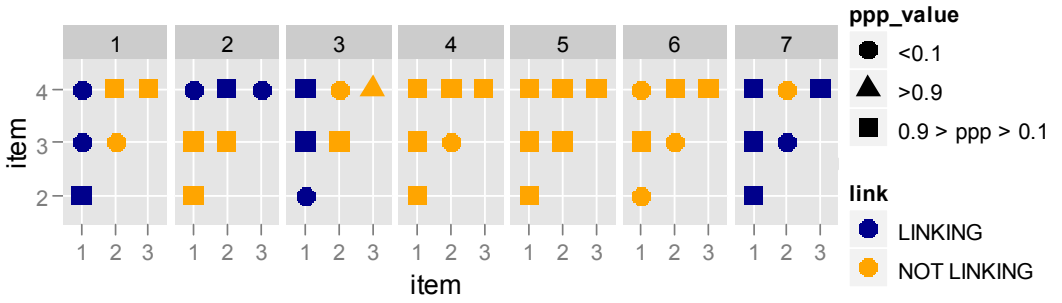


Figure 6.3: PPP tests for the odds ratios for the November Physics test under the 2 parameter model

6.6.4 Population dependence

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 6.4). 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 6.4: 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 strong suggestion, therefore, is that there is an interaction between further study in GCSE Science and the test items. Certain elements of knowledge may have been forgotten while elements of synthesis, analysis and understanding may have improved. Within this sparse design, however, it is difficult to estimate the impact of this interaction on the equating. More linking items could have resulted in some such estimation as specific items could be removed and the impact on the equating measured.

6.6.5 School effects

With such small sample sizes it is difficult to estimate the instability in the item parameters introduced by the use of only a few centres. If there were no centre effects then the item facilities would be reasonably similar between centres. Figure 6.4 shows a plot of the item facilities from one centre against another. While the scatter for Biology and Chemistry would lead to the conclusion that the candidates from these centres are of a reasonably similar ability, the scatter for Physics would imply that the candidates from centre B are relatively weaker. The relative strengths and weaknesses of the candidates could reflect the teaching as much as the ability of the candidates; both contribute to instability in item parameters.

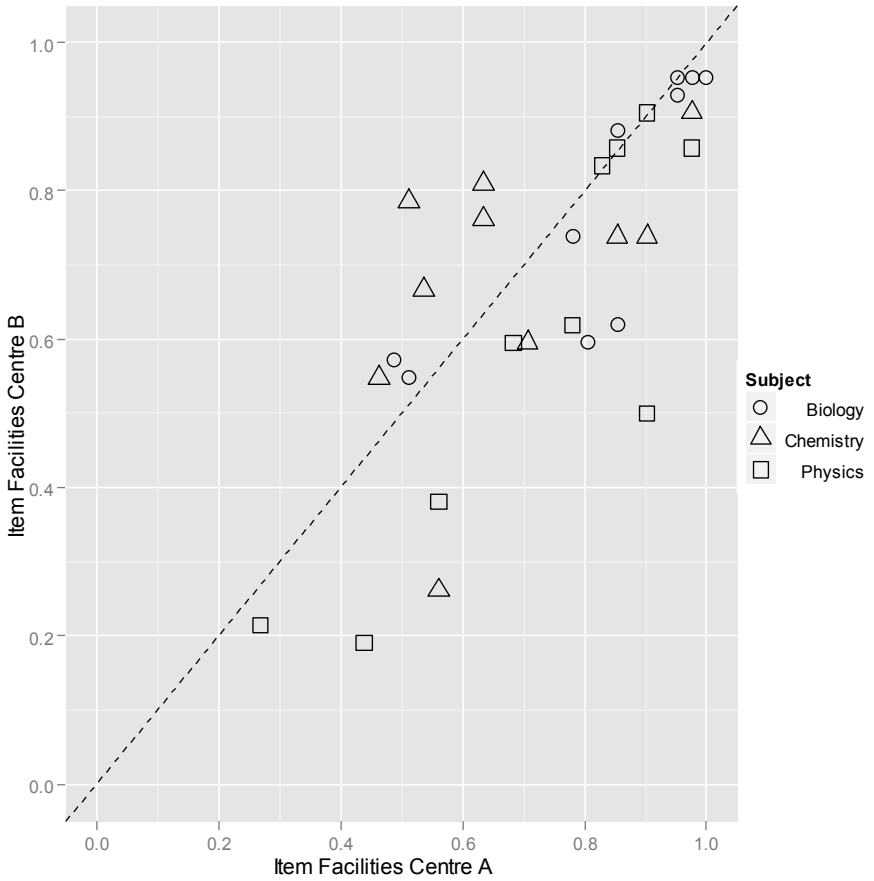


Figure 6.4: A comparison of item facilities for two of the trial centres on the anchor paper.

6.6.6 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 per cent 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 6.5. 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 6.5: Number of items used in the equating

	June		November		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 15-year-olds in June and 15-year-olds in November as the basis of the comparison with the test equating. Unfortunately prior achievement measures were not available for the 15-year-old cohort in November as this would have provided an interesting comparison.

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 was determined for each marginal population for each test. Figure 6.5 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 produced a pass rate of 71.60 per cent for 15-year-olds. The closest match on the expected cumulative distribution created from the sample of 15-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.

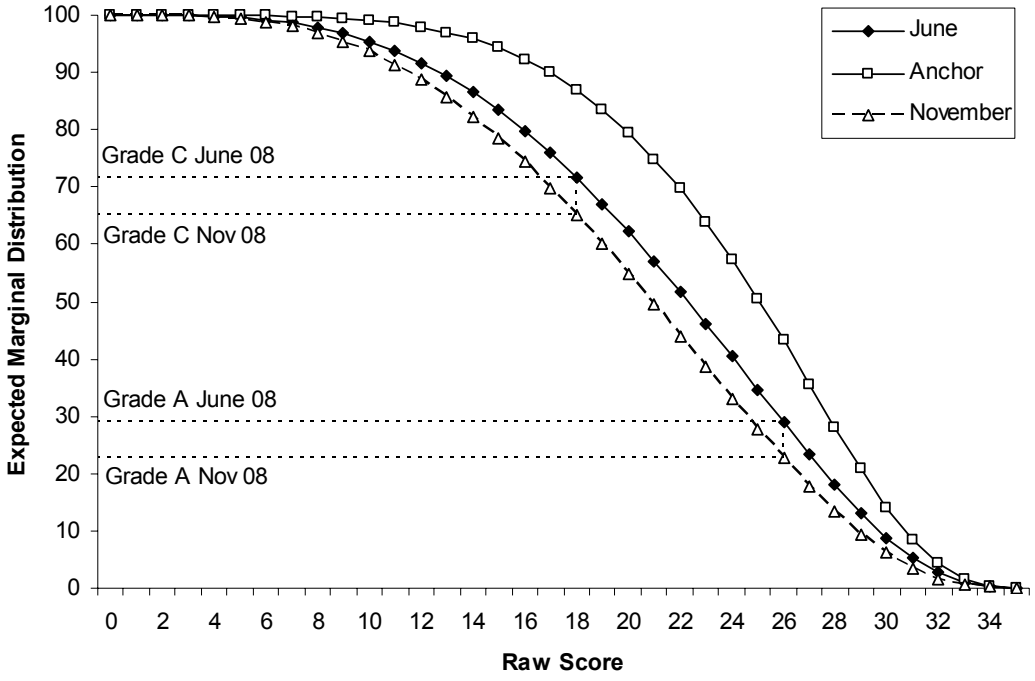


Figure 6.5. Equipercentile equating between marginal populations on the June Chemistry live test.

The results from the test equating, summarised in Table 6.6, suggest that the performance of candidates in November was worse than in June. These results suggest between five and eight per cent 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 population dependence and context effects already noted, the consistency of the findings across both subjects adds some weight to the finding.

Table 6.6: 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

6.7 Discussion

The purpose of this chapter was to consider whether a post-equating design could have a practical use in maintaining unit standards in the public examinations system in England. Its major finding was that there is an interaction between length of study in GCSE Science and Science test items. Certain elements of knowledge may have been forgotten while elements of synthesis, analysis and understanding may have improved. This replicates the finding of Cook and Paterson (1987).

To some extent, however, the study failed due to some mistakes in its design. The context of items was not respected across linking forms, and as these items display weak local independence the item parameters across forms were not invariant. This could have been predicted had an analysis on local independence been available at the time the experiment was designed. Visual inspection of the items for local independence clearly is no guarantee of that independence.

The sparsity of the design in terms of the number of common items and the low sample size also threatened the success of the study. Fewer than the

recommended 20 per cent of common test items between forms was used. This meant that where items did display differential functioning between forms they could not be discarded without a substantial loss of information to the test equating link. The lack of linking items also makes it difficult to discern which items are functioning differentially and which are not. A further consequence of the sparse design is that no estimation of the impact of the population invariance, school effects or context effects could be made.

In retrospect, therefore, the following recommendations could be made:

- That linking items that are presented in coherent blocks are not removed from their context even if they appear not to be dependent upon it
- That redundant items should be built into an equating design so that they can be removed if they are suspected of performing differently due to context effects, population dependence or school effects
- That a matrix sampling approach is adopted which allows more items to be linked and school effects to be monitored
- That the narrowest construct definition possible should be used. For example, Biology should be equated as Biology not as Science!

While many of these suggestions exist in the literature, it is difficult to appreciate their impact until the consequences of their violation become apparent in practice. Those behind the NAEP anomaly must have had a similar sinking feeling as the reasons for the failure of their linking test in 1985 seemed so easy to predict with hindsight.

A fringe benefit of the study once again relates to validity. The cross information in the Physics questions was not apparent on visual inspection. In other words, they appeared to have good face validity. In practice, however, it is clear that candidates could select the correct answer to the efficiency of a lamp through a process of elimination rather than from knowledge of the correct equation to apply. Improving the design in this case, however, is not straightforward. Presenting the item in isolation led to very able candidates using all the relevant information to come up with the wrong answer. Irrelevant information in the stimulus is a common source of construct irrelevant difficulty. Perhaps the process of elimination could be seen as a validation of candidates' knowledge therefore rather than cross-information.

The limitations of this study, however, should not reflect upon the feasibility of introducing a post-equating design into public examinations in England. A post-equating approach allows some estimation of performance standards between units taken in different modular series. In order to achieve this estimation assumptions are, of course, required. Apart from the basic requirement of unidimensional IRT, that the items must comprise a coherent scale, the item parameters must be stable across equating samples and test versions. While this research found that item parameters were not always stable across samples due to context effects, school effects and maturity effects, more money and time invested in developing anchor tests could yield more precise results.

A less obvious and unexpected advantage of the IRT approach is that a meaningful discussion on the test equating information could be held with examiners. As the test equating results are derived from information on the test items the examiners themselves designed, they could make meaningful comparisons

between their expectations of the difficulty of items and their actual difficulty. No such meaningful discussions can be had on statistics derived from prior achievement as the examiners have no basis on which to engage in those discussions. This is encouraging from a practical perspective.

Of course, it would be wrong to understate the practical difficulties involved in a post-equating approach. According to the current design of GCSE Science thirty-six anchor tests per year would yield the most precise results. Under this design every aspect of every subject within the qualification at every level between every session could be equated. This design is simply not practicable. A less complex specification design, however, would require fewer anchor tests. GCSE Science is exceptionally complex, as most qualifications are made up of far fewer units. Traditionally GCSEs are made up of two written units which are only available twice per year. If the qualification is not tiered then four anchor tests per year would suffice. A-levels have a standardised four unit design with each unit being offered twice per year. Again, one anchor test per session per unit would suffice. Overall fewer units and fewer sessions both make a post-equating design more feasible. Key subjects could also be prioritised.

The structure of certain tests, with questions grouped around a context, contributes to difficulties in equating. Candidates are trained that there is no irrelevant information in a GCSE context and therefore attempt to use every piece of information to answer the questions that follow. Even when the number of questions is reduced they will still persist in attempting to use all the information presented in the context although only one aspect of that information may provide them with the correct answer. More focused anchor tests or longer anchor tests could ensure that the contexts are preserved intact in equating designs. Of course IRT requires every

test to comprise some form of scale; tests that consist of one or two items would prove challenging to model in an IRT framework.

One area that cannot necessarily be addressed by more investment of time and money is the population dependence of the post-equating design. More schools could be recruited to reduce the school level effects and matrix sampling of items employed, but the choice of an appropriate cohort to undertake the anchor tests will always prove challenging. This study used candidates one year further on in their study; they found items testing scientific literacy relatively easier than the live candidates did. In a modular situation care would have to be taken to monitor interactions between length of study and the content of items.

Given the potential cost and complexity of the post-equating design, as well as the issues of population dependence, it seems worth exploring potential alternatives. Equating designs built on item banks provide an efficient and cost-effective method of pre-testing. While, for the purposes of this thesis, it is impossible to develop and try out an item bank it is possible to consider the ethical issues of pre-testing in live examinations. The next chapter will therefore consider the opportunities available in the future for IRT methods of test equating.

7. Test Equating for the Future

The post-equating design described in the previous chapter has some obvious practical limitations. Every specification would require a specifically designed post-equating experiment in order for test forms to be linked. This approach would provide useful evidence to the system of public examinations in England as it stands, but allows little scope for innovation in the way in which examinations are delivered. More flexible equating designs which are based on item banks can support more flexible delivery of examinations so examinations can be delivered more frequently and in personalised forms. This flexibility, however, can require compromises in security and raises some difficult ethical issues. The purpose of this chapter is to review the case for personalised assessment and consider how stakeholders would react to the changes that personalised assessment, underpinned by test equating approaches, requires.

