1

CONTEXTUALISING THE COMPARABILITY OF EXAMINATION STANDARDS

Paul E. Newton

Abstract

The purpose of this introductory chapter is to set the scene for the chapters that follow. This will involve providing:

1. a sense of the social and political climate in which debates over comparability are conducted in England (Part 1)
2. a description of the unique organisational, structural and procedural components of examining in England (Part 2)
3. an introduction to what comparability might actually mean in this context (Part 3)
4. an indication of steps taken, through regulatory mechanisms, to facilitate the comparability of examination standards (Part 4).

1 The public face of comparability

This book is concerned with comparability: the application of the same standard across different examinations. In England, webs of comparability link a multitude of examinations: between examining boards; over time; across subjects; and so on. In fact, the number and variety of such links imply comparability challenges on a scale that is probably unparalleled in the rest of the world. Comparability dominates debate over the quality of large-scale educational assessments in England, often to the exclusion of other technical concerns. Furthermore, the extent to which comparability has been achieved is frequently a matter for high-profile national debate, often drawing in senior politicians and spokespeople from a variety of stakeholder groups. Desmond Nuttall once described comparability as 'the English disease' (Murphy & Broadfoot, 1995, p. 54). It is certainly an enduring fixation of educational discourse here.

The roots of this fixation can be traced back to the roots of examining in England. Yet, despite the controversies of comparability having been very well rehearsed, debate is as alive today as it ever has been. A glance at any of the major national newspapers during mid-August of any year, when public examination results are officially announced, clearly illustrates this.

We begin with an illustration of the significance of comparability in England, by considering how it tends to be represented through news media, such as national newspapers. Consider the following quotation, for example:

> Pressure is today mounting on the government's exams watchdog to take action over "easier" A levels.
>
> […] a leading exam expert and designer of the original modular A level, this morning blamed the qualifications and curriculum authority (QCA) for failing to ensure that all A levels are comparable in difficulty.
>
> Calling the results "depressing reading", [he] claimed that a 5.3% fail rate in English compared with nearly 20% in mathematics proved that there were different standards for different subjects.
>
> Echoing allegations already levied by headteachers, [he] said: "Students are choosing easier A levels. If people are switching from mathematics and science to the easier subjects it will mean that we will lose our ability to compete in the 21st century economy – we will lose our technological, manufacturing and engineering base."
>
> He went on to blame the QCA for failing to address the problems: "The QCA have once again failed to ensure comparability between subjects – it is their job to do so. They need to find a way of getting people examining different subjects together to discuss how they will examine to the same standards.
>
> He admitted that such a task was "not easy to do", but claimed that the QCA hadn't even begun to look into the question. He added that he thought that to make results comparable, mathematics needed to be made a little easier, and the arts subjects brought into line with the harder sciences and languages.
>
> Curtis (2003)

The quotation was taken from one of England's major national newspapers, the *Guardian*, during 2003. In different years, different comparability concerns arise. The default criticism tends to be that standards in our major national examinations have been allowed to slip over time, an inference that is drawn from evidence of continually increasing pass rates (Warmington & Murphy, 2004; Newton, 2005a). In 2003, though, the major story was an apparent lack of comparability between A level standards in different subjects, as inferred from evidence of differential pass rates.

The quotation is useful for highlighting a range of points that need to be appreciated in order to understand the necessity of monitoring the comparability of examination standards in England and, consequently, the need for this book. These points include the following:

- comparability is a major topic of national debate, filling many newspaper column-inches each year; it is therefore of major political significance

- comparability has genuine real-world importance, since it has a direct impact on students – from their subject choices to their life chances – with implications for the structure of society itself
- even those notionally 'within' the system, including senior examiners, are not averse to criticising it
- criticisms can often be fairly naïve, tending to be based on evidence from pass rates alone
- calls for action can also be fairly naïve, assuming that there must be a straightforward solution to any identified comparability challenge.

In short, comparability is a concern of major social and political significance in England. However, at the same time, many commentators fail to grasp quite how complicated the underlying issues are. To some extent, this is understandable, since comparability is a very complex and enigmatic concept indeed.

Unfortunately, the chapters that follow will not solve the complex enigma of comparability. Instead, what they will do is to help trace its roots, to help unpack its meanings, and to present a state-of-the-art account of the techniques that have been used to investigate it. They explore these issues from a peculiarly English perspective, in a context which is characterised, in particular, by the operation of multiple examining boards, providing parallel examinations, and competing with each other for market share.

2 England's qualifications system

The particular challenges of comparability in England can only be appreciated in relation to the unique organisational, structural and procedural characteristics of its qualifications system. These will be introduced below.

2.1 England

In writing this book, we debated whether it should be centrally concerned with comparability in the four 'home nations' of the United Kingdom (England, Scotland, Wales and Northern Ireland), or about comparability in England, Wales and Northern Ireland, or simply about comparability in England. This wasn't a straightforward debate, for a range of reasons.

On the one hand, England, Wales and Northern Ireland share a common public examination system. This is based primarily around the Advanced level (A level) and the General Certificate of Secondary Education (GCSE), in the context of which most recent comparability monitoring work has been undertaken. Moreover, UK comparability monitoring studies usually include qualifications offered in both Wales and Northern Ireland.

On the other hand, the impetus for comparability monitoring work has tended to come mainly from England, where GCSEs and A levels are offered by a number of

different boards. In both Wales and Northern Ireland, there is only one examining board, and comparability has been less of an enduring fixation there than in England.

The system in Scotland offers different qualification titles from those offered in England, Wales and Northern Ireland – including Standard Grade, Highers and Advanced Highers – although the assessment approaches employed and the overall qualifications structure are similar. On occasion, qualifications from Scotland have featured in UK comparability monitoring studies; some of the technical development has also stemmed from Scotland (for example, Kelly, 1976).

On balance, it was decided that the situation in England – including the operation of multiple examining boards, providing parallel examinations, and competing with each other for market share – was sufficiently distinct and salient to justify focusing the book on this country alone. Having said this, the following pages will make reference to the context in Wales and Northern Ireland where relevant.

2.2 England's comparability contexts

To understand the nature of comparability in England, the characteristics of England's principal assessment systems need to be understood: our tests, examinations and qualifications. Although often used quite loosely, the conventional application of these terms is as follows:

- tests – non-certificated National Curriculum tests, particularly at ages 11 and 14
- examinations – the major 16+ (school-leaving) and 18+ (university entry) certificated examinations
- qualifications – the full range of certificated educational assessments on offer, from school-leaving examinations to licensing tests and degrees.

These national assessments, and the extent to which they have been the subject of comparability monitoring exercises, are explained below.

Comparability of National Curriculum test standards

A National Curriculum was introduced in England during 1988. For the first time ever, it set out programmes of study in a range of subject areas as a statutory entitlement for all students in state-maintained schools. It also specified standards of attainment according to which all students would be assessed.

Prefaced by a pre-compulsory Foundation Stage, the National Curriculum is delivered in four compulsory Key Stages, through which students progress by age:

- Key Stage 1 (Years 1 and 2) ending at age 7
- Key Stage 2 (Years 3 to 6) ending at age 11
- Key Stage 3 (Years 7 to 9) ending at age 14

- Key Stage 4 (Years 10 and 11) ending at age 16.

During each of the four key stages, there is a statutory requirement upon schools to deliver five programmes of study – for English, mathematics, science, information and communication technology and physical education. For a range of other subjects – including design and technology, history, geography, modern foreign languages, art and design, music and citizenship – programmes of study are available, but their delivery is only statutory at certain key stages.

From its origins in proposals from the Task Group on Assessment and Testing (DES/WO, 1988), assessment of the National Curriculum was intended to be based primarily upon teacher judgement. Even now, National Curriculum tests are only administered at the end of Key Stages 1–3, and only for certain subjects, which means that most National Curriculum assessment (in Years 1 to 9) is still based upon teacher judgement. The introduction of the National Curriculum and, in particular, its assessment have been discussed at some length by both Daugherty (1995) and Shorrocks-Taylor (1999).

With a growing emphasis upon target setting and school accountability during the 1990s, tests of the National Curriculum grew in prominence. Since 1995, particular importance has been placed upon results from National Curriculum tests in English, mathematics and science, which are administered at the end of Key Stages 2 and 3, and are attempted by almost all students in England (since the tests are statutory). A single national test is administered for each subject, in each key stage, each year (unlike the situation that exists for public examinations).

To enable the comparison of schools, student-level results are aggregated to school level. These figures are published nationally and locally, and have been the subject of an extensive debate about their legitimacy (for example, Goldstein, 2001; Statistics Commission, 2005). Schools are compared in terms of whether their results are better or worse from one year to the next and in terms of whether their results are better or worse than those in other schools. Schools establish performance targets for their students to aspire to, and schools are set performance targets to which they must aspire.

Results are not only aggregated to school level, but also to local authority level and to national level. The figures are used as an index of the overall quality of education in England, and the Government has its own set of performance targets, expressed in terms of the percentage of students who attain specified levels of attainment in each of the National Curriculum tests at each key stage.

Clearly, the high stakes associated with test results – particularly for schools and politicians – necessitate a high degree of comparability of standards between test versions, since a new version of each test is produced each year. Each test undergoes a rigorous development process, and a variety of experimental and judgemental techniques is used to help ensure the maintenance of test standards over time.

Although national testing has operated for over a decade now, there has only been one formal comparability monitoring exercise, which explored comparability between tests administered in 1996 and versions administered in 1999, 2000 and 2001 (Massey *et al.,* 2003). The lack of monitoring work in this area explains why the title of this book specifies techniques for monitoring the comparability of examination standards, rather than also extending to test standards.

Comparability of GCSE examination standards

GCSE examinations (which are not statutory) were introduced in 1988. GCSEs – rather than National Curriculum tests – are the principal means by which 16-year-olds are assessed at the end of Key Stage 4, the end of compulsory education in England. GCSEs tend to involve students studying a subject-specific syllabus over a period of two years, and a typical student might study eight to ten GCSEs.

By present-day standards, the GCSE was a long time in the making. It was announced, many years before it was finally delivered, as 'a common system of examining at 16+', which would supersede both the existing General Certificate of Education Ordinary level (GCE O level) and the Certificate of Secondary Education (CSE).

The O level had been in existence since 1951 and was awarded on a subject basis. However, it was a relatively exclusive examination, targeted primarily at the most able 20% of students nationally. In 1965, the CSE was introduced to cater for students of somewhat lower levels of attainment, notionally the next 40%. The aim of the GCSE was to cater for all students, resulting in the widespread adoption of a form of differentiated assessment based upon 'tiers' of entry.

Nearly all students in state-maintained and independent schools study for GCSE examinations. They are available in a wide range of subjects, from astronomy to manufacturing to Welsh as a second language; although not all students will have access to all courses because schools and colleges can only offer a limited range. Students have an element of choice over which subjects to study although, for the majority of GCSE students in schools, a core of subjects is compulsory. Almost all GCSE examinations are offered by several examining boards, which operate in competition with each other; this represents a particular driver for comparability monitoring work.

As for National Curriculum test results, GCSE results are aggregated from student level to school level, to local authority level and to national level, and are used for student, school and national target setting and for accountability purposes. Moreover, as a school-leaving examination, the stakes associated with GCSEs are high for students as well. Again, then, the demands for comparability of standards are high.

A considerable amount of the early comparability monitoring work focused on O level and CSE examinations, and a considerable amount of the more recent work has focused on GCSE.

Comparability of A level examination standards

According to figures from the Department for Education and Skills (DfES, 2006), 77% of 16-year-olds remained in full-time education in England during 2004–5. Of the same cohort of 16-year-olds, 42% were studying at least one general A level or AS level (as their highest qualification), with an additional 3% studying at least one vocational A level (as their highest qualification). Equivalent figures for the corresponding cohort of 17-year-olds were: 63% (full-time education); 35% (general A/AS); 5% (vocational A level).

Since the 1950s, the A level has functioned as England's principal pre-university examination, catering primarily for students on two-year programmes. Although other pre-university examinations are available, including the International Baccalaureate and more vocationally related qualifications, the A level has remained the most popular. Having said this, while the substantial majority of young entrants to undergraduate study do so with traditional A levels (84% in 2001), the majority of entrants aged 21 or older now do so without traditional A levels (71% in 2001). In fact, mature students now comprise the majority of the undergraduate population (see Admissions to Higher Education Steering Group, 2004, p. 16).