7.1 On-demand testing

In '2020 Vision: Report of the Teaching and Learning in 2020 Review Group' Christine Gilbert, on behalf of the group, presented the following vision of what personalised teaching and learning might look like in a 2020 school to the Secretary of State:

> *"Personalising learning means, in practical terms, focusing in a more structured way on each child's learning in order to enhance progress, achievement and participation. All children and young people have the*

> *right to receive support and challenge, tailored to their needs, interests and abilities. This demands active commitment from pupils, responsiveness from teachers and engagement from parents."* (Gilbert, 2006, p. 3)

Teachers, according to the 2020 vision, are experts in the analysis of data, and use a mixture of formative and summative assessment to ensure that no student falls behind. All learners, regardless of socio-economic background, gender or ethnicity will achieve high standards, possess functional skills in English and Mathematics and understand how to learn, think creatively, take risks and handle change. Teachers will operate a fast-response system to ensure learners do not fall off their upward trajectory and parents will become their child's co-educators. The vision is clearly aligned with the Every Child Matters: Change for Children policy agenda (http://www.everychildmatters.gov.uk) designed to protect children and young people from harm and help them achieve what they want in life. While the report did not touch on a different role for high-stakes assessment, the concept of learning tailored to individual needs presents an opportunity for high-stakes assessment to evolve.

The implications of this policy agenda for National Curriculum Tests (NCTs) were drawn out in the consultation document 'Making Good Progress' in which the Department for Education and Skills (DfES) set out the case for making NCTs available on a when-ready basis. The emphasis is placed on the engagement of all, and the progression of all:

> *"The model could be a powerful driver for progression, raising expectations for all pupils, motivating them, bringing a sharp focus on 'next steps' and perhaps especially benefiting those who start the key stage with lower attainment than their peers, or are currently making too little progress."* (Department for Education and Skills, 2006)

More timely assessment information is a challenge to the existing infrastructure of high-stakes assessment in the UK which has been designed around a single major session each year. These sessions are designed and timetabled years in advance, tests for that series written at least two years in advance and entries gathered many months in advance. Once the tests have been sat in June the marking, awarding and marking review processes mean that most candidates will receive their results some three months after sitting a test, although appeals against results may not be settled until the following year. Such a labour intensive process is clearly not conducive to delivering more timely assessment information.

Modernisation of the infrastructure of assessment is ongoing, however, and leading to some changes. At AQA, for example, most marking of short response answers is now being done on-screen, the training of markers is being done on-screen, the technical infrastructure for remote awarding is being set up (at EdExcel it has been in place for some time) and the first on-screen tests themselves are also available. The days of large volumes of examination scripts being shipped from the awarding body to the examination centres to the markers and back to the awarding body are coming to an end. It is now possible to imagine a situation where candidates take their examinations on-screen, their answers are marked on-screen, and results delivered electronically. With technological change comes the potential

for improved efficiency in processes. The responsibility for the quality of marking, for example, used to reside with senior examiners. They would mark and select certain scripts to be used to standardise and quality control the marking of other examiners at certain points before and during the marking period. On-screen marking has made this process obsolete through peer-pairing of items by which two markers will mark the same item and any discrepancies are highlighted. This is a much more flexible system that can be fine-tuned to achieve the desired balance between quality and efficiency (Pinot de Moira, 2009c).

Is it also possible therefore that test equating can be combined with on-screen testing and on-screen marking to deliver improved efficiency to the assessment process, but as with on-screen marking (see for example Fowles, 2009), the sort of tests that are conducive to this approach need to be carefully evaluated. Test-equating does hold the promise, at least for short-answer tests, of the grade boundaries being known before a test is taken. This would be an essential component of establishing tests on a when-ready basis. In the absence of large stable cohorts the item information rather than the cohort information becomes central to maintaining standards. While the efficiency gains, however, of being able to maintain standards without the need to convene a committee are obvious, however, the case for on-demand testing is less so.

The experience from the US suggests that large scale achievement tests whose scores are only needed once a year are the worst suited to testing on-demand (Wainer, 2000). The most popular dates for the SAT®, which is available on demand, are a Saturday morning in December and in January. This is the latest date at which results are necessary for college admissions, giving students the most time available to study. In the UK entrance to universities is actually determined largely

by predicted grades rather than achieved grades, so arguably there is a need to determine the achievement of candidates at an earlier point. It is not the inefficiency in the assessment system that means that candidates do not know their grades at the point of application for admission, however, as awarding bodies have now shown that they can meet the deadline required for a system of Post Qualification Admissions. Further, with the current A-level system there is no reason why candidates could not finish their studies early, in January rather than June of their second year of study, but the vast majority continue to study new material and leave their final certification until the end of their two-year course. High-stakes achievement tests seem most suited by their very nature to mass administration on certain dates. Low-stakes tests where item security is not an issue, licensing tests where results are required immediately and vocational tests which offer more realistic simulations of skills required are identified as better candidates for testing on-demand (Wainer, 2000).

On-demand testing is, however, a spectrum. The existing modular systems are more on-demand than the systems they replaced. These modular systems are popular and marketed on the premise that they offer timely information on candidates' progress,

> The modular Mathematics specification 2381, for teaching from September 2007, offers a flexible, modular route to GCSE Mathematics. It is intended to motivate students by giving both formative and diagnostic feedback from the modular tests throughout the course, enabling teachers and students to identify any weaknesses and remedy them. (Edexcel, 2009)

Does this mean that high-stakes testing is finally bridging the gap between formative and summative assessment and making formative testing irrelevant?

Surprisingly the debate on whether an on-demand high-stakes assessment could play a significant formative role was played out in the UK in the 1980s when a paper-based on-demand system of mathematical modules known as the Graded Assessment in Mathematics Project (GAIM) was developed with the original intention of providing an alternative path to a GCSE (Brown, 1989). Based solely on coursework, even its critics agreed that it provided an interesting and excellent basis for curriculum development. The authors of the programme claimed that it was the continual flow of diagnostic information that delivered excellent outcomes: its critics attributed increased outcomes to a flawed equating model (Noss, Goldstein & Hoyles, 1989). The technical argument is hard to resolve as the outcomes from different modes of assessment will always be difficult to equate. The argument against the theoretical standpoint that better outcomes were to be expected due to the diagnostic features of the GAIM assessment model is, however, worth repeating. Critics of these gains argued successfully (the GAIM model was never accepted for GCSE certification) that schemes that attempt to provide both grading and diagnostic information are fundamentally unviable and educationally unsound.

The evidence for educational gain through formative assessment comes from a particular model that prioritises dialogue and reflection which builds the self-esteem of the learner (Black & Wiliam, 1998). When diagnostic feedback is accompanied by a grade the feedback loses its worth: grading encourages the suppression of a student's weaknesses and a concentration on maximising assessment ratings or test scores (Noss et al., 1989). Grading dulls the message about what it means to improve, so summative assessment has limited use where teachers have little control

over setting the assessment content or marking (Black, Harrison, Lee, Marshall & Wiliam, 2003). As for the claims of greater student motivation, the wider psychological literature originally suggested that the provision of extrinsic rewards is likely to have a damaging effect on intrinsic motivation (Deci, 1975). Since Deci's original study countless studies have been conducted to examine the effects of reward on intrinsic motivation under different contingencies and in different circumstances. There are some who argue that under certain circumstances, such as low academic interest, reward can increase intrinsic motivation (Cameron, 2001). Even if this were true, however, this model would seem to assume a performance standard in which everyone makes the grade,

> When students work hard within a stringent reward contingency and then do not get a reward, the experience is likely to be highly detrimental both because the contingency tends to be controlling and because not getting the reward will probably be experienced as failing. Thus, even if the intrinsic motivation of the few who receive a reward is not diminished, it seems quite likely that the intrinsic motivation of those who do not receive a reward will be destroyed. (Deci, Ryan, and Koestner, 2001, p.48)

In a modular examination setting therefore, it would be important to evaluate the size of the effect on motivation as, in practice, it may be negligible or it could dominate.

More worrying still, in the context of GAIM, Noss et al. (1989) argued that the provision of both grading and diagnostic information in a single scheme can be extremely damaging when a hierarchical model of learning is, without theoretical or

empirical underpinning, turned into a recipe for curriculum sequencing. If on-demand testing leads to smaller, more carefully defined steps through a curriculum, the epistemological and psychological distortions produced by this didactical transposition should be made explicit. To manage this risk, much closer links with pedagogy are required to avoid the potentially damaging consequences of ill-conceived modularisation. Nor will this process be simple, as there are no clearly agreed steps to learning that work for all children in all areas of learning. An examination that samples a curriculum after two years of learning, ignorant of the path which that learning has taken, clearly poses far fewer risks to pedagogy.

There are probably cheaper and more effective ways of delivering formative assessment than more regular high-stakes assessment. Any claims that the diagnostic nature of high-stakes tests has led to genuine gains in understanding that feed through to increased outcomes should be viewed with scepticism. Tests have uncertainties associated with their outcomes, and where the best results can be banked candidates are likely to improve their scores by re-taking simply through chance. As Black (2007) stated simply, "test again and again and again – standards will go up". The case for formative gains achieved through the modular system is weak; the evidence of candidates improving their scores through re-takes is strong (Spalding, 2009).

Recently, there have been moves to tighten the system of modular assessment to prevent these gains. For new GCSEs, introduced from 2009, there is a limit on one re-take per candidate per unit. A maximum of four units is permitted for GCSEs (and each unit must have a weighting of at least 20 per cent of the marks of the total qualification) and candidates must take at least forty per cent of their assessment in the final sitting. This clearly restricts the scope for further modularisation. Indeed, a

change of government could lead to a rolling back of modularisation, which they hold partly to blame for a decline in examination standards,

> Modularisation of A-levels has contributed to (this) loss of depth, and encouraged compartmentalisation of knowledge, reducing the opportunity for students to develop the ability to integrate information. (Sykes, 2010, pp. 12-13)

The current coalition government while in opposition planned to restrict the number of assessments being taken and the time spent in preparation for those examinations (Ryan, 2010).

7.2 Beyond linear testing

Test-equating is a pre-requisite to the establishment of calibrated item banks. These banks could be used to facilitate on-demand tests in a traditional paper and pencil format. Once an item bank has been established, however, new possibilities in testing open up.

7.2.1 Computer Adaptive Testing

Computer Adaptive Testing (CAT) was conceived by Lord (1980) as a way of providing an individually tailored test that could be mass–administered. An adaptive test is one that adapts the difficulty of the questions offered to candidates to suit their ability as illustrated by their response pattern. Thus, if a candidate fails to answer a question correctly an easier question is presented. If this question is answered

correctly a more difficult question is presented. This process continues until the candidate's ability is measured to a predetermined degree of accuracy. Green (1983) outlined the major advantages expected of CAT as improved test security and an appropriate level of challenge for all candidates. Improved test security was expected as any one candidate would only see a small proportion of the total questions in the test pool: if this pool is large then learning the pool would be analogous to learning the subject (Wainer, 2000). An appropriate level of challenge would ensure that time was not wasted on questions that were too easy or too hard for candidates: the brightest would be challenged while the weakest would not be discouraged.

The appropriate level of challenge has indeed proved a popular feature of CATs. In the United Kingdom the largest operational CAT is the Computer Adaptive Baseline Test (CABT) offered by the Curriculum Evaluation and Management Centre at Durham University. In 2005 over 100,000 adaptive tests of mathematics and vocabulary were delivered to 11 to 16-year-olds using a Rasch-based adaptive algorithm. The tests have proved reliable psychometrically with a test-retest reliability above 0.9 and been welcomed by teachers as improving the testing experience of students (R. Coe, personal communication, 2009). Used as a baseline test, however, the CABT has the advantage of being delivered in a low-stakes environment.

In a high-stakes environment CATs have proved to have significant security flaws. The problem with CATs is that item selection algorithms do not choose all items with equal likelihood and a very small proportion of the item pool accounts for a large amount of the items administered (Wainer, 2000). A common finding is that between 15 and 20 percent of the item pool accounts for more than 50 percent of the test items being administered. This occurs because the distribution of difficulty of

items in the item pool tends to match the ability distribution in the population. The result is that the tests delivered overlap considerably, especially for the most able students. These able students are precisely the students who can reproduce the items they are asked most accurately. According to Wainer (2000) Kaplan Educational Centres were able to exploit this flaw to methodically steal a large proportion of the item bank being delivered by Educational Testing Services (ETS) for the American high-stakes test, the Graduate Record Examinations, the largest operational CAT in the world. Modelling how this was possible McLeod and Schnipke (1999) found that by asking eight candidates to memorise the difficult items they received a low-scoring examinee could use this information to increase their score by three standard deviations.

Although the case against Kaplan was never proven, and ETS denied that the security of the GRE had been compromised (Frantz & Nordheimer, 1997), item exposure, which models ways in which item pools can be more effectively utilised and test overlap limited, has become a major field of study. ETS withdrew the CAT version of the GRE from 2007 in favour of linear tests, citing security concerns as the main reason, although it has now been reinstated (Educational Testing Service, 2009). There are many other successful CATs, but these are generally employed in fast-moving technology fields which require detailed knowledge that can be easily varied. Building up the research expertise to support a high-stakes CAT (ETS released the first report on the adaptive version of the GRE in 1995) would, however, take substantial resources.

7.2.2 Multi-Stage Testing

Because of the security concerns surrounding adaptive testing, there is a perceived need for item-level adaptive tests to be constructed "on-the-fly" using some form of automated test assembly but this presents limitations to their use. Complex specifications may need to be relaxed for their use as it may not be possible to meet all competing requirements simultaneously, while design flaws can cause unintended test assembly issues. Some examinations may also have content requirements that are difficult to quantify or implement as rules. To deal with these concerns, and to ensure that stakeholders have sufficient input into the test construction process, an alternative design known as Multi-Stage Testing (MST) has been implemented (Mead, 2006).

Multi-Stage Testing has the same aim as CAT: to shorten the test length while optimising discrimination. While CATs require complex algorithms to be built into test players to decide the selection of the subsequent item on every case, MSTs have built-in paths that lead candidates through a series of testlets. Depending on a candidate's score on a particular testlet, they are directed to a subsequent testlet. Figure 7.1 illustrates a 1-3-3 module computer adaptive sequential test (CAST) configuration (Luecht, Brumfield & Breithaupt, 2006). The possible routes through the seven testlets are indicated by the solid and dashed lines. Most examinees are expected to follow the solid pathways; the dashed lines compensate for unexpected performance. Some pathways, for example from 2E to 3H are precluded. The seven testlets and the associated routing rules are packaged together in units called panels. Figure 7.1 depicts multiple panels which can be assigned to examinees just like multiple test forms.