A levels are offered by the same examining boards that offer GCSEs, again on a competitive basis. Students have complete freedom of choice over which subjects to study at A level, constrained only by the range that is delivered in their school or college. Their subject choices are likely to be influenced by career aspirations and (for those who plan sufficiently in advance) by the entry requirements of the higher education institutions to which they intend subsequently to apply.

The A level was originally targeted at only a fraction of the small percentage of the nation that studied for O levels. Partly due to its longevity, but perhaps mainly due to its exclusivity, it became known as the Gold Standard. In recent years, the A level has been offered to many more students, and is now taken by over a third of the national cohort. Students generally study for A level examinations in the two years immediately following GCSE, and many then progress directly to university, contingent on having attained a satisfactory profile of A level grades.

Assessment at A level changed quite radically with the reforms of Curriculum 2000, which constituted a response to concerns expressed in a major review of 16–19 qualifications conducted during the mid-1990s (see Dearing, 1996). From the early 1990s onwards, there had been a growing trend for syllabuses to be converted from 'linear' to 'modular', the main difference being the staging of the assessment: linear courses culminate in a suite of 'terminal' examination papers; modular courses offer assessments throughout the two-year programme. With the Curriculum 2000 reforms, all A level courses were 'modularised'. Generally speaking, this meant that the new A levels consisted of three units which corresponded to the work of the first year, and three units which corresponded to the work of the second year.

The modularisation of A levels has meant that students can now be examined in smaller chunks, at two examination sessions each year, shortly after having studied

the relevant content of each chunk. More significantly, once students have been examined on all three units of the first year, they are eligible to 'cash in' their units for the award of an AS level. The AS level is both a free-standing qualification and the first half of an A level. Students who wish to complete the full A level study their second three (A2) units and then cash them in (alongside their AS units) at the end of the second year. Students typically study four AS subjects in their first year, then continue with three of them in year two, to convert them into full A levels.

The A level is a high-stakes qualification for students, being the principal tool for university selection in England. Higher education (HE) departments use A level results in different ways. Some offer places to students contingent on their attaining a fairly specific profile of grades (for example, grade A in a mathematical subject and at least grade C in two other A level subjects, with a preference for quantitative ones). Others offer places contingent on a less specific profile (for example, at least grade C in at least three A level subjects). The Universities and Colleges Admissions Service has established a points system (the UCAS Tariff) to report achievement for entry to HE in a numerical format. In particular, it enables comparisons between applicants with different types and volumes of achievement. So, for example, a grade A at A level is worth 120 points, the same as a straight pass on a BTEC National Diploma. Similarly, a grade D at A level is worth 60 points, the same as a grade A at AS level. Some HE departments will make an offer simply in terms of the UCAS Tariff; for example, a score of at least 240 points, which equates to two grade As at A level, or four grade Ds.

As with the earlier tests and examinations, results are aggregated to school, local authority and national level, with associated stakes for individuals and institutions at each level. The effective operation of the system is a similarly high-stakes concern for those high-ranking civil servants and ministers who have ultimate responsibility for it. Reflecting the high profile of these uses and consequences, a considerable amount of comparability monitoring work in England has focused upon A level standards.

Comparability of other qualification standards

In addition to A level, which is known as a 'general' qualification, many other qualifications are available to students in post-compulsory education, typically of a more 'vocational' nature.

In a recent report (QCA, 2005a), the regulators for England, Wales and Northern Ireland identified three distinct types of vocational qualification:

1. National Vocational Qualifications (NVQs) are work-related, competence-based qualifications. They reflect the specific skills and knowledge needed to do a job effectively, and show that a candidate is competent in a particular area of work.

2. Vocationally Related Qualifications (VRQs) are also work-related, but tend to be less job-specific. They reflect the general skills and knowledge needed in the workplace.

3. Other General [vocational] Qualifications (OGs) are practical, but may not be particularly work related. They reflect skills like dance, music and drama, and tend to include graded examinations as part of their assessment.

NVQs constitute the largest proportion of vocational qualifications. Unlike GCSEs and A levels, they tend not to be assessed through large-scale examinations. Instead, NVQ assessment is individualised and relies upon each candidate being able to demonstrate to an assessor that s/he has acquired the competencies required of the job, as specified in the National Occupational Standards for each NVQ.

In the context of tests and examinations, comparability tends to relate to the test or examination itself, that is, whether it results in the award of a given level or grade to the right group of students. In the context of qualifications such as NVQs, where assessors award passes, levels or grades directly, comparability does not have the same meaning, and it can not be investigated in the same way. For this reason, the comparability of standards between vocational qualifications has not been monitored using the kind of techniques discussed in the following chapters.

2.3 England's examinations

As indicated above, most of the comparability work in England has focused on the major 16+ and 18+ public examinations: O levels, CSEs, GCSEs and A levels. The major examinations currently available – GCSE and A level – are offered, in England, by three examining boards:

1. Assessment and Qualifications Alliance (AQA)
2. Edexcel
3. Oxford, Cambridge and RSA examinations (OCR).

They are also offered by boards in Northern Ireland and Wales:

4. Northern Ireland Council for the Curriculum, Examinations and Assessment (CCEA)
5. Welsh Joint Education Committee (WJEC).

The examining boards are nowadays known by the more general term 'awarding bodies', the latter extending also to organisations that award qualifications other than public examinations. Throughout this chapter, and throughout the book, they will generally be referred to as examining boards, with the more general term reserved for reference to all organisations that award qualifications.

GCSE and A level examinations

Each GCSE and A level is classified within a subject group, as illustrated in Tables 1 and 2. Within each of these subject groups may fall a range of subject titles; and examinations for each subject title are often offered by more than one examining board. So, for example, the category Other Modern Languages includes Italian (as

Table 1 Number of (full course) GCSEs sat in the UK during summer 2006 (aggregated across boards)

Subject category	Male	%	Female	%	Male and female	Cumulative % A* to C (M & F)
Mathematics	371,875	50	378,695	50	750,570	54
English	362,007	50	359,755	50	721,762	62
English Literature	275,845	48	296,316	52	572,161	68
Science: double award	238,097	50	241,692	50	479,789	58
Design and Technology	203,118	55	168,554	45	371,672	59
French	104,825	44	131,364	56	236,189	65
History	118,082	51	113,575	49	231,657	67
Geography	118,849	56	94,620	44	213,469	66
Art	86,035	41	126,322	59	212,357	72
Religious Studies	69,184	43	90,497	57	159,681	71
Physical Education	99,614	65	53,212	35	152,826	61
Information and Communication Technology	60,888	56	48,713	44	109,601	62
Drama	37,369	37	63,439	63	100,808	71
Science: single award	47,884	50	48,490	50	96,374	25
German	42,567	47	47,744	53	90,311	69
Business Studies	51,452	58	37,905	42	89,357	60
Statistics	35,751	52	32,580	48	68,331	71
Spanish	25,287	41	36,856	59	62,143	69
Music	31,048	51	29,620	49	60,668	73
Science: Biology	33,717	56	26,365	44	60,082	88
Media/Film/TV Studies	28,718	50	28,803	50	57,521	61
Science: Chemistry	32,800	58	23,964	42	56,764	90
Science: Physics	33,031	59	23,004	41	56,035	91
Home Economics	3,042	7	43,486	93	46,528	54
Business and Communication Systems	21,106	51	20,534	49	41,640	57
All other subjects	8,359	24	25,859	76	34,218	54
Other Modern Languages	12,932	44	16,256	56	29,188	82
Other social sciences	8,009	29	19,994	71	28,003	57
Classical subjects	8,507	52	7,798	48	16,305	88
Humanities	7,779	48	8,345	52	16,124	47
Expressive arts	3,492	35	6,574	65	10,066	55
Other sciences	4,816	49	5,003	51	9,819	53
Welsh: second language	4,183	44	5,388	56	9,571	69
Welsh: first language	2,525	48	2,687	52	5,212	70
Welsh Literature	1,896	46	2,271	54	4,167	70
Mathematics (additional)	1,709	52	1,573	48	3,282	91
Economics	2,293	72	878	28	3,171	73
Irish	1,130	44	1,430	56	2,560	83
Social Science	443	35	820	65	1,263	49
Other technology subjects	1,002	90	116	10	1,118	37
All subjects	**2,467,488**	**49**	**2,534,094**	**51**	**5,001,582**	**62**

Data provided by Simon Eason, AQA

Table 2 Number of GCE A levels sat in the UK during summer 2006 (aggregated across boards)

Subject category	Male	%	Female	%	Male and female	% grade A (M & F)
English	26,821	31	59,819	69	86,640	22
General Studies	27,450	47	31,517	53	58,967	12
Mathematics	34,093	61	21,889	39	55,982	44
Biology	22,597	41	32,293	59	54,890	24
Psychology	13,485	26	39,136	74	52,621	18
History	23,634	50	23,310	50	46,944	25
Art and Design subjects	13,195	31	28,794	69	41,989	30
Chemistry	20,393	51	19,671	49	40,064	31
Geography	17,694	54	14,828	46	32,522	26
Media/Film/TV Studies	14,116	46	16,848	54	30,964	14
Business Studies	18,080	59	12,568	41	30,648	17
Physics	21,408	78	5,960	22	27,368	29
Sociology	6,488	24	20,833	76	27,321	21
Sport/PE studies	13,640	62	8,194	38	21,834	14
Technology subjects	11,086	59	7,598	41	18,684	17
Expressive Arts/Drama	5,360	29	13,312	71	18,672	18
Religious Studies	5,619	31	12,586	69	18,205	27
Economics	11,714	67	5,741	33	17,455	32
Law	6,029	40	9,212	60	15,241	20
French	4,624	32	10,026	68	14,650	35
ICT	9,052	64	5,156	36	14,208	9
Political Studies	6,722	59	4,623	41	11,345	29
Music	5,648	54	4,759	46	10,407	18
All other subjects	4,649	47	5,314	53	9,963	17
Mathematics (further)	5,106	70	2,164	30	7,270	57
Other modern languages	3,039	43	3,970	57	7,009	44
Spanish	2,133	33	4,387	67	6,520	37
Computing	5,629	90	604	10	6,233	15
German	2,369	38	3,835	62	6,204	38
Classical subjects	2,690	43	3,496	57	6,186	37
Science subjects	3,068	73	1,141	27	4,209	23
Communication studies	665	31	1,449	69	2,114	22
Home Economics	83	8	1,004	92	1,087	18
Welsh	182	19	771	81	953	20
Irish	109	33	220	67	329	49
All subjects	**368,670**	**46**	**437,028**	**5**	**805,698**	**24**

Data provided by Simon Eason, AQA

well as Arabic, Japanese, Russian, Urdu and so on), and GCSE Italian is offered by both AQA and Edexcel.

The scale of examining

Public examinations are administered on a large scale. Almost all students in England will take at least one GCSE examination during their final year of compulsory education. To appreciate this scale, the numbers of students in a typical UK age cohort are illustrated by the figures in Table 3.

Table 3 The mid-2004 projection, from the Government Actuary Office, of the number of 18-year-olds in the UK during 2007

Country	Projection
England	665,240
Scotland	64,923
Wales	40,872
Northern Ireland	26,097
E, W, NI	**732,209**
UK	**797,132**

From Table 1, it can be seen that the number of GCSE mathematics and English examinations sat each year is of the same order as an annual student cohort for England, Wales and Northern Ireland combined. (Here, 'sat' means the number of students awarded either a passing grade, A* to G, or a fail, U, in each subject.) The very small proportion of the final year cohort that does not take English and mathematics is compensated for by a similarly small number who sit the examination either as 'early entry students', or later as 're-sit students' or as 'returning adults'. On the other hand, GCSEs in certain subject areas are awarded to a small number of students, of the order of only a few thousand for subject groups such as Economics and Irish.

During the final years of compulsory education, students are allowed to exercise (a limited amount of) choice over which subjects to study for GCSE. This means that the groups of students studying each subject are not necessarily equally representative of the student cohort. For example, although for most of the major subject areas the gender balance is fairly evenly split, this is not universally true. And, for certain subject groups, gender-biased subject preferences are quite extreme. Using an arbitrary criterion of a gender imbalance greater than 40:60, subject preferences appear to be gender-biased for 9 of the 40 subject groups in Table 1. The most extreme trends are evident for subjects in the Home Economics and Other Technology groups, which are heavily biased towards female and male students respectively (see Stobart *et al.*, 1992, for an interesting discussion of gender-biased subject choice at GCSE).