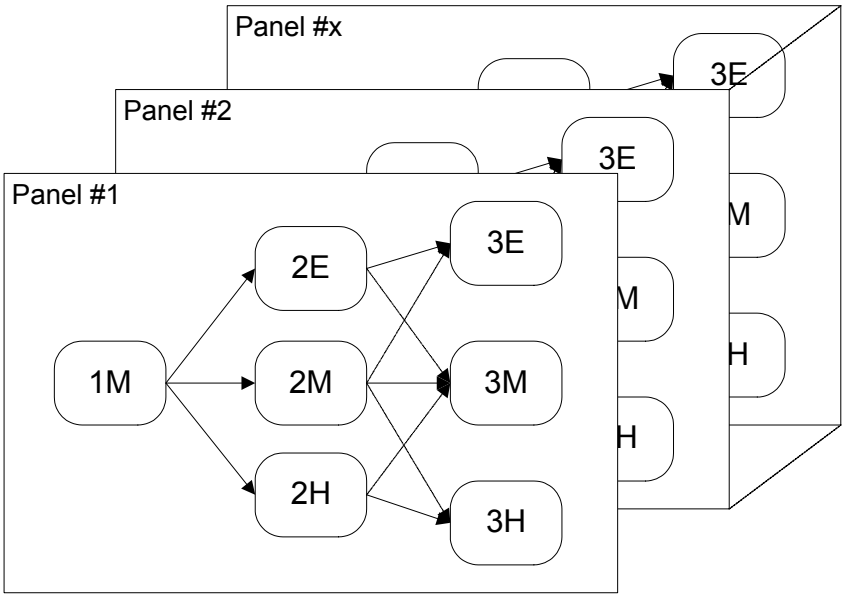


Figure 7.1: Design for a 1-3-3 computer adaptive sequential test configuration with multiple panels. E=relatively easy; M=moderately difficult; H=relatively hard. The different panels represent different versions of the same test that may be used when test forms are spiralled to different samples of candidates.

This design alleviates many of the problems found with CAT: MST developers should never get back their results and find that 20 per cent of the items in the pool have been used on 80 per cent of examinees' tests as is common in CAT settings (Wainer, 2000). All tests can be subject to the same quality review procedures as are currently in place for general qualifications, and the test delivery software does not have to handle complex scoring and item selection algorithms. This approach would seem particularly suited to general qualifications which have struggled with the problem of differentiation since they were re-launched in 1988 with the brief to emphasise positive achievement while retaining optimal discrimination (Good & Cresswell, 1988b). The current approach, tiering, has significant technical flaws (Wheadon & Béguin, 2010) and has been criticised for the need to allocate candidates early on in their course of study to a level (or tier). MST leaves the decision until the last possible moment, and makes the judgement on

objective information available at the time of testing. The decision to use such models concerns more than just the tests themselves: whether or not candidates of all levels of ability should follow the same syllabus, an assumption to some extent implicit in this model, has wide-ranging implications for education.

7.3 Ethical questions

The achievement of personalisation and flexibility clearly requires changes to test delivery that may be contentious. The need for some overlap between the test-takers and the tests offers the assessment agency choices that may be more, or less, palatable to stakeholders. One question that arises, for example, is the extent to which pre-testing can be built into live testing to support a test equating design. Would candidates mind? Would their teachers mind? Would their parents mind? It is also extremely challenging to make IRT test equating designs as secure as the current system that keeps all test items secret until the day and time of the designated test, and then releases all those items for public scrutiny, never to be re-used. To what extent are stakeholders prepared for the security of testing to be sacrificed, in some degree, in exchange for more flexibility? To start to understand how stakeholders in the system (teachers, pupils and examiners) value the way in which qualifications are currently delivered and what they perceive as the benefits of a more flexible system requires a qualitative approach. It is important that such stakeholders understand the advantages and disadvantages of the system that is being built, for where public confidence is grounded in false beliefs, those beliefs are likely to be challenged and trust threatened (Newton, 2005b). In the competitive market for

general qualifications, a loss of confidence could also be accompanied by a loss of market share.

7.4 Methodology

The aim of this part of the research, therefore, was to explore opinions, attitudes and values of teachers, pupils and examiners regarding the flexibility of examination delivery. While other groups are obviously of interest – parents, governors and the regulator – to some extent these groups are all guided by the views and experience of teachers and pupils. It is possible to conceive of a matrix of qualification level, school sector and subject that should be sampled to gain a view across qualification, sector and subject, but such a matrix is beyond the scope of this research. Instead the focus was placed on the examination of GCSE Science in the state sector. GCSE Science lends itself to a flexible, modular approach and attracts the largest entries from the state sector. Focus groups were chosen as the appropriate methodology as they allow for a full discussion through which individuals can debate issues and thereby achieve clarity on their own positions (Stewart & Shamdasani, 1990).

7.4.1 Participants

7.4.1.1 The learner focus groups

In total, sixteen learners participated in the learner focus groups, which were held on three separate occasions. The first group consisted of four first year university students who were asked to draw on their experiences from school. It was hoped that they could offer a mature, but recent, perspective on examinations in general. An opportunistic sample was taken from a local university without particular reference

to the subjects they had studied or the type of school or college they had attended. This is the only group who were not asked to draw on experience of GCSE Science. The second and third learner groups consisted of six year 10 pupils from a selective state school, and six year 10 pupils from a non-selective state school, all studying for the modular GCSE Science syllabus.

7.4.1.2 The teacher focus groups

The teacher focus groups were held on two separate occasions: the first with three teachers from the GCSE Science department of a selective state school, the second with nine participants on a GCSE Science day. These nine participants all came from the state sector; their schools varied in size from a small state boarding school to the largest comprehensive school in the county with over 2,000 pupils. To supplement these groups, the Deputy Head from a Special school was interviewed separately.

7.4.1.3 The examiner focus group

Only three examiners were recruited for the examiner focus group: the three Principal Examiners with current and direct involvement in compiling the modular GCSE Science A question papers. This group was chosen as they had experience of working with a large number of different structures for science at this level, going back to the 1970s, and are currently responsible for producing the examination papers for the new syllabus.

7.4.1.4 Sample size

The last round of focus groups with the teachers and pupils revealed enough redundancy in the data to imply that the sample size, for GCSE Science at least, was

adequate. This pragmatic approach to sample size, albeit in a more systematised manner, is common in grounded theory (see for example, Charmaz, 2006). For this reason, no more focus groups were held. In terms of generalising beyond GCSE Science, the pupils' views did not seem to relate to particular subjects or levels, however examiner and teacher views for other subjects and levels, and teachers' views from different sectors on certain aspects, could be quite different to those reported here.

7.4.1.5 Group dynamics

The group dynamic aspect of focus groups was problematic for the first teacher group as the presence of the Head of Department may have led to motivated responding. Given more time and resources, interviews may have led to more authentic responses. For pupils, the same risk of motivated responding is present, but in their case interviewing is likely to lead to more passive responses. The examiner focus group used three examiners who had worked together for many years: as it was their shared experience that was of interest, and there was no hierarchical structure present, this was not considered problematic. As a close group, however, they may represent a shared perspective; ideally other teams would have been included to see if there were other genuinely different perspectives.

7.4.2 Procedure

For all the focus groups, the aims of the research were clearly shared with participants, but the scene-setting was deliberately vague in order to elicit broad areas of concern rather than points of specific detail. For the pupil focus groups, two researchers were present, one researcher acting as a facilitator and the other as an

observer. The roles of the researchers and their employment by an awarding body were explained to the pupils. For the teacher and examiner focus groups, an expert on the technical aspects of the topic was also available, where the participants sought clarification. The expert mode was deliberately limited in order to allow the participants to explore their own conceptions of more flexible examination delivery. The facilitators' role was to present unstructured questions and intervene only to maintain the productivity of the groups. In this case, the participants would be brought back to the original topic, moved to another area that had been highlighted in the discussion, or moved on to a new topic altogether. The idea was to maintain as fluid a discussion as possible to capture key areas of concern rather than dictate them. The questions were adapted for the different groups to reflect the different roles that the teachers, pupils and examiners play in assessment, and were refined after each focus group session.

The examiners were presented with two scenarios for on-demand testing, and asked to think through and discuss the implications of each. In scenario 1, tests would contain a built-in 'anchor' of around ten items, which would be repeated across series in order to maintain standards. The candidates would not be told which items comprised the anchor and the items in the anchors would not contribute to their scores. In scenario 2, each candidate's test would contain one or more randomly allocated pre-test items from an item-bank. The candidates would not know which items these were, and they would not contribute to their scores. The pre-tested items would then constitute the live items for the subsequent session.

The pupil focus groups were initiated by the facilitator asking for ideas on what on-demand testing may mean, and what advantages and disadvantages of different levels of on-demand there could be. They were asked to focus on the

frequency of testing they would like to see, the issue of re-sits, who should choose when they are ready, how different systems would affect the way in which they prepare for examinations, the involvement of parents and whether they would feel pressurised by different levels of on-demand. Following this general discussion, they were presented with some aspects of the scenarios of on-demand that had been presented to the examiners and were asked to comment on the fairness of these scenarios and whether they thought they could induce cheating.

The teacher focus groups began with a prompt on whether three test windows a year are enough and whether some students are more ready than others at these set test windows. Different levels of flexibility were then outlined and advantages and disadvantages discussed. In particular, the teachers were asked to comment on the perceived impact of changes on teaching and learning, school logistics, and the way in which pupils and their parents may react to the changes. Following this general discussion, they were presented with the same aspects of the scenarios of on-demand as the pupils and were asked to comment on the fairness of these scenarios and whether they thought they could induce cheating.

Consent was obtained from all participants in the study to record and transcribe the focus groups; participants were assured their responses would remain confidential and would be anonymised in any resultant reports. The data were analysed to identity dominant themes. Two researchers, one who had acted as the expert and one who had acted as the facilitator with the teacher groups, considered the transcripts separately and coded the data into themes. They then conferred and reached agreement on these themes. While this approach is subjective, it provides a useful descriptive overview of the data, and it is reassuring to note that there was little disagreement over key themes.

7.5 Results

Ten dominant themes were derived from analysis of the transcripts. The first seven themes relate to how the provision of a more flexible examination system would impact on the success of the pupils, namely: examination stress, level of flexibility, the gap between teaching and testing, group examination preparation revision and post-mortems, parental involvement, re-sits, and good data. The remaining three themes - fairness, practical considerations, and security - relate to issues associated with the mode of testing necessary for a flexible system to be implemented. Each of the ten dominant themes will be discussed in turn.

7.5.1 Reducing examination stress

The dominant themes that emerged from the focus groups were unsurprisingly linked to how the provision of a more flexible examination system would impact on the success of pupils. Both teachers and pupils referred frequently to the tiredness and stress induced by examinations and how reducing this stress would lead to better examination results. There appeared to be three major sources of stress: the amount of teaching to be done before the examination, the compressed nature of the examination timetable and uncertainty about results. Both teachers and pupils referred to panic revision and cramming as they ran out of time on syllabuses:

> *'we rushed through three lessons in one and a mock and no extra practice, we could have done with a fortnight.'*
>
> Karen, Science teacher

Part of the problem was identified in the difficulty of planning teaching for modules with unequal weights. Some can be taught in plenty of time for the examination while others are inevitably rushed. Another aspect of the problem lay in dealing with unexpected absences:

> *'whilst maybe four or five out of the eight classes were ready at the beginning of November we might have had a second sitting for the classes who have missed lessons due to trips and fire bells.'*
>
> Karen, Science teacher

When it came to the examinations themselves, the teachers were particularly aware of the stress induced by the examination timetable on their weakest candidates, whose re-takes were timetabled on the same day as examinations they were entering for the first time. Even a couple of days' grace, they felt, would make the difference as re-take pupils were described as 'exhausted' and 'frazzled' by the experience. The pupils agreed with the need to space out examinations, and were unanimous in preferring a modular system, feeling that the alternative is:

'too stressful 'cause you've got all the other exams at the same time.'

Sophie, year 10

The modular system also means that some results are in the bank – which reduces the stress of the final examination period:

'you already know all your grades so you know that it takes the pressure off if you've done well in the other ones.'

Ellie, year 10

To mitigate the stress of examinations, pupils and teachers referred to the therapeutic process of discussing the examinations after they had occurred and a sense of group solidarity that develops amongst the pupils:

'it's still like scary to take it all together but you know like everybody's in the same boat, you're all doing it at the same time.'

Jane, first year undergraduate

7.5.2 Level of flexibility

Given the popularity of the current modular system, it was not surprising that the offer of more examination dates was welcomed. Specific gaps were identified for certain syllabuses where an examination opportunity would have made sense, and there was general enthusiasm for a system which could offer assessment windows

every two months. Separate Science candidates, for example, follow a different schedule to Single Science candidates, but they take the same core modules with the same choice of examination dates. Different examination dates may allow the demands placed on these different streams to be better managed. The teachers also felt that if more dates were available the dates they chose could be fine-tuned to allow for different rates of progress by different classes.

There was little appetite amongst the pupils, and outright opposition from the teachers, however, to any system that would allow pupils within the same class to take examinations at different times. Most pupils, particularly those with the most maturity (the university students), saw the potential pitfalls of being offered this utmost level of flexibility and choice. Comments did range, however, from the optimistic,

> *'you'd know when you're ready and you'd probably pass,'*
>
> Alice, year 10

to the foreboding,

> *'you will get the people who will do it, you know for the short term just go "oh, can't be bothered" but then really suffer for it in the long term when they get to their exams.'*
>
> Sophie, year 10

The pupils were worried about the level of responsibility implied by too much flexibility:

'I think having to choose a date is just more stress than there needs to be, a fixed date you've got something to aim for,'

Sam, year 10

and agreed that having a fixed date to aim for reduced their stress levels. They were also worried about whether they may get left behind:

'when other people go ahead I'd be left behind … … I'd feel like worse they're going on and doing exams and I'm not ready for it yet. So it's better to do it all at once,'

Jane, first year undergraduate

and the impact of other pupils being left behind:

'you're going to end up with kind of people at different stages which is going to make the teaching less efficient.'