At A level, since students tend to study only three or four subjects, and since the A level cohort is around a third of the size of the national cohort, the total number of A levels sat is considerably smaller. (Here, 'sat' means the number of students awarded either a passing grade, A to E, or a fail, U, in each subject.) The subject groups with the highest number of awards at A level include English, General Studies, Biology, Mathematics and Psychology, all with over 50,000 awards made during summer 2006 (see Table 2). At the other extreme, only just over 300 awards were made for A level Irish.

At A level, the evidence of gender-biased subject preference is even clearer, with 22 of the 35 subject groups exceeding the gender imbalance criterion of 40:60. Certain subject groups are clearly male dominated, such as Computing, Physics and Further Mathematics; others are clearly female dominated, such as Home Economics, Psychology and Sociology.

In all, around 6.5 million GCSE and A level awards are made each year. Since each award represents performance across a number of examination components, the number of individual examination papers, projects, practicals, pieces of coursework and so on that are taken, marked and graded is much higher.

Finally, returning to the point noted at the beginning of this chapter, Tables 1 and 2 indicate quite clearly the apparent differences in attainment between students in different subject areas. For example, 91% of the GCSE physics cohort was awarded grade C or above in 2005, in contrast with only 47% of the GCSE humanities cohort. Similarly, 57% of the A level further mathematics cohort was awarded grade A, in contrast with only 9% of the A level information and communication technology (ICT) cohort. For further useful analyses of examination statistics, see Vidal Rodeiro (2005) and Claessen (2005).

Setting examinations

Both GCSEs and A levels require students to be assessed through a series of examination components, typically comprising some combination of written papers, practical tasks and coursework projects. Written papers are designed, marked and graded by examining boards. This process is known as 'external' assessment. Practical tasks (for example, physics investigations, or music performances) tend to be designed by examining boards, and might be marked by students' teachers or by visiting examiners from the board. Coursework projects are often designed by teachers, or by students in collaboration with teachers, within tightly defined parameters laid down by examining boards. Coursework is usually marked by students' teachers, and marks awarded to coursework are moderated by examiners appointed by the board. This process is known as 'internal' assessment. Even when marks are awarded internally, the grading process is ultimately external. (Moderation and grading processes will be explained shortly.) The nature and role of coursework in public qualifications has recently received considerable scrutiny, following a review of coursework at GCSE (QCA, 2005b). Concerns over fitness for purpose, backwash effects upon teaching and potential for cheating have led to the exclusion

of coursework from mathematics and to a refocusing and tightening of arrangements in many others.

Examinations are structured differently at GCSE and A level. A level examinations are modularised, and examinations are available at set points throughout both years, normally in January and June. GCSE examinations tend still to be terminal, with examinations available only at the end of the course in June of the second year, though some have elements of modularisation with examinations available at other times.

GCSEs, unlike A levels, tend to offer differentiated written papers and tasks, that is, alternative 'tiers' of entry. The idea of tiering is to provide all students with a suitably challenging assessment experience. In subjects where differentiation occurs primarily by outcome (i.e. where a single assessment is accessible to all students) tiers are not used; for example in, history, music, humanities and so on.

Nowadays, the typical GCSE will have two tiers of entry: a higher tier targeted at the higher grades (A* to D); and a foundation tier targeted at the lower grades (C to G). This differentiated system allows for students on the higher tier to attain one of the lower grades, should they happen to perform slightly lower than anticipated (E). However, students who fail to achieve grade E on the higher tier fail the examination entirely (Ungraded); and students can only achieve at best a grade C, no matter how well they actually perform.

Examination papers and their associated mark schemes begin development some two years prior to their being administered. They are prepared by Principal Examiners, who work for examining boards on a freelance basis. Principal Examiners are usually appointed from the ranks of experienced examiners, and will generally have undergone an informal apprenticeship with a board over a period of years, as well as more formal training.

Written papers at both GCSE and A level involve a range of assessment formats, which might include selected-response questions (for example, multiple choice) and short- or long-answer constructed-response questions (which might be either structured or unstructured); questions may sometimes require the use of maps, diagrams, tables, graphs, photos and so on. Typically, written papers comprise some balance of mainly short- and long-answer questions, with relatively few selected-response questions. In the past, it was not uncommon for candidates to be allowed to choose between optional questions; this practice is far less usual nowadays.

In summary, the tradition of examining in England has emphasised the importance of validity by assessing:

- a reasonably wide range of knowledge, skill and understanding
- knowledge and understanding through written expression

- skill through performance
- depth of knowledge, skill and understanding in specific areas.

As a consequence, though, the tradition has sacrificed an element of reliability, given the problems inherent in assessments designed to more 'authentic' specifications, problems such as sampling error and marking error. Incidentally, despite the prevalence of comparability monitoring over the years, there has been no tradition of reliability monitoring in England. There is a small body of published work containing evidence of marking reliability (for example Murphy, 1978; 1982; Newton, 1996), but virtually none containing evidence of the overall reliability of examination results (Black & Wiliam, 2006; cf. Please, 1971; Willmott & Nuttall, 1975).

Marking examinations

The number of examiners required to mark all GCSE and A level examinations each year is somewhere in the region of 60,000; the majority are employed to mark GCSE written papers. Examiners are employed by examining boards under temporary contracts, during examination periods. They are often practising or recently retired teachers, and they are required to have a substantial professional understanding of the subject area for which they are marking.

Examiners score student performance according to mark schemes developed alongside the examination paper or task, and refined once the first responses are seen. Students' examination performances are referred to as 'scripts' where this can mean either a student's response to a single written paper, or to the full corpus of work completed by a student for an examination (across all components).

Examiners, for each syllabus within each examining board, are structured hierarchically. At the top is the Chief Examiner, with responsibility for the entire examination. Beyond the examination for which s/he has responsibility, the Chief Examiner will report to a Chair of Examiners, who oversees standards across different syllabuses within a single subject area, or group of related subjects. Each Chair of Examiners reports to the awarding body Accountable Officer.

At the next level down from the Chief Examiner, with responsibility for individual components – including the preparation of assessment tasks and mark schemes – are Principal Examiners. Each Principal Examiner may be supported by Team Leaders, each of whom will train, and manage the marking of, a number of Assistant Examiners. This hierarchy is represented in Figure 1.

Before any marking takes place, all examiners undergo training and standardisation in the application of the mark scheme. Again, this is hierarchically managed, as more senior markers train and standardise less senior members of their team. The aim is to ensure that the Principal Examiner's application of the mark scheme is applied by all examiners. During the marking period, samples of work are collected from examiners and the marking is checked by their supervisor. The supervisor will give feedback to

modify inappropriate application of the mark scheme and, ultimately, may decide that an examiner's marking is too discrepant to allow her/him to continue.