Nick, first year undergraduate

Only one teacher was confident her pupils would be self-motivated enough to set their own targets; the other teachers felt that the pace of lessons would suffer as pupils would not be motivated to take their examinations early.

Teachers were more outspoken on the negative impact of a system which offered flexibility at an individual pupil level. They echoed the need for dates as targets, as the dates promote group solidarity and allow pace and focus to be maintained in achieving those targets:

'I think it would be an absolute nightmare I really do. What I'm looking at is if I'm looking at a class that I take, and half of them have had an exam and the other half haven't, what do I do with the half that have? 'Cause they now switch off and then. I think it would have to be a whole class at a time … … I know where you're going with the flexibility but I don't think it's going to work,'

Brian, Head of Science

'you still have to teach a class as a whole. You can't let some students skip straight ahead, take a bunch of exams and move onto other stuff whilst others haven't even taken the first set of exams yet, so then you'd get some students whizzing ahead of others but there is still 24 kids in the same classroom. You'd have problems with the pace of the group as a whole. So if all students in the class are heading for a common date, and we all move on together.'

Karen, Science teacher

All agreed that any system which involved pupils within a class taking examinations at different times was a recipe for mayhem which would achieve little or no benefit.

7.5.3 Testing when fresh

Taking a broader view of success as effective learning rather than good examination results, a potentially harmful side-effect of more flexible delivery of examinations is

the loss of synopticity – pupils learning and being tested on discrete units may not draw on and benefit from earlier learning. This could lead to superficial coverage of topics and more emphasis on teaching to a test. The benefits of further learning having the potential to improve performance on earlier topics was only raised as an issue by the university students:

> *'once I had moved further on then, the more experience with it made me understand the other stuff much clearer,'*
> Nick, first year undergraduate

> *'I took most of my AS level modules again in A2 cause I didn't do very well at AS cause – I don't know, but it was so much easier to understand once you'd done the whole topic, so it was pointless me taking the modules in the first place, you know.'*
> Jane, first year undergraduate

but even amongst themselves they disagreed:

> *'erm well I quite liked it when I could actually move on from something that I'd understood and I didn't like to keep going over things when I was getting it so I would have kind of liked to have been able to move on to the next stuff*

> *Yeah, just do an exam and move on'*
> Tom and Emma, first year undergraduates

The GCSE students certainly wanted to be tested while their learning was 'fresh':

> *'it's much easier to do it when you've just done a topic so then you can just quite easily do the test and you won't have to revise as much 'cause you'll be able to remember everything you've been taught,'*
>
> Claire, year 10

> *'it's nice and easy, I've got, you know, my modules and so I only have to revise a small bit and then there's the test.'*
>
> Luke, year 10

One pupil even described the process of moving on and then taking an examination on the earlier content as confusing. Synopticity was not mentioned by the teachers, although such concerns are probably related to the level of study, the subject, and possibly even the syllabus. One teacher at least, however, was unconcerned about the gap between any teaching and testing, feeling that effective revision could compensate.

7.5.4 Group examination preparation, revision and post-mortems

A key aspect of success in examinations was identified as effective revision. Neither pupils nor teachers wished to sacrifice class examination preparation, revision and examination post-mortems in order to achieve flexibility. While most would prefer to revise alone, all saw the need for some group revision:

'it would be nice to know that you could revise with a group of other people if you get stuck.'

Jane, first year undergraduate

If pupils were entered at different times for an examination there was a worry that this group revision would suffer:

'they'd give revision sessions but I think if only one of you's doing it then I don't think a teacher would stay behind just for one person,'

Alice, year 10

and the lessons would suffer:

'and then other people are doing revision and other people have still got to learn.'

Alice, year 10

If everyone was given a different examination paper the pupils and teachers worried that the examination post-mortem would suffer:

'you can't actually go, you can't give meaningful feedback to them if every question paper in the room is different,'

Brian, Head of Science

but it was unclear why this post-mortem was felt to be important. Some of the pupils assumed, for example, that they would receive different mock examinations – this is unlikely to be the case.

7.5.5 Parental involvement

One teacher felt that one aspect of the success of their pupils was down to effective communication with parents, a process that would be made more difficult with flexible assessment:

> *'in this school parents are pretty much on the kids' cases so if they know the exam dates they can be sure they're revising at home, and I think that's crucial too, and I think it has a big impact on our success.'*
>
> Veronica, Key Stage Co-ordinator

From the pupils' perspective the level of involvement and understanding of their parents varied:

> *'mine don't really understand the situation at the moment let alone with them happening throughout the year so they don't really know when we're doing the actual thing from when you're doing the mocks and when we are doing practice ones,'*
>
> Claire, year 10

'...mine, mine are just like yeah ok, revise and then they don't really get it, 'cause they don't have a clue what I'm learning about at any point and they don't have a clue when I have tests

'Ah no, no, no, my mum, my mum gets involved.'

Sophie and Luke, year 10

Accountability to parents, however, was real to both pupils and teachers, and this aspect did raise concerns in relation to a more flexible system:

'erm, I'm not sure, I think she'd probably assume that I'd just put it off until the very end of the year and then have a panic revision and try and cram,'

Luke, year 10

'the people who do the same option would do the exam at the same time. Otherwise we would have parents on the phone, a nightmare.'

Brian, Head of Science

7.5.6 Re-sits

More flexible examinations offer the opportunity for more re-sits. In general, the pupils didn't feel this would change their attitude towards re-sits in any way:

'it's more like a safety net. It's a good thing to have, just so you know that if there is some sort of mitigating circumstances then you can always do it again some other time rather than just completely fail,'

Nick, first year undergraduate

'if you haven't passed it in like three times then you just know you're never gonna pass it,'

Claire, year 10

One pupil, however, felt that more re-sit opportunities would have an adverse effect:

'well if you had unlimited you'd just say, always think, yeah but you'd always think, you'd never try as hard as you can in any of the tests 'cause you'd always think if I screw it up this time I'm gonna just do it next time.'

Matt, year 10

7.5.7 The quality of feedback

The final aspect of success, related to a more flexible examination timetable, was the availability of good results data. One worried that too much flexibility would lead to poorer data. If different classes are taking examinations at different times:

'heads of departments in Key stages will find it really difficult to have an overview of the year group.'

Veronica, Key Stage Co-ordinator

Different tests for different pupils could also impact on the perceived quality of data:

'is it fair to re-organise teaching groups when they didn't all take the same exam?'

Karen, Science teacher

Others, however, were excited by the prospect of more data:

'I'm just thinking actually it's a great way to show progress…,'

Neil, Deputy Head

especially if the data is timely:

'I would like to see that. 'Cause I think there's quite a delay from when they actually do the exam and they get their results.'

Brian, Head of Science

Overall, it would seem that more flexibility would be welcomed, certainly at a year group and stream level, and, with more reservations, at a class level. This flexibility would allow the teachers to plan more effectively and reduce the stress placed on their pupils. More flexibility, however, cannot be achieved without certain

changes being made to the way in which examinations are currently delivered. To what extent are teachers and their pupils prepared to accommodate these changes?

7.5.8 Fairness

A very flexible system might involve different pupils taking different examinations in the same subject at the same time. The pupils were slightly unsettled by this, but didn't appear to hold strong views:

> *'if at the end of it, say, me and my friend did a different exam but she got like much higher marks than what I did I'd probably be like 'oh I had a harder exam', probably try and blame it on the exam,'*
>
> Ilona, year 10

> *'the grade boundaries always change anyway so it's not a huge deal,'*
>
> Sophie, year 10

> *'it's the same at A-levels when you took another module, and when you re-took. There are different questions there so no really.'*
>
> Grace, first year undergraduate

Different papers for different classes aroused more suspicion amongst the teachers, however:

> *'and it would be seen as "oh you got the easy paper that's why you got better marks". It would give us another thing for parents to hit us over the head with.'*
>
> Brian, Head of Science

The examiners welcomed the idea of different questions being used on different papers in the same session so some questions could be pre-tested, even if those questions did not contribute to the pupils' final marks:

> *'if you ask me for my ideal world I'd construct tests entirely from pre-tested items,'*
>
> Kevin, examiner

> *'if it's only one item per test and if it's the only way that we're gonna get pre-testing then I'd be happy.'*
>
> Kevin, examiner

They did note, however, that testing time placed limits on the amount of pre-testing that could be achieved in this way. Teacher views on pre-testing in live tests were mixed:

'it concerns me a little bit in the fact that, yeah you tell the children you're looking at a mark a minute, suppose they get stuck on one of the questions that gains no marks. And that will then have an adverse, a disproportionate adverse effect on their overall grade,'

Neil, Deputy Head

'I have to say conversely I wouldn't have a problem with that at all. If you were really talking about one or two questions very short ones that wouldn't get marked I wouldn't have a problem with that at all.'

Judith, Science teacher

One teacher felt that the kids would have to be kept in the dark about this, however:

'I wouldn't have a problem with that no. We helped you with the pre-testing last year. No we would be happy with that, as long as the kids weren't aware that it wasn't being marked, we'd keep that very quiet.'

Brian, Head of Science

It seems very unlikely that keeping it quiet would be more widely acceptable.

7.5.9 Security

More re-use of questions may be required in order to achieve more flexibility in the examination timetable. Examiners were sanguine about this:

> *'so they've got to learn the science to answer the questions – even if they've seen them before…'*
>
> Jacob, examiner

while the pupils professed to only remember questions that they couldn't answer correctly,

> *'the only ones that you do remember are the ones that you think you got wrong or that was really hard.'*
>
> Luke, year 10

Some felt that more re-use of questions would not be so different from the present situation,

> *'...I mean, things like RS, you've got every single question is almost identical every year, and you've got past papers and past papers and our teachers just says you know, 'this is a question that always comes up – learn it', so, I think people would say 'what did you have' 'what was the answer' all that but, if they don't actually get back the answers within the week then it's not*

really too much of a problem 'cause they might be telling them the wrong answer.'

Sophie, year 10

The examiners felt that trust in the teaching profession was key:

'well in the same way that when we had pre-tests within schools they were administered by someone from the board the teachers from the school were sort of allowed in the room they could cast an eye over a pre-test to see what direction things were moving,'

Kevin, examiner

'I think so ninety-nine percent of the teachers are professional… and it doesn't matter what system you operate if a teacher's determined to be a rogue… …he or she will be a rogue, and I think you've got to base any system on the fact that teachers are professionals … if you take away that assumption you might as well give up… seriously.'

Kevin, examiner

It may be, however, that as retired teachers, the examiners' view was of a teaching profession in an age before accountability and social networking, but one examiner who was also still a practising teacher agreed:

'the security of an exam paper will always be compromised if somebody wants to compromise it,'

Neil, Deputy Head

Other practising teachers were less complacent, however:

'you'd be putting the teacher in an awkward situation. Where the teacher would know they could go to another colleague, another school and get some advantage for their pupils who they want to do better. So we need less temptation almost rather than more temptation,'

Adam, Science teacher

Technology was seen to play a role in the heightened need for security:

'even if papers are not photocopied and posted to everybody, ideas still transfer around the country relatively easily…'

Karen, Science teacher

'... we all have networks, associates and friends in other schools,'

Karen, Science teacher

'Yeah you could just whack it on Facebook or something like that. You could get the questions pretty easily then.'

John, first year undergraduate

7.5.10 Practical Considerations

More flexible testing would require careful timetabling. Many comments related to systems that would be unworkable:

"sorry miss, I'm missing your lesson I've chosen to have my test then,"'

Sophie, year 10

'ah, a minute's notice no, it wouldn't work in school, 'cause you've planned to do this and guess what they've all gone to Colchester Zoo,'

Neil, Deputy Head

'people turning up on the wrong day for exams, people think they are going to enter for an exam when they are not,'

Neil, Deputy Head

'if we were taking out half a year group, only the triple award scientists and that would then leave half classes in French, Maths, English, we would have absolute mayhem.'

Brian, Head of Science

Teachers felt that the only way that the system could work was with examinations being taken in lessons, given good inter-subject co-ordination and careful forward planning:

> *'you could get round that by being able to take the exams in the classroom so you wouldn't move them all,'*
>
> Karen, Science teacher

> *'schools work better with dates in diaries,'*
>
> Neil, Deputy Head

> *'flexibility is great, but it needs to be planned flexibility,'*
>
> Neil, Deputy Head

> *'we'd have to know well in advance if they were going to be able to choose a date,'*
>
> Judith, Science teacher

> *'I would say that you wouldn't be able to schedule science exams independently of other departments.'*
>
> Judith, Science teacher

While these suggestions seem to offer an attractive solution, they do depend - assuming that more flexible assessment can only be delivered on-screen - on computers being available in classrooms.

7.6 Discussion

The findings reported here can be treated as fairly robust for the delivery of a more flexible examination structure for GCSE Science, for schools in the state sector. The sample size is small, but the level of redundancy in the data suggests that it is adequate. Views may obviously change, however, once a system is up and running. Care must also be exercised with any generalisations to other sectors of education, subjects or levels.

Overall, the consensus is that more flexibility would be welcomed in planning examination sessions for year groups, streams within year groups, and even classes within streams, provided that examinations can be taken in class time, and any technology required is available in that class time. Schools could exploit this flexibility to prepare pupils more fully for their examinations, test what is learned in a timely fashion, and reduce the stress of examinations for their pupils.

There is very limited support for any form of personalised approach which would allow pupils in the same class to progress at different rates. This is seen as a recipe for chaos which would be a nightmare to timetable and to teach. Rather than ensuring that no child is left behind it could end up isolating pupils and depriving them of the support they need. It is perceived that motivation, morale and results would all suffer. New methods of educating may be required before new methods of assessing are introduced.