Figure 1 The marking hierarchy

Examination scripts, primarily written documents, are normally sent directly from schools and colleges to examiners, who mark the scripts at home. After the marking process is completed, examiners return the marked scripts to examining boards, alongside mark-sheets that contain a record of the marks awarded. These marks are either input manually or, as is more common nowadays, input automatically using optical mark-reading technology. Scripts are not returned to schools or students unless a request is formally made.

Where components are internally assessed, samples of teacher-marked work are checked by examining board moderators – who are trained to apply the agreed standard – and a teacher's marks may be brought in line with those of the moderator if they are sufficiently discrepant. Where teacher-assessment marks are consistently harsh or lenient, they are brought into line through statistical methods. Where they are inconsistently discrepant, the full corpus of work from a school may need to be moderated. A similar training, standardisation and management hierarchy exists for internally assessed components as for externally assessed components with Principal Moderators, Team Leaders and Assistant Moderators.

This approach to marking was once described by the chief executive of the regulatory authority for England as a 'cottage industry' (HoC/ESC, 2003, Ev 57). It is certainly a cumbersome and expensive process, and is vulnerable not simply to marking inaccuracy, for example in the manual addition of marks (threatening product quality), but also to other types of threat, such as postal strikes, theft of delivery vans,

depot fires and the like (threatening product delivery). For reasons such as these, there has been a drive to modernise the system – and the marking process in particular – through the development of electronic technologies. Script scanning and web-based marking are now being successfully introduced.

Grading examinations

Grading is the process by which mark scales are divided into mark bands, such that marks within each band represent a particular grade. The minimum mark required for the award of each grade is known as the 'grade boundary' mark. Grading, therefore, involves the identification of grade boundary marks.

This process is important because even ostensibly parallel versions of the same subject examination can differ, from one year to the next, in terms of how easy or hard it is to achieve marks. If (near the top of the mark range) the year 2 examination is five marks easier than the year 1 examination then, to apply the year 1 standard to the year 2 exam, the (grade A) grade boundary will need to be raised by five marks. The process of adjusting grade boundaries is what allows examining boards to claim comparability.

In contrast to many international high-stakes tests and examinations, the vast majority of GCSE and A level examinations are not pre-tested. As such, the identification of grade boundary marks is largely post hoc, as described in Chapter 3. Specially designed linking studies are not undertaken for a range of reasons. A principal reason is feasibility, given the huge numbers of examinations involved. The risk to confidentiality of the materials is another driver, as is the fact that materials are publicly available once administered live. Equally, though, since GCSEs and A levels employ complex assessment procedures, and since their content, process and statistical frameworks tend to be revised on a fairly regular basis, pre-test linking studies might often return quite contestable grade boundary recommendations anyhow.

Reporting examination results

Each year, students entered for the summer examination session complete their final examinations by the end of June. By mid-August, following an intense period of marking and grade awarding, the results are ready to be returned to students. Schools are sent notification of grades achieved by individual students, to whom this information is then forwarded (some boards send results directly to students). A level results are also sent directly to the Universities and Colleges Admissions Service, where the data are used to identify whether students have satisfied entry criteria set by the departments to which they applied. Certificates are produced somewhat later.

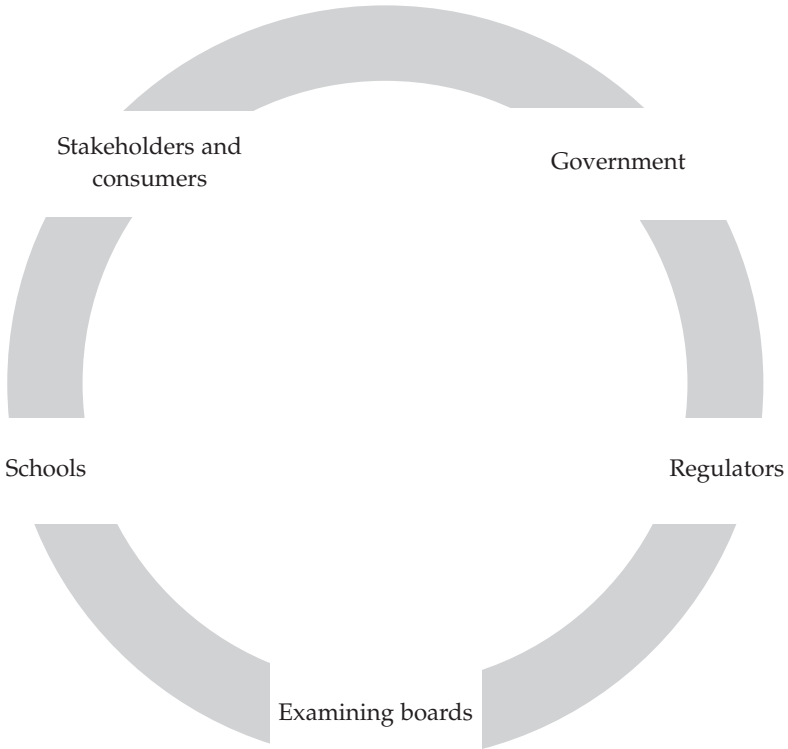
Where a school is of the opinion that a student has not been awarded the correct grade, it is at liberty to challenge the award. These challenges, known as Enquiries About Results, may lead to a student's grade either remaining the same, being raised or (less frequently) being lowered. If a school disagrees with the outcome of the enquiry, it is at liberty to lodge an Appeal. Appeals are directed, in the first instance,

to the relevant examining board. If a school is unhappy with the outcome of its Appeal, the school may escalate it to an independent body, the Examinations Appeals Board.

2.4 England's organisational and qualifications structure

Both the organisational structure and the qualifications structure influence the nature of comparability challenges in England.

Figure 2 Five degrees of separation



Organisational structure

It is possible to identify at least five degrees of separation between the general public as stakeholders in England's qualifications system and the general public as consumers of its products (see Figure 2). At the highest level, stakeholders exert power over central government, influencing qualifications policy and practice through national elections, voting on the basis of the declared political intentions of the major parties. Once in power, central government is able to influence qualifications policy and practice directly through the DfES. It also exerts power indirectly, since only qualifications and their associated courses that the DfES approves may be funded from public money.

The regulators influence policy and practice at the next level down. They are instituted and funded by central government and their purpose is essentially to ensure that the system operates in the best interests of the stakeholders and consumers. In England, responsibility for regulation of the GCSE and A level system lies with the Qualifications and Curriculum Authority (QCA); a similar role is undertaken in Wales and Northern Ireland, respectively, by the Department for Education, Lifelong Learning and Skills (DELLS) and the Council for the Curriculum, Examinations and Assessment (CCEA). These organisations also have a role in advising government on qualifications policy, and in helping government and examining boards to put qualifications policy into practice.

Associated with the QCA is an organisation called the National Assessment Agency (NAA). It is technically part of the QCA, and is located in the same building, although it operates 'at arm's length'. The NAA is charged with the task of coordinating the modernisation of the examination system, as well as with the development of National Curriculum tests (which are regulated by the QCA, hence the need to operate 'at arm's length').

The examining boards function at the next level down, with direct operational responsibility for processing GCSE and A level examinations. As noted earlier, they include:

- Assessment and Qualifications Alliance (AQA)
- Edexcel
- Oxford, Cambridge and RSA examinations (OCR)
- Northern Ireland Council for the Curriculum Examinations and Assessment (CCEA)
- Welsh Joint Education Committee (WJEC).

In addition to running GCSE and A level examinations, these five examining boards also offer a range of other, typically more vocational, qualifications. To reflect this broad base they are now known as awarding bodies, or more specifically as the Unitary Awarding Bodies.

The five examining boards operate in competition with each other. In the past, this competition focused particularly upon the content and structure of the syllabuses that they offered. Nowadays, with much tighter restrictions upon syllabus content and structure, competition focuses more upon the quality of service provided, including the quality of support provided to teachers. Whereas students have an element of choice concerning which subjects to study (complete choice at A level and some choice at GCSE), the choice of which syllabus will be delivered, and hence which examining board is chosen, is typically exercised by the subject departments of schools and colleges.

The examining boards collectively run the Joint Council for Qualifications (JCQ), in collaboration with two other boards: the Scottish Qualifications Authority (SQA) and City and Guilds. JCQ has responsibility for coordinating shared: guidance to schools; understanding of good practice in examining; research; public statements; and statistical releases.

Although only five examining boards offer GCSE and A level examinations, the regulators for England, Wales and Northern Ireland recognise over 100 awarding bodies. Many of these are small, typically offering vocational or occupational qualifications in a narrow field. In addition to the recognised (i.e. regulated) awarding bodies, it has been estimated that 900 or so unrecognised (i.e. unregulated) organisations also award qualifications in the UK (PricewaterhouseCoopers, 2005).

Examining boards deal directly with schools and colleges ('centres'), which manage the processes of registration and administration on behalf of primary consumers, the students ('candidates').

Qualifications structure

In recent years, considerable work has been done to rationalise England's qualifications system. This has involved, in particular, the creation of a National Qualifications Framework (NQF). The NQF provides a formal structure for those qualifications that are offered by recognised awarding bodies and that have been formally accredited by the regulators (more follows on the accreditation process). It is intended to help learners make informed decisions about the qualifications they want to pursue, by comparing the levels of different qualifications and by identifying different progression routes.

The NQF has nine levels, and each accredited qualification is located at one of these. Examples of qualifications located at each level include:

Level 8	Specialist awards (PhD degree level)
Level 7	Level 7 Diploma in Translation (Masters degree level)
Level 6	Level 6 National Diploma in Professional Production Skills (Honours degree level)
Level 5	Level 5 BTEC Higher National Diploma in 3D Design
Level 4	Level 4 Certificate in Early Years
Level 3	A level, Level 3 Certificate in Small Animal Care
Level 2	GCSE Grades A* to C, Level 2 NVQ in Agricultural Crop Production
Level 1	GCSE Grades D to G, Level 1 NVQ in Bakery
Entry level	Entry Level Certificate in Adult Literacy

A database of qualifications accredited to the NQF – the National Database of Accredited Qualifications – can be found from the QCA website (www.accreditedqualifications.org.uk). It does not include qualifications offered by universities and colleges of higher education, since regulation of the higher education sector is the responsibility of the Quality Assurance Agency (QAA).

3 The very idea of comparability

This chapter opened by characterising comparability – the application of the same standard across different examinations – as an enduring fixation of educational discourse in England. That shouldn't be taken to imply that comparability is a uniquely English concern; far from it – the challenges of comparability are truly international. However, they are not necessarily understood, manifested, or responded to in similar ways, and they do not always assume the same social or political significance elsewhere (for example, Wolf, 2000).

The idea of applying the same standard across different examinations is a deceptively complex one: deceptively complex because it seems so simple.

3.1 The 'straightforward' situation: parallel versions

It is easiest to understand comparability when the same standard has to be applied across two examinations that differ only in terms of the specific questions posed. In this situation, the two versions of an examination will have been constructed according to identical criteria, in order to represent the same content or process framework of assessment objectives and to satisfy the same statistical blueprint for results. Expressed less technically, this means that they will have been designed to assess exactly the same thing (for example, attainment in A level physics), using exactly the same number, kind and quality of questions. With this level of control, so the story goes, the only factor that might still vary somewhat is the overall difficulty of each examination, that is, how many marks a random sample of students from the same population would tend to achieve on average.

Figure 3 is intended to illustrate this situation, given two versions of the same imaginary A level physics examination, from 2006 and 2007, each marked from a total of 80. For each version, it represents both the mark scale (from 0 to 80) and the grade scale (from U to A). In England, examination standards are associated with grades rather than with raw marks. To decide which students are awarded which grades, the raw mark scale is divided into a series of bands. So, any student who scored between 0 and 18 marks on the 2006 version would have received a U for the component, whereas any student who scored 19 or more would have received at least grade E. The lowest raw mark for the award of each grade is known as the grade boundary mark; thus, in 2006, the E/U grade boundary mark was 19.