The key concern of teachers, which is not necessarily shared by their pupils or by examiners, is the need to maintain a certain level of security for the examinations. This concern appears to be heightened by the potential for mass communication that technology offers. Security leaks will spread quickly, and

teachers will talk. Security concerns outweigh those of fairness. There appeared to be no serious concerns, for example, regarding the delivery of different tests to different pupils or classes, or even the pre-testing of questions in live tests, although it is preferable that these questions be marked. It appears, therefore, that the main challenge is designing a system that has enough overlap between test-takers and tests to satisfy a test equating design, but that also maintains an acceptable level of security.

In conclusion it would seem that there is enough evidence here to suggest that schools will not hold up their hands in horror at standards or ethical issues raised by children taking different versions of tests as part of test equating designs. Flexible delivery also appears to promise rewards, but the practical and technological issues require that a substantial investment is made in the infrastructure of on-screen testing to deliver robust test equating designs. The future gains obviously need to be balanced against the current cost of development.

8. Conclusion

8.1 Overview

The purpose of this thesis was to consider the contribution that Item Response Theory can make in the process of maintaining standards in England's qualification system. The current theoretical and practical positions were outlined alongside their challenges and limitations. The element of most concern arising from this review is the lack of an objective measure of performance standards within the English public qualifications system. The case made for the Rasch model and Item Response Theory being able to provide this objective measure of performance standards was then examined. Theoretically these models do appear to provide objective measures of performance. In practice, however, they are based on stringent assumptions that rarely hold. Detailed analyses of model choice and model fit revealed that great care must be taken in the design of IRT test equating experiments if these models are to prove useful. The most robust designs require the equating experiment to become part of live testing; this combination raises some difficult ethical issues. Exploration of these ethical issues through focus groups showed that they were not insurmountable, but that they would require careful handling. When the test equating data is not collected as part of live testing, the equating sample may not be representative of the live sample, so robustness studies are required. Finally, some more thought is given to the practical challenges facing any introduction of IRT test equating and the research programme that would be needed to support this introduction into the process of maintaining standards in England's qualification system.

8.2 Review of findings

Chapter 1 considered the current theoretical and practical positions that are adopted by awarding bodies in the English public examination system. It appears that all definitions of standards and their supporting practical implementations fail at some level to satisfy what is required of them. The public in England expect examinations to be representative of a particular level of performance standard; to be useful in rank ordering candidates for selection purposes; and to provide some indication of whether local and national standards are rising or falling. Judgemental approaches to maintaining standards, through which examiners carefully scrutinise the performance of candidates, appear to provide some guarantee of performance standard, but bias in the judgement introduced by the varying difficulty of test questions means that any pure judgemental approach would cause substantial swings in test outcomes from year to year. This fluctuation conflicts with the requirement to provide relatively stable rank orders of candidates within a reasonable time frame of, say, five years. It is not unreasonable to expect that a grade A this year in a particular qualification is equivalent to a grade A last year. Failure to provide this equivalence would undermine the use of grades by employers or universities. Statistical approaches to maintaining standards based on prior achievement can guarantee relatively stable rank ordering of candidates within a short time frame, however, they are based on the underlying assumption that performance does not improve or deteriorate over time. This invalidates their use as an indicator of whether local or national standards are rising or falling, and lays the system open to claims of the maintenance of a statistical illusion of performance standards.

As both of these positions represent compromises, there is clearly a need for investigations of alternative sources of evidence that could contribute to the maintenance of standards which meet both the requirement for an objective measure of performance standards while ensuring relatively stable outcomes over time. Chapter 2, therefore, considered the theoretical claims of the Rasch model and Item Response Theory against this requirement to provide an objective measure of performance standards. It also considered some of the practical difficulties involved with introducing the experimental designs required to enable Item Response Theory methods of test equating.

Chapter 2 suggested that the Rasch model has the strongest theoretical claim to an objective measure of performance standards. Unlike Item Response Theory models, the Rasch model makes no assumptions regarding ability distributions. This allows it to separate estimation of ability from the estimation of question difficulty. However, this objectivity is obtained through the observance of rigid assumptions. It requires, among other things, unidimensionality, no guessing, items with the same discrimination and items that perform consistently with respect to variables such as gender, age and education.

OPLM provides a more adaptable model as it allows items with different discrimination to be modelled. It is described as an extension to the Rasch model as it retains the summed score of candidates as a sufficient statistic for ability. The flexibility, however, is gained over the Rasch model through an increase in mathematical complexity. This complexity, which can make the model appear opaque, could be one reason why it has received little attention in the literature of Item Response Theory.

Mixed Rasch models and Item Response Theory models have been developed to deal with most of the restrictions of the Rasch model. As more parameters are added to Item Response Theory models, however, they become increasingly complex. This makes them difficult to estimate, their fit difficult to evaluate, and their interpretation less obvious. Clearly, the simplest model which provides reasonably accurate predictions is to be preferred.

Turning to test equating designs it is clear that theoretically they promise exactly what is required. Under the equity requirement candidates should be indifferent to whether they are tested on either of two alternate equated test forms. Translated into the requirements of the public examination system in England, this means candidates' grades would be equivalent whether they took last year's, this year's or next year's test. The equity requirement is, however, a theoretical principle that cannot be achieved in practice. Candidates would only be indifferent to which test form they had taken if the tests were perfectly reliable. Lacking perfect reliability there will always be some aspects of particular test forms that attract particular candidates. Estimating the extent to which this is the case and to which the equity requirement has been breached is extremely difficult.

While observance of equity may not be tested empirically, other aspects of IRT methods of test equating do lend themselves to quality control measures. It is possible, therefore, to evaluate the impact of breaching most of the assumptions of IRT models on the test equating. In order to equate tests, however, using IRT methods a test equating design must be in place. Chapter 2 also reviewed these test equating designs and considered their relative strengths and weaknesses.

In the US a variety of test equating designs are used for IRT and non-IRT equating. The emphasis there is to ensure that all equating is done on the basis of

item data that is collected if not in a live testing situation then as close as possible to live testing as can be achieved. This ensures that candidates are representative of the live testing sample and are sufficiently motivated. The desire to gather equating information in live testing, however, conflicts with security concerns that dictate that new items should receive as little exposure as possible before going live.

Both security and ethical concerns have led the Netherlands to take an alternative approach, which is described as post-equating. All the equating is done on the basis of item data collected from equivalent populations after all the items have been used live. This has the advantage of preserving the security of items and restricting experimentation in live tests, but the disadvantage that the equating population may not be representative of the live test population. A series of robustness studies was required to allay concerns that the equating results were not compromised by the use of a non-representative population.

It is clear, therefore, that for IRT methods of test equating to be practically useful in the context of public examinations in England, the tests must show reasonable fit to the models and a test equating design needs to be in place. Chapters 3 and 4 therefore dealt with the issue of fit while Chapters 5 to 7 dealt with the issues of different test equating designs.

The study of model fit is complicated by the existence of multiple tests of fit as well as two paradigms in which to apply these tests. The Rasch paradigm is not particularly interested in whether a model fits: fit is described as an idealisation that will never be achieved. The emphasis is placed on diagnosing sources of misfit and solving the misfit. The IRT paradigm uses statistical tests of fit which indicate whether more complex models are needed. If a model does not fit then a more complex model is estimated instead.

Standard chi-square Rasch tests of fit certainly proved effective in identifying some examples of poor test design: poor and confusing stimuli; poor distracters; mark schemes causing disordered thresholds; and items that failed to discriminate. Used in isolation, however, tests of fit at an item level were shown to be potentially misleading. The observed score distribution for tests that displayed no misfitting items according to Rasch chi-square tests was shown, by use of the Lord & Wingersky recursive algorithm, to be poorly estimated. Failure to reproduce observed score distributions is of particular concern in a test equating context as imputations are made from such distributions. Under the IRT paradigm, global fit statistics rejected the use of the Rasch model or OPLM for modelling of the tests analysed in almost every case. Graphical illustrations of the observed and expected score distributions, and Item Response Functions with empirical response functions, however, showed that the size of the misfit was often negligible.

Inspection of Item Response Functions clearly revealed that the observed performance at item level could be better modelled by OPLM than the Rasch model. Under OPLM the discrimination of items that failed to discriminate, perhaps through flaws in their design, could be set at a lower level than other items. In other words, the discrete classification of items into levels of discrimination produced better short term predictions of the empirical Item Response Functions.

The latent dimensionality of the item responses was examined by a comparison of the observed and expected values of the second eigenvalue of the tetrachoric correlations matrix under the Rasch model. The analysis revealed that these data loaded on a substantial second factor that was not predicted under the Rasch model. Comparisons with the popular Rasch Principal Components Analysis of variance are favourable; although it was suggested that for one test the Rasch

Principal Components Analysis failed to detect dimensionality due to poor fit of the Rasch model.

In Chapter 4 model fit was pursued in a Bayesian framework. This allowed more complex models to be estimated and evaluated against tests of model selection and tests of model fit that do not depend on asymptotic analysis. Looking forward to the test equating that would be attempted in Chapter 5, which would use items at the end of tests as the basis of test equating, a mixture Rasch model, the Multi-Class Mixture Rasch Model for test speededness, was estimated in order to evaluate how stable the item parameters at the end of a test appeared to be. The MMRM identifies candidates whose ability estimates decline towards the end of a test; the imputation is that they do not have enough time to complete the test. If enough candidates fail to complete the test in time, item parameters can be poorly estimated.

The MMRM suggested that there was a substantial proportion of candidates failing to complete certain tests. This proportion was sizeable enough to suggest that there would be a noticeable impact on the estimation of item parameters where those items were located towards the end of a test. A purified sample of those candidates whose ability estimates did not decline towards the end of the tests was obtained for use in the equating in Chapter 5.

Chapter 3 suggested that the observed score distribution was poorly predicted by the Rasch model for certain shorter tests. These tests all had a characteristic testlet design in which a stimulus was followed by four items related to the stimulus. It was hypothesised that these items may display weak local independence, and this may have caused the poor predictions under the Rasch model. The tests also used multiple-choice formats; so guessing may have been responsible for the poor predictions under the Rasch model. To test these hypotheses a two-parameter Testlet

Model and a three-parameter IRT model were estimated in a Bayesian framework and model fit compared against the Rasch model and the standard two-parameter IRT model.

Short term predictions did appear to improve slightly under the Two-parameter Testlet model, but even under this model the observed score distribution was poorly predicted. The three-parameter model produced degraded predictions. This cannot be taken as evidence, however, that guessing does not affect item responses as it may have been due to poor estimation of the pseudo-guessing parameter.

In these analyses, Posterior Predictive Model Checking (PPMC), which examines the features of the original dataset that are preserved by a model, proved very useful in diagnosing causes of misfit. The graphical displays associated with PPMC, which provide relevant probability distributions of the statistics of interest, proved more intuitive than asymptotic calculations which tend to be complex and therefore opaque. In particular, the PPMC checks replicated findings that the Rasch model does not preserve point biserial correlations of items. PPMC checks applied to Testlet Response Theory Models (TRTM) have not been reported elsewhere and are therefore of interest. Odds ratios, which measure the association between item pairs in terms of the observed and expected ratios of response patterns, were used to understand how successful TRTM models are in modelling weak local dependence between items in testlets. Where weak local independence existed in the observed data between items the TRTM was successful in predicting the weak local independence. However, the TRTM applies a single testlet parameter across all items in a testlet. This means that it poorly predicts the odds ratios between item pairs that do not display weak local independence within a testlet. This finding has interesting

diagnostic uses as it is suggested that PPMC checks on a TRTM model can inform test design where weak local dependence is intentionally sacrificed in order that a test can access higher skills of synthesis and interpretation. If test items remain conditionally independent it suggests that this single higher skill is no longer being tested.

Overall, the analyses of model fit in Chapters 3 and 4 suggested that modelling the discrimination of items discretely can improve the fit of IRT models and produce better short term predictions. This would suggest that OPLM would provide the most accurate equating predictions. Chapter 5 compared the predictions derived from OPLM against those from the Rasch model to evaluate the practical consequences of using a more complex model in the equating. It also considered the impact of the poor estimation of item parameters at the end of the test that was suggested by the analysis in Chapter 4.

The focus of Chapter 5 therefore was the impact of poor model fit on predictions derived from equating tiered GCSEs. As the tiers share common items they can be equated using a non-equivalent groups design. Use of a purified sample of candidates whose ability estimates did not decline towards the end of the test where the linking items were placed on one form in combination with the more flexible OPLM method of equating produced predictions that were substantially different from those derived from the Rasch model using a representative sample of candidates. It is clear, therefore, that while tiers can be equated successfully, care must be taken with choice of model and equating design.

From the analyses undertaken in Chapter 5 the following set of principles relating to test design and the equating design, supported by the general literature, were suggested for equating GCSE tiers:

i. That mark schemes and question presentation for linking items should be identical across test forms
ii. That linking items should be located early in the test forms
iii. That linking items should consist, where possible, of coherent blocks of items, but should remain representative of the subject as a whole
iv. That more than the minimum number of linking items should be included in case they perform differently across test forms (the rule of thumb is for 20 per cent of items to be common between forms to be equated)

Additional analyses suggested a variety of context effects that could degrade the quality of predictions derived from equating. These included differences in the maximum mark allocated to a linking question; different cognitive clues that were present in items preceding linking items; and ambiguity in questions that only affected higher ability candidates.

Chapter 6 turned to the wider issue of maintaining unit standards over time. A post-equating experiment was undertaken with the aim of understanding some of the potential difficulties in equating tests taken at different points in a course of study. From this experiment a set of principles relating to equating over time, again supported by the general literature, were suggested:

i. That linking items presented in coherent blocks are not removed from their context even if they appear not to be dependent upon it
ii. That redundant items should be built into an equating design so that they can be removed if they are suspected of performing differently due to context effects, population dependence or school effects

iii. That a matrix sampling approach is adopted which allows more items to be linked and school effects to be monitored

iv. That the narrowest construct definition possible should be used

The experiment also replicated the finding that there can be interactions between recency and duration of instruction and item difficulty.