Figure 3 A representation of the process of maintaining standards

2006 Raw mark scale

Grade	Mark
A	80
:	:
E	25
E	24
E	23
E	22
E	21
E	20
E	19
U	18
U	17
U	16
U	15
:	:
U	0

2007 Raw mark scale

Mark	Grade
80	A
:	:
25	E
24	E
23	E
22	E
21	E
20	U
19	U
18	U
17	U
16	U
15	U
:	:
0	U

The implication from Figure 3 is that lower attaining candidates tended to find the examination easier in 2007 than in 2006. Putting this more formally, students of a certain level of attainment would, on average, have scored around two marks more on the 2007 version of the examination than on the 2006 version. Given this fact, had the E/U grade boundary on the 2006 version (19 marks) simply been carried over to the 2007 version, then it would have been easier to get a grade E in 2007 than in 2006. To ensure comparability of standards between versions, the E/U grade boundary needed to be raised by two marks, to adjust for the fact that the 2007 examination tended to be two marks easier for low attaining students.

The English are certainly not the only ones who go to great lengths to attempt to achieve comparability of standards across parallel versions of tests and examinations. Psychometricians in many countries are engaged in this kind of enterprise, albeit from slightly different perspectives on occasion. In the United States, for example, this kind of enterprise would tend to be known as equating. (See Kolen & Brennan, 2004, for a state-of-the-art discussion of techniques for equating, written largely from a North American perspective.) Instead of identifying different grade boundary marks for different versions, marks on subsequent versions would be scaled to the standard of the first. In effect, using Figure 3 as an example, a raw mark of 21 on the 2007 version would be scaled to an adjusted score of 19; that is, students who achieved a raw mark of 21 would be awarded an adjusted score of 19. When equating, the entire raw mark range is scaled in this way. The idea here is that scores

should carry the standard, such that any particular score should have the same meaning across all versions. For English examinations, only grade boundary marks require adjustment.

3.2 A complex situation: non-parallel versions

The idea of applying the same standard across different examinations becomes much harder to understand when the versions in question differ in more respects than simply the specific questions posed and, hence, overall difficulty. In the North American literature, this situation would tend to be described as 'scaling to achieve comparability' or, more generally still, as 'linking'. In fact, the term 'linking' is used in both a general and a more specific sense. Kolen & Brennan (2004), for example, describe linking as an entire continuum that has 'equating' as an ideal extreme; for two tests to be equated, they must (at least) have been built to identical content and statistical frameworks. However, they also note the tendency to apply the term more specifically, to that part of the continuum to which 'equating' does not apply: particularly, when tests have been built to different content and statistical frameworks.

Imagine that between 2006 and 2007 the A level physics syllabus had changed substantially, for example around one-third of the old content had been dropped and replaced with new content (and, perhaps, less of it). In addition, while still being marked from a total of 80, the kinds of questions asked had also changed substantially, for example, many of the old multiple-choice questions had been replaced with a single essay question. Although, notionally, the same construct is assessed by both versions – attainment in A level physics – in reality, the assessed constructs may be substantially different.

In situations like this, it is generally accepted that the two versions cannot be equated, in any strict sense. Why not? Primarily, because students who studied the earlier syllabus will have attained somewhat different kinds of knowledge, skill and understanding from those who studied the later syllabus. That is, marks on the 2006 version will reflect a certain level of 'old physics', while marks on the 2007 version will reflect a certain level of 'new physics'. As such, no adjusted score – or grade boundary – could ever carry exactly the same meaning across versions.

To a purist, it might simply seem impossible to apply exactly the same standard across different versions of an examination, when those versions differ more than in terms of difficulty. Purist logic, though, can be challenged by a more pragmatic 'common sense'. This common sense would argue that there must be some way in which something like the same standard could be applied, even across substantially different versions of an A level physics examination.

The most fundamental challenge of comparability, then, is to identify the sense in which something like the same standard can be applied across versions of examinations that differ more than in terms of difficulty. Unfortunately, comparability theory is not at all well developed, although theoretical work has begun in both the United States and in England (for example, Christie & Forrest,

1981; Mislevy, 1992; Linn, 1993; Cresswell, 1996; Wiliam, 1996; Fitz-Gibbon & Vincent, 1997; Baird *et al.*, 2000; Goldstein, 2000; Newton, 2005b).

As noted above, in the United States, where much of the theoretical and empirical work has been undertaken, the idea of linking – rather than equating – is often used to convey the essence of comparability: it is somehow like equating, but not quite as good, or as strong, or as statistically robust, or something like that. That might seem too casual a definition for an academic text; unfortunately, though, there is no clear academic consensus over what comparability/linking actually means, and explanations offered do tend to be quite woolly.

Regardless of how the theory of linking or comparability might differ from that of equating, the practice is essentially the same. It either involves adjusting grade boundary marks on version 2 to reflect the standard of version 1, or it involves adjusting all marks on version 2 to reflect the standard of version 1.

So far, in this section, we have considered the case of two versions of an A level physics examination, based upon substantially different syllabuses and examined in substantially different ways. In this situation, it might seem reasonable to assume that there must be some way in which something like the same standard can be applied. But what about a more extreme case, for example, the application of the same standard across different subject examinations?

At least in England, it still seems commonsensical to many stakeholders that there must be some way in which something like the same standard can be applied across (let's say) A level physics and A level psychology. After all, they have the same qualification title (A level), which implies that they somehow represent attainment at the same educational level. More importantly, their grades are often treated equivalently by university selectors when making admissions decisions: particularly, for subjects like law or psychology, which select from students who have made very different A level subject choices.

There would certainly seem to be a social imperative, in the English context, to apply the same standard across examinations in different subjects. Exactly what this might mean, though, is very far from obvious. Similar comparability challenges are faced around the world. The states of Australia, for example, have grappled with this challenge for many years, and have adopted essentially pragmatic solutions, based upon complex statistical adjustments. What those states tend not to be clear on, though, is the precise logic of the adjustments nor exactly what the adjusted scores mean (see Nuyen, 1986; Partis, 1997; McGaw *et al.*, 2004).

3.3 A highly complex situation: multiple non-parallel versions

The English qualifications system is probably unique in the complexity of the comparability challenges that it faces or, put another way, in the number and diversity of examinations whose standards are at least notionally linked. The case of GCSE usefully illustrates the point.

At the time of writing this chapter, the examining board OCR had three distinct syllabuses for mathematics at GCSE. Syllabus A was the standard option, with two externally examined written papers (40% weighting each) and an internally assessed coursework component (20% weighting). Both of the written papers were offered at three levels – foundation tier, intermediate tier and higher tier – with overlapping grade ranges. Syllabus B was the 'Maths in Education and Industry' option