Finally, Chapter 7 considered opportunities that item banks derived from a test equating methodology opened up in assessment. Firstly on demand testing was discussed along with the possibilities of high-stakes tests being used in formative ways. While pure anytime anywhere on-demand testing does not seem suitable for high-stakes achievement tests there is clearly potential for more flexibility to be introduced into the public examination system in England. The case for adaptive testing and multi-stage testing was also examined. It was suggested that multi-stage testing has both better security and better opportunities for control of test design than adaptive testing and represents a feasible alternative to tiering for the future. The security issues involved in adaptive testing seem hard to overcome.

Sustaining item banks, however, is only possible through pre-testing and test equating in live test sessions. Some of the ethical issues that arise were considered through a series of focus groups. While the opportunities offered by item banks of more flexible testing were broadly welcomed, security and ethical concerns were an issue.

8.3 Summary

8.3.1 The challenge ahead

Clearly, neither judgemental nor statistical approaches to the maintenance of standards satisfy the requirements of the public. The combination of judgemental and statistical indicators is hard to defend theoretically and is a recipe for confusion and obfuscation. In judging whether the standards of different awarding bodies are in line different weight can be placed on different sources of evidence. An awarding body can therefore defend its standards by using a different definition of standards to the others. Recourse to performance descriptors, for example, could be used to defend increases in outcomes that are statistically improbable. Without a clear framework within which the evidence is accorded a certain weight the maintenance of standards between awarding bodies and over time becomes opaque.

For many years awarding bodies have defended their procedures through a sociological perspective that regards grades as illocutionary speech acts; the equivalent of a football player being offside not because of where they were standing in relation to the play but because the referee had declared them offside. Just as in the world of football, television replays are making this position untenable, neither awarding bodies nor the examiners they employ can any longer lay claim to unassailable authority. Examination statistics are being published in greater detail than ever before, and there are plans to publish examplar material from each examination session. In this climate, a robust framework for maintaining standards is critical.

If all awarding bodies were to adopt an IRT approach, which has been shown by this thesis to be feasible, then the evidence for the maintenance of performance

standards could be open to objective external scrutiny in a way that the qualitative records simply do not lend themselves to. After establishing equivalence, through an experimental or statistical link, awarding bodies could maintain performance standards using IRT procedures. Performance could then be monitored through objective measurements of equating error and the statistical likelihood of increases and decreases in performance calculated.

Of course a simpler alternative is to adopt a diktat model in which only one of the purposes of examinations would be accepted. If qualification to a particular standard was considered to be of paramount importance then the fluctuations in outcomes and differences between awarding bodies associated with judgemental approaches would be a necessary but unfortunate side-effect. If, on the contrary, rank ordering was considered to be of paramount importance then a statistical approach would suffice. Standards between awarding bodies would be easily monitored through readily available statistics. Genuine differences, however, in cohort ability from one year to the next would be suppressed so the system would no longer support evaluations of local or national progress. There is no reason, however, why this evaluation should be undertaken by awarding bodies; indeed their involvement in the system would seem a conflict of interest.

8.3.2 Concluding comments

The recent US-focussed Handbook of Statistics on Psychometrics suggests that a profitable area for future research in psychometrics in the US is the investigation of approaches to the maintenance of standards over time that are based on the use of prior achievement data, the methodology that is prevalent in the UK. It may seem somewhat ironic therefore that this thesis has started from the limitations of UK-

based prior achievement approaches to the maintenance of standards over time and suggested more research into the kinds of test equating that is prevalent in the US..

Test equating designs are complex and require investment in infrastructure, cooperation from examinees and researchers with advanced psychometrics skills. When prior achievement data is readily available predictions for every test within every specification within every awarding body can be produced by a single researcher with a degree of mathematical competence within a matter of weeks. This same researcher, given two more weeks, can also produce objective measures of inter-awarding body comparability for every test within every specification. If no cost is incurred in the collection of the prior achievement data then it would seem perverse to pursue models that offer a theoretical advantage that may be compromised by the practical details of its implementation. Working within this framework rank ordering of candidates over successive years in a specification can be preserved to a remarkable degree of accuracy. The other claims laid on the assessment system can be satisfied by design, in the case of the validity of the performance standard required, and by experimentation, in the case of relative local and national trends in performance. In this framework test equating can be used to maintain standards across tiers of qualifications and IRT more generally can be used in validity studies.

England, however, is relatively unique in its insistence on externally assessing every pupil in every school at ages 10 and 13. This provides the public examination system with an extraordinarily rich and robust dataset on which to base its prior achievement approaches. Sadly for the awarding bodies, however, the system of national external testing at ages 10 and 13 is being dismantled. The subsidy on the prior achievement approach is effectively being removed. It is

conceivable that within 10 years no prior achievement data will be freely available to awarding bodies. Unless alternative approaches are available a return to norm referencing and the vagaries of judgement could be the only recourse. It is unlikely that a public that has grown used to the stability promoted by the current system would accept the vacillations in outcomes that were commonplace before prior achievement data was readily available. For this reason, therefore, it is suggested that the public examination system in England should continue research in test equating that could provide evidence on the maintenance of standards over time.

8.4 Suggestions for further research

8.4.1 The reliability of marking of constructed response items

The scope of test equating is obviously limited to constructed response items. The advent of on-screen marking in which individual scripts are distributed amongst a pool of markers suggests that the marking can be considered to be randomly equivalent. This suggestion, however, needs validation so a sensible limit can be placed on the maximum marks of questions considered for equating.

8.4.2 Predictions derived from constructed response items

As constructed response items represent only one mode of assessment in the English public examination system predictions derived from equating portions of assessments need to be assessed for their validity. Improvement in the performance of candidates on constructed response items may not necessarily be accompanied by improvements in the performance of candidates on essay questions, for example.

8.4.3 The robustness of post-equating samples

The main obstacle to introducing post-equating approaches is in the generalisability of findings derived from a post-equating sample. Investigations into whether there are pools of available students who are prepared for tests but not actually taking those tests need to be undertaken. The robustness of data gathered in a post-equating framework must also be established.

8.4.4 The ethics of pre-testing and test equating in live test sessions

The most robust data and the cheapest data for test equating is obtained when the test equating takes place in live test sessions. Experimentation within live testing, however, raises some difficult ethical questions. The potential impact on candidates' scores of using unmarked equating sections within tests should be evaluated. The quantitative data should be supplemented by qualitative data relating to the experience of candidates taking such equating sections and teachers preparing candidates for tests that contain these equating sections.

8.4.5 IRT and validity

Although an examination of the validity of public examinations was not the primary purpose of this thesis, it is replete with validity evidence. Contrary to fears that the Rasch model in particular restricts validity through its rigid assumptions this thesis has shown how fitting a variety of Rasch and IRT models can shed light on interesting test design issues. It is hoped that regardless of the success of any equating programme IRT will be used increasingly to establish the validity of tests on an objective basis.

8.4.6 Performance standards over time

Regardless of the relative advantages and disadvantages of embedding IRT procedures in the operational procedures of setting grade boundaries, experimental research on performance standards over time is of national interest. Any study would need to take particular care with sampling of both questions and candidates.

8.4.7 Multi-Stage Testing

It is suggested that Multi-Stage Testing could be implemented to replace the current system of tiering. A study of classification accuracy under the tiering and the multi-stage models could add weight to this argument.

References

Agresti, A. (2002). *Categorical data analysis*. Hoboken, New Jersey: John Wiley & Sons.

Akaike, H. (1973). 2nd International Symposium on Information Theory (pp. 267–281). Budapest: Akad´emiai Kiad´o.

Albert, J. & Ghosh, M. (2000). Item response modeling. In D. K. Dey, S. K. Ghosh & B. K. Mallick (Eds.), *Generalized linear models: A Bayesian perspective* (pp. 173-193). New York: Marcel Dekker.

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.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 143-166). Maple Grove, MN: JAM Press.

Assessment and Qualifications Alliance. (2008). *Mathematics 4301 Specification A 2008*. Manchester: Assessment and Qualifications Alliance. Retrieved from http://store.aqa.org.uk/qual/pdf/AQA-4301-W-SP-08.PDF

Baird, J. (2000). Are examination standards all in the head? Experiments with examiners' judgements of standards in A-level examinations. *Research in Education*, *64*, 91-100.

Baird, J. (2007). Alternative conceptions of comparability. In P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 124-165). London: Qualifications and Curriculum Authority.

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

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

Baird, J., Fearnley, A., Fowles, D., Jones, B., Morfidi, E. & White, D. (2001). *Tiering in the GCSE: A study undertaken by AQA on behalf of the Joint Council for General Qualifications*. Joint Council for General Qualifications.

Beaton, A. E. & Zwick, R. (1990). *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 Reading Anomaly* (Revised). National Assessment of Educational Progress, Educational Testing Service, Princeton, NJ.

Béguin, A. A. (2000). *Robustness of Equating High-Stakes Tests*. (Master's thesis) University of Twente, Enschede, Netherlands. Retrieved from http://cito.nl/share/poc/dissertaties/dissertationbeguin2000.pdf

Béguin, A. A. & Glas, C. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541-561.

Béguin, A. A., Wheadon, C., Meadows, M. & Eggen, T. (2007, November). *Comparability of high-stakes assessments: the role of standard setting*. Paper presented at the 8th annual conference of the Association for Educational Assessment (AEA) Europe, Stockholm.

Binks, J. (2002). *Official Response to the Science and Technology Parliamentary Committee Inquiry: Science Education from 14-19*. London: Confederation of British Industry.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Black, B. & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, *23*(3), 357-373.

Black, P. (2007, May). *Can we design a supportive assessment system?* Paper presented at the Chartered Institute of Educational Assessors, London.

Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Maidenhead, UK: Open University Press.

Black, P. & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*(2), 139-148.

Bock, R. D. & Moustaki, I. (2007). Item response theory in a general framework. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26* (pp. 469-513). Amsterdam: Elsevier.

Bolt, D. M., Cohen, A. S. & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, *26*(4), 381-409.

Bolt, D. M., Cohen, A. S. & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, *39*(4), 331-348.

Brennan, R. L. (2008). A discussion of population invariance. *Applied Psychological Measurement*, *32*(1), 102-114.

Brown, M. (1989). Graded assessment and learning hierarchies in mathematics: An alternative view. *British Educational Research Journal*, *15*(2), 121-128.

Cameron, J. (2001). Negative Effects of Reward on Intrinsic Motivation - A Limited Phenomenon: Comment on Deci, Koestner, and Ryan (2001). *Review of Educational Research* 71(1): 29-42.

Charmaz. K, (2006). *Constructing grounded theory*. London: Sage.

Chen, W. & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

Christie, T. & Forrest, G.M. (1980) Standards at GCE A-level : 1963 and 1973 : a pilot investigation of examination standards in three subjects. Basingstoke: Macmillan Education.

Cockcroft, W. (1982). *The Cockcroft Report (1982): Mathematics counts*. London: Her Majesty's Stationery Office. Retrieved from http://www.educationengland.org.uk/documents/cockcroft/

Coe, R. (2007). Common Examinee Methods. In P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 331-367). London: Qualifications and Curriculum Authority.

Cohen, J. (1988) *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.

Cook, L. L. & Paterson, N. (1987). Problems related to the use of conventional and Item Response Theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, *11*(3), 225-244.

Cresswell, M. J. (1997). *Examining judgements: Theory and practice of awarding public examination grades (Doctoral dissertation)*. London: Institute of Education, University of London.

Cresswell, M. J. (2000). The role of public examinations in defining and monitoring standards. In *Educational Standards* (pp. 69-104). Oxford: Oxford University Press for the British Academy.

Cresswell, M. J. (2010). *Monitoring general qualification standards: A strategic view from AQA*. Manchester: Assessment and Qualifications Alliance.

de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement*, *33*(6), 465-485.

Deci, E. L. (1975). *Intrinsic motivation*. New York: Plenum Press.

Deci, E. L., Ryan, R. M., Koestner, R. (2001). The Pervasive Negative Effects of Rewards on Intrinsic Motivation: Response to Cameron (2001). *Review of Educational Research*, 71(1): 43-51.

Department for Education and Skills. (2006). *Making Good Progress: How can we help every pupil to make good progress at school?* Nottingham: DfES Publications

Dorans, N. J. (1990). Equating methods and sampling designs. *Applied Measurement in Education*, *3*(1), 3.

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., Pommerich, M. & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.

Drasgow, F. & Lissak, R. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, *68*, 363-373.

Eason, S. (2003). *Cashing-in of curriculum 2000 AS and A-level results*. Manchester: Assessment and Qualifications Alliance.

Eason, S. (2007). *GCE Information and Communication Technology (5521 / 6521): Conflict of unit standards between the January and June examinations series*. Manchester: Assessment and Qualifications Alliance.

Eason, S. (2008). *Perceived conflict between GCE unit awarding outcomes from the January and June examinations series: A worked example based on AS Psychology B (5186)*. Manchester: Assessment and Qualifications Alliance.

Eason, S. (2009). *GCSE Sciences: Candidates' unit-entry behaviour and the impact on overall subject awards – June 2008 and June 2009*. Manchester: Assessment and Qualifications Alliance.

Eason, S. (2010). *Predicting GCSE outcomes based on candidates' prior achieved Key Stage 2 results*. Manchester: Assessment and Qualifications Alliance.

Ecclestone, K. (2006). *Assessment in post-14 education: The implications of principles, practices and politics for learning and achievement* (No. 2). The Nuffield Review of 14-19 Education. The Nuffield Foundation. Retrieved from http://www.nuffield14-19review.org.uk/files/documents125-1.pdf

Edexcel. *Mathematics (2381) Modular*. Retrieved August 3, 2009, from http://www.edexcel.com/quals/gcse/gcse-leg/maths/2381/Pages/default.aspx

Educational Testing Service. (2009). GRE Details: Test Takers. *Educational Testing Service*. Retrieved from http://www.webcitation.org/5jIDsSaIr

Eignor, D. R., Stocking, M. L. & Cook, L. L. (1990). Simulation results of effects on linear and curvilinear observed- and true-score equating procedures of matching on a fallible criterion. *Applied Measurement in Education*, *3*(1), 37-52.

Engineering Council (2000). *Measuring the Mathematics Problem*. London: The Engineering Council.

Fawcett, J. (2005). Criteria for evaluation of theory. *Nursing Science Quarterly*, *18*, 131-135.

Feyerabend, P. (1988). *Against Method* (Rev. ed.). Verso: London/New York.

Fischer, G. H. (2007). Rasch models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26* (pp. 515-585). Amsterdam: Elsevier.

Fowles, D. (2009). *A concurrent approach to estimating the reliability of electronic marking of long form answers*. Manchester: Assessment and Qualifications Alliance.

Fox, J. & Wyrick, C. (2008). A mixed effects randomized item response model. *Journal of Educational and Behavioral Statistics*, *33*(4), 389-415.

Frantz, D. & Nordheimer, J. (1997, September 28). Giant of exam business keeps quiet on cheating. *New York Times*. Retrieved from http://www.nytimes.com (http://www.webcitation.org/5e78yqhcW)

Gilbert, C. (2006). *2020 Vision: Report of the teaching and learning in 2020 review group*. Department for Education and Skills. Nottingham: DfES Publications

Glas, C. & Falcon, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87-106.

Good, F. & Cresswell, M. J. (1987). *Grade awarding judgements in differentiated examinations*. Manchester: Assessment and Qualifications Alliance.

Good, F. & Cresswell, M. J. (1988a). *Differentiated assessment: Grading and related issues*. London: The Secondary Examinations Council.

Good, F. & Cresswell, M. J. (1988b). *Grading the GCSE*. London: Secondary Examinations Council.

Green, B. F. J. (1983). Notes on the efficacy of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of Modern Psychological Measurement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society: Series B*, 29(1), 83-100.

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

Hanson, B. A. & Béguin, A. A. (1999). *Separate versus concurrent estimation of IRT item parameters in the common item equating design*. ACT Research Report Series, PO Box 168, Iowa City, IA 52243-0168. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED438310

Harker, R. & Tymms, P. 2004. The Effects of Student Composition on School Outcomes. *School Effectiveness and School Improvement,* 15(2): 177-199.

Hitchcock, C. & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *The British Journal for the Philosophy of Science*, *55*(1), 1-34.

Holland, P. W., Dorans, N. J. & Petersen, N. (2007). Equating test scores. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26* (pp. 169-203). Amsterdam: Elsevier.

Ireson, J., Hallam, S. & Hurley, C. (2005). What are the effects of ability grouping on GCSE attainment? *British Educational Research Journal*, *31*(4), 443-458.

Jones, B. (2002). *Clerical errors in marking - Manchester office - year 2001 summer examinations*. Manchester: Assessment and Qualifications Alliance.

Jones, B. (2005). *Analysis of predicted outcomes for six GCE science units in the January and June 2004 examination series*. Manchester: Assessment and Qualifications Alliance.

Jones, B. (2008). *Statistical predictions for GCE new specification AS units in January 2009: A discussion paper*. Manchester: Assessment and Qualifications Alliance.

Jones, B. (2009a). *Awarding GCSE and GCE - time to reform the Code of Practice?* Manchester: Assessment and Qualifications Alliance.

Jones, B. (2009b). *Setting standards in the new GCE specification AS and A2 units in January 2010*. Manchester: Assessment and Qualifications Alliance.

Keifer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, *27*, 887-903.

Kim, J. & Bolt, D. M. (2007). An NCME instructional module on estimating Item Response Theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*(4), 38-51.

Kolen, M. J. (1990). Does matching in equating work: A Discussion. *Applied Measurement in Education*, *3*(1), 97-104.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and Practices* (2nd ed.). New York: Springer.

Laming, D. (2004). *Human judgment*. Cengage Learning EMEA.

Lawrence, I. M. & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. *Applied Measurement in Education*, *3*(1), 19-36.

Linacre, J. M. (1994). Sample size and item calibration (or Person Measure) stability. *Rasch Measurement Transactions*, *7*(4), 328.

Linacre, J. M. (2004a). Equating constants with mixed item types. *Rasch Measurement Transactions*, 18(3), 992.

Linacre, J. M. (2004b). Rasch model estimation: Further topics. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement*. Maple Grove, Minnesota: JAM Press.

Linacre, J. M. (2008). A user's guide to WINSTEPS® MINISTEP: Rasch-Model Computer Programs (Program Manual 3.66.0.)

Liu, J., Harris, D. & Schmidt, A. E. (2007). Statistical procedures used in college admissions testing. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26* (pp. 1057-1091). Amsterdam: Elsevier.

Livingston, S. A. (2004). *Equating test scores (without IRT)* (ETS Rep. No. LIVINGSTON). Princeton, NJ: Educational Testing Service.

Livingston, S. A., Dorans, N. J. & Wright, N. K. (1990). What combination of sampling and equating methods works best? *Applied Measurement in Education*, *3*(1), 73-95.

Lord, F. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F. & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. & Wingersky, M. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*, 452-461.

Luecht, R., Brumfield, T. & Breithaupt, K. (2006). A testlet assembly design for adaptive multi-stage tests. *Applied Measurement in Education*, *19*(3), 189-202.

Lundgren-Nilsson, Å., Tennant, A., Grimby, G. & Sunnerhagen, K. (2006). Cross-diagnostic validity in a generic instrument: An example from the functional

independence measure in Scandinavia. *Health and Quality of Life Outcomes*, *4*(55).

Mair, P. & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1-20.

Mariano, L. T. & Junker, B. W. (2007). Covariates of the rating process in hierarchical models for multiple ratings of test items. *Journal of Educational and Behavioral Statistics*, *32*(3), 287-314.

Maris, G. & Bechger, T. (2007). Scoring open ended questions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26* (pp. 663-681). Amsterdam: Elsevier.

McLeod, L. D. & Schnipke, D. L. (1999, April). *Detecting items that have been memorized in the computerized adaptive testing environment*. Presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Mead, A. (2006). An introduction to multi-stage testing. *Applied Measurement in Education*, *19*(3), 185-187.

Meyer, L. (2009a). *Principles of standard setting*. Manchester: Assessment and Qualifications Alliance.

Meyer, L. (2009b). *Putting education policy into practice*. Manchester: Assessment and Qualifications Alliance.

Molenaar, I. W. (1983). Some improved diagnostics for failure in the Rasch model. *Psychometrika*, *48*, 49-72.

Moreno, K. & Segall, D. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters & J. R. McBride (Eds.), *Computerized adaptive*

testing: From inquiry to operation (pp. 169-174). Washington, DC: American Psychological Association.

Mroch, A. A., Bolt, D. M. & Wollack, J. A. (2005). *A new Multi-Class Mixture Rasch Model for test speededness*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

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

Newton, P. (2005b). The public understanding of measurement inaccuracy. *British Educational Research Journal*, *31*(4), 419-442.

Newton, P. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy and Practice*, *14*, 149-170.

Newton, P. (2008). Comparability monitoring: Progress report. In P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 452-476). London: Qualifications and Curriculum Authority.

Nietzsche, F. W. (trans. 2004). *Human, all too human*. Cambridge: Cambridge University Press.

Noss, R., Goldstein, H. & Hoyles, C. (1989). Graded assessment and learning hierarchies in mathematics. *British Educational Research Journal*, *15*(2), 109-120.

Patz, R. J. & Junker, B. W. (1999). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*(4), 342-366.

Petersen, N. (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?* Manchester: Assessment and Qualifications Alliance.

Pinot de Moira, A. (2009a). *The effects of maturity: Evidence from linear GCSE specifications*. Manchester: Assessment and Qualifications Alliance.

Pinot de Moira, A. (2009b). *Introduction of the new AS and A-level qualifications: Predictions for the winter 2009 awards*. Manchester: Assessment and Qualifications Alliance.

Pinot de Moira, A. (2009c). *Marking reliability & mark tolerances: Deriving business rules for the CMI+ marking of long form answer questions*. Manchester: Assessment and Qualifications Alliance.

Poirier, D. J. (1988). Causal relationships and replicability. *Journal of Econometrics*, *39*, 213-324.

Pollitt, A. (1985). *What makes exam questions difficult?: An analysis of 'O' grade questions and answers*. Edinburgh: Scottish Academic Press.

Pollitt, A., Ahmed, A. & Crisp, V. (2007). The demands of examination syllabuses and question papers. In P. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 166-210) London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (2009). *Code of Practice*. London: Author.

R Development Core Team. (2010). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

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

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *The Danish Yearbook of Philosophy*, *14*, 58-94.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, *9*(4), 401-412.

Rizopoulos, D. (2006). An R package for latent variable modelling and Item Response Theory analyses. *Journal of Statistical Software*, *17*(5), 1-25.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, *12*(4), 1151-1172.

Ryan, M. (2010, March 24). Tories want traditional A- level to 'restore confidence'. *BBC*. Retrieved from http://news.bbc.co.uk/1/hi/education/8583440.stm

Scharaschkin, A. & Baird, J. (2000). The effects of consistency of performance on A-level examiners' judgements of standards. *British Educational Research Journal*, *26*, 343-357.

Schmeiser, C. (2004). *Reaffirming our raison d'etre: The ACT assessment*. Paper presented at the annual meeting of the American Psychological Association, Honolulu.

Schmitt, A. P., Cook, L. L., Dorans, N. J. & Eignor, D. R. (1990). Sensitivity of equating results to different sampling strategies. *Applied Measurement in Education*, *3*(1), 53.

Sinharay, S. (2005). Assessing fit of unidimensional Item Response Theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*(4), 375-394.

Sinharay, S., Johnson, M. S. & Stern, H. S. (2006). Posterior predictive assessment of Item Response Theory models. *Applied Psychological Measurement*, *30*(4), 298-321.

Skaggs, G. (1990). To match or not to match samples on ability for equating: A discussion of five articles. *Applied Measurement in Education*, *3*(1), 105-113.

Smith, R.M. *Fit Analysis in Latent Trait Measurement Models*. In E.V. Smith & R.M. Smith (Eds.), *Introduction to Rasch Measurement* (pp. 73-92). Maple Grove, MN: JAM Press.

Smith, R.M., Schumacker, R.E. & Bush, M.J. (2000). *Examining Replication Effects in Rasch Fit Statistics*. In M. Wilson, G. Engelhard (Eds.), *Objective Measurement: Theory into Practice* (pp. 303-318). Stamford: Ablex.

Spalding, V. (2009). *GCSE Science A: The size and effect of 'If at first you don't succeed, try, try, again'*. Manchester: Assessment and Qualifications Alliance.

Spiegelhalter, D., Best, N., Carlin, B. & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*: *Series B*, *64*(4), 583-639.

Spiegelhalter, D., Thomas, A., Best, N. & Lunn, D. (2003). WinBUGS User Manual (Version 1.4) [Computer manual]. Cambridge: MRC Biostatistics Unit, Institute of Public Health. Retrieved from http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/manual14.pdf

Stewart, D. & Shamdasani, P. (1990). *Focus groups: Theory and practice*. Newbury Park, CA: Sage.

Stringer, N. (2008, September). *An appropriate role for professional judgement in maintaining standards in English general qualifications*. Paper presented at the

34th annual conference of the International Association for Educational Assessment (IAEA), Cambridge, UK.

Stringer, N. (2010). *Setting and maintaining GCSE and GCE grading standards: the case for contextualised cohort-referencing*. Manchester: Assessment and Qualifications Alliance.

Sturtz, S., Ligges, U. & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, *12*(3), 1-16.

Swaminathan, H. & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. Journal of Educational and Behavioral Statistics, 7(3), 175-191.

Swaminathan, H., Hambleton, R. K. & Rogers, H. J. (2007). Assessing the fit of Item Response Theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26* (683-718). Amsterdam: Elsevier.

Sykes, R. (2010). *The Sir Richard Sykes Review*. Retrieved from http://www.conservatives.com/News/News_stories/2010/03/~/media/Files/Downloadable%20Files/Sir%20Richard%20Sykes_Review.ashx

Tennant, A. & Pallant, J. F. (2006). Unidimensionality matters! (A Tale of Two Smiths?). *Rasch Measurement Transactions*, *20*(1), 1048-51.

Traub, R. (1983). A priori considerations in choosing an item response model. In *Applications of Item Response Theory*. Vancouver: Educational Research Institute of British Columbia.

Tymms, P. & Fitz-Gibbon, C. (2001). Standards, achievement and educational performance: A cause for celebration? In J. Furlong & R. Phillips (Eds.), *Education, reform and the state: Twenty-five years of politics, policy and practice* (pp. 156-173). London: RoutledgeFalmer.

van Rijn, P., Verstralen, H. & Béguin, A. A. (2009). *Classification accuracy of multiple-test based decisions using Item Response Theory*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Verhelst, N. & Glas, C. (1995). The One Parameter Logistic Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Wainer, H. (with Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., et al.). (2000). *Computerized Adaptive Testing: A Primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H., Bradlow, E. T. & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.

Wheadon, C. & Béguin, A. A. (2010). Fears for tiers: Are candidates being appropriately rewarded for their performance in tiered examinations? *Assessment in Education, 17(3)*, 287-300.

Wheadon, C., Spalding, V. & Tremain, K. (2008). *GCSE English A: Comparability between tiers*. Manchester: Assessment and Qualifications Alliance.

Whitehouse, C. & Eason, S. (2007). *Pseudo-aggregation for GCSE Science A (4461)*. Manchester: Assessment and Qualifications Alliance.

Wickham, H. (2009). ggplot2: elegant graphics for data analysis. New York: Springer.

Wise, S. L., Plake, B. S. & Mitchell, J. V., Jr. (1990). Editor's Note. *Applied Measurement in Education*, *3*(1), 1-2.

Wollack, J. A., Youngsuk, S. & Bolt, D. M. (2007). *Using the testlet model to mitigate test speededness effects*. Presented at the annual meeting of the National Council on Measurement in Education, Chicago.

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

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Wright, B. D. & Stone, M. H. (1999). *Measurement Essentials* (2nd ed.). Wilmington, DE: Wide Range.

Yamamoto, K. & Everson, H. (1997). Modeling the effects of test length and test time on parameter estimation using the HYBRID model. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 89-98). New York: Waxmann.

Yi, Q., Harcourt Assessment, Harris, D. & Gao, X. (2008). Invariance of equating functions across different subgroups of examinees taking a science achievement test. *Applied Psychological Measurement*, *32*(1), 62-80.

Zeng, L. & Kolen, M. J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement*, *19*, 231-240.

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice*, *10*(3), 10-16.

Appendix A: An Example of a 'How Science Works' Question

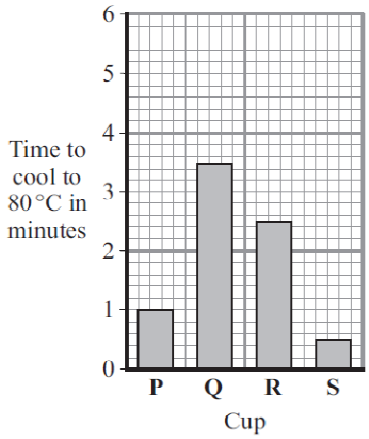
QUESTION EIGHT

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		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	HSW	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		15	10
Copper and recycling	Chemistry	HSW	15	9
Whooping cough	Biology		11	10
Solar cell panels	Physics	HSW	10	11
Smoking and disease	Biology		9	16
Vitamin C	Biology	HSW	7	11
Insulation	Physics		7	11
Drug trials	Biology	HSW	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

Appendix C: Description of Research Project for Participants

www.aqa.org.uk

ITEM BANK/ON-DEMAND TESTING FOCUS GROUP

DESCRIPTION OF RESEARCH PROJECT FOR PARTICIPANTS

This project explores some of the issues in the UK context of using item banking/on-demand testing. The rapid progress in ICT has led to the development of computer-based item banks. Computer-based item banks can be used to create on-demand electronic tests. On-demand tests provide the flexibility that is suited to the personalisation agenda being pursued by the Department for Children, Schools and Families, and is likely to become increasingly apparent in the education system. It is worth investigating, therefore, the requirements and concerns of stakeholders in this field.

The main aims and objectives of this research project are:

1. To ascertain the desired functions from stakeholders of an item bank/on-demand test.

2. To ascertain the concerns of stakeholders using item bank/on-demand tests.

3. To establish safeguards required for item bank/on-demand tests.

AQA needs help from stakeholders to achieve these objectives.

Evaluation

Stakeholders will be engaged in a two hour focus group, held in London. Three focus groups will be conducted comprised of representatives of; those who will be taking on-demand tests, those who will be end-users of on-demand tests, and those who will provide on-demand tests. The purpose of the group will be a structured discussion to ascertain experience, requirements and concerns in this field.

Appendix D: Parental Consent Form

ASSESSMENT and QUALIFICATIONS ALLIANCE

ON-DEMAND TESTING
STUDENT FOCUS GROUP CONSENT FORM

Please note the focus group will be recorded by audiotape. The data collected from the focus group will be analysed and used to contribute to a report on students and teachers perspectives on On-Demand testing. Once the data entry is complete for this research project, the entries will be made anonymous so that individual teachers' and pupils' identities are protected and opinions and attitudes are unattributable.

The data will be handled in accordance with the Data Protection Act (1998) and the AQA's current registration particulars under that legislation. The data will be used only for the purpose for which they are being collected.

	Yes	**No**
1. I am willing for my child to take part in the student focus group	☐	☐
	Yes	**No**
2. Is it acceptable to you that the focus group be audiotaped?	☐	☐

Signature: ..... **Date:** .....

Print name: .....

Position: .....

Please sign and print your name and give this form to your child so that they can return this form to Victoria Spalding on **12th November 08**.

Thank you

Appendix E: Student Consent Form

ASSESSMENT and QUALIFICATIONS ALLIANCE

ON-DEMAND TESTING
STUDENT FOCUS GROUP CONSENT FORM

Please note the focus group will be recorded by audiotape. The data collected from the focus group will be analysed and used to contribute to a report on students' and teachers' perspectives on On-Demand testing. Once the data entry is complete for this research project, the entries will be made anonymous so that individual teachers' and pupils' identities are protected and opinions and attitudes are unattributable.

The data will be handled in accordance with the Data Protection Act (1998) and the AQA's current registration particulars under that legislation. The data will be used only for the purpose for which they are being collected.

	Yes	**No**
1. I am willing to take part in the student focus group	☐	☐

	Yes	**No**
2. Is it acceptable to you that the focus group be audiotaped?	☐	☐

Signature: ..... **Date:** .....

Print name: .....

Please sign and print your name and return this form to Victoria Spalding on **12th November 08**.

Thank you

Appendix F: Teacher Consent Form

ASSESSMENT and QUALIFICATIONS ALLIANCE

ON-DEMAND TESTING
TEACHER FOCUS GROUP CONSENT FORM

Please note the focus group will be recorded by audiotape. The data collected from the focus group will be analysed and used to contribute to a report on students and teachers perspectives on On-Demand testing. Once the data entry is complete for this research project, the entries will be made anonymous so that individual teachers' and pupils' identities are protected and opinions and attitudes are unattributable.

The data will be handled in accordance with the Data Protection Act (1998) and the AQA's current registration particulars under that legislation. The data will be used only for the purpose for which they are being collected.

	Yes	**No**
1. I am willing to take part in the teacher focus group	☐	☐
	Yes	**No**
2. Is it acceptable to you that the focus group be audiotaped?	☐	☐

Signature: ..... **Date:** .....

Print name: .....

Position: .....

Please sign and print your name and return this form to Victoria Spalding on **12th November 08**.

Thank you

Appendix G: Student Focus Group Session Plan

Stage 1 and 2a Topical areas for Group Discussion

Introduction

- State who we are
- State purpose of study – for Ofqual report
- Briefly describe on-demand testing
- State ground rules of discussion – everyone talk, interested in all opinions, don't talk over anyone
- Get everyone to say their name (makes transcription easier)

Stimulus material

Views on on-demand testing

- What do you think is good about on-demand testing?
- What do you think is bad about on-demand testing?

General Topical areas for more detailed discussion

Test Pressure

- Frequent vs. end of year?
- Re-sits?
- Who chooses when student is ready?
- Parents will force children to take examinations
- Students will be over pressurised by the new regime

Re-Sits

- How many?
- Should there be a limit?

Revision

- Do you prefer to revise in groups?
- What if only a few of you were taking the test?
- What if there were no past papers?
- If you could would you help each other in the test?

Specific Topical areas for more detailed discussion

End of year vs. on-demand

- Is it easier to remember a topic just after you have learnt it?
- Do you think you will have acquired all the skills you need if you take the test earlier?
- Are there any subjects that you don't think this will work for?
- What would you prefer?
- When would you choose?

Different papers for different people

- How would you feel about taking a unique test?
- Are you worried that the test your friend takes may be easier than the one you take?

Location

- Would you be happy taking your test in test centres?
- Would you prefer to take your test in a room at school?
- Would you prefer there to be 'test days' e.g. every Friday?

Appendix G: Student Focus Group Session Plan

Security

- Do you think that some students may cheat? How do you think they could do this?
- Are you concerned that your paper could be lost?

Tailored route through education

- If you could take tests at any time how do you think school systems might work?
- How would you feel about choosing when you studied your different subjects and when you took your tests?
- If you have the choice, would you prefer individual study or classroom study with everyone?
- More able students will want to broaden their examination range
- More students will seek extra tuition from commercial tutorial companies

(*Extra questions from JISC)*

Students will be over pressurised by the new regime

Parents will force children to take examinations

More students will seek extra tuition from commercial tutorial companies

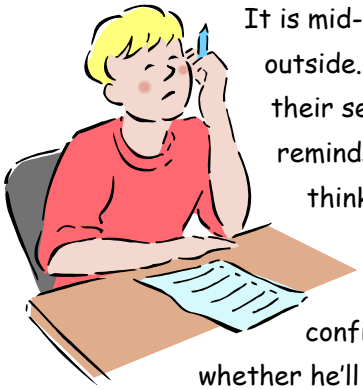
More able students will want to broaden their examination range

Appendix H: University Students' Focus Group Stimulus Material

Scenario A

Becca is waiting to enter the school hall to take her English GCSE exam. It's her last exam of the week and is worth about 60 *per cent* of her final grade. The other 40 *per cent* was coursework, where Becca wrote an essay analysing the themes in the book her class had studied; *Lord of the Flies*. Everybody from Becca's year is here, including all her friends, standing outside waiting to be let in to the school gym, which has been cleared of its usual clutter of basketballs and sports bibs. In their place is row upon row of identical chairs and tables, a clock at the front, and the exam invigilator's table, where all the exam papers are sat. It's nearly time for the exam and Becca checks once more that she has all her pens and stationery. The door to the hall is opened and Becca and the rest of her year file in and take their seats while the invigilator writes the start and finish times on the whiteboard at the front.

Scenario B



It is mid-November and it's just starting to get cold and frosty outside. Paul and his friends wait outside the sports hall to go into their second GCSE Maths modular exam. Paul feels a bit nervous but reminds himself that he has done well on his first modular exam and thinks back to all the practice papers he did. As his friends have pointed out, this exam is shorter than end of year exams so there is less to remember. All the same, Paul doesn't feel as confident about this exam as he did for the first one and wonders whether he'll have to re-sit this one when he takes the third and fourth module exams. It would mean more revision in March, when the next set of modular exams take place. He thinks it would be more stressful to revise for two at once, especially as he doesn't get a break or a holiday before going back to class to start thinking about the next set of tests, but he supposes it might be worth it for a better mark.

Scenario C

Jodie and Callum are waiting to take their GCSE Science exam. It doesn't feel like exams they've taken before; normally the whole year group is waiting with them but today it's just Jodie, Callum and a couple of kids from Mr Gregory's Science class. Their teachers entered them for the exam because they felt that they were ready to take it. Some people took the exam a few weeks ago, when they were ready. Some people still haven't taken it. Jodie likes the idea that she can take the test when she feels she is able to – it's nice to have it out of the way. Callum though, feels worried. He likes to plan and spread his revision out over the year, and he feels anxious that he hasn't had enough time to prepare. He worries vaguely that he might be in the wrong place; they aren't in the sports hall as usual because there are so few of them. He tells Jodie his fears, who reminds him that he can always re-sit later on, and that they have covered this topic fairly recently in class and he did well then.

How do you feel about exams?

We run GCSE and A levels, and we'd like to know what you think of them.

Circle the answer closest to how you feel in each case.

1. How nervous do you get about exams?
    a. I show cucumbers the meaning of cool
    b. A little bit – generally I'm fairly calm
    c. Pencil-chewing, hair-tearing-out kind of nervous
    d. Quite nervous
    e. It depends on the exam

2. What do you do when you leave the exam hall?
    a. Meet up with all my friends and discuss how it went
    b. Go home and watch TV
    c. Go out and have fun!
    d. Start revising for the next one…

3. How would you like to revise for your exams?
    a. Past papers – and lots of them!
    b. Write out notes from the textbook
    c. Reading and rereading the textbook
    d. Having revision sessions with your friends
    e. Revision? I'm not sure I understand the question…

4. What kind of exam would you prefer?
    a. Short answer
    b. Long answer
    c. Multiple-choice
    d. Practical/oral

Please turn over…

5. How many re-sits would you like?
    a. Unlimited – the more the merrier
    b. As many as you can before the end of school
    c. 3 strikes and you're out
    d. Just the one
    e. None at all

6. What type of assessment would you prefer?
    a. More coursework, less exam
    b. More exam, less coursework
    c. Half coursework, half exam
    d. All exam
    e. All coursework

7. When would you prefer to take your exams?
    a. All at the end of the school year
    b. Spread throughout the school year
    c. Every Friday
    d. When I feel I've revised enough

8. How would you prefer to be taught?
    a. In class, with the same exam for everyone in the school hall
    b. In class, doing mini-exams at the end of each topic
    c. In small groups, taking exams when the teacher thinks you're ready
    d. By a tutor, taking exams when you choose to

Thank you for taking the time to fill this in.

Appendix J: Teachers' Focus Group Session Plan

Stage 2b – Focus Group

Introduction

- Introduce researchers
- We are running a project reviewing on-demand testing
- We would like your views
- Everyone introduce themselves
- Ground rules

Opening Questions

- How many times a year can your students enter their GCSE exams (e.g. just in June, or in November and March as well?)
- Are some students more ready than others at these set test windows?
- Do you find that the current 3 test windows a year are enough?

Describe On-Demand testing

Transition Questions

- What do you think about exams being more frequent and covering smaller chunks of the syllabus?

- If test windows were more frequent do you think that students would benefit?
- Do you think more able students will want to broaden their examination range?

- How would you decide when a student was ready?
- Do you think parents might force children to take examinations?

- What are the implications of students being ready at different times?
- It will become compulsory for children to stay in education until they are 18. If students were to take their exams earlier, what would you do with them after their exams?

- Do you think more able students will seek extra tuition from commercial tutorial companies

Key Questions

- Questions will need to be repeated over time– how do you feel about that?
    - Do you think this will be a security risk?

(May need to explain live test –pre-test)

- How would you feel about questions being included in the exam which are not going to be marked?
- How would you feel about past papers not being released, only one specimen paper or a set of specimen questions?
- How would you feel if your class took a different paper from other centres but was graded independently?
- What support would you want from AQA?

Serendipitous Questions

Ending Questions

- Give a summary - Is this an accurate summary?
- Have we missed anything?

Appendix K: Post-equating Non Equivalent Groups Anchor Test

The test used for the post-equating design in Chapter 6 is included in Adobe Reader Format in the accompanying CD ROM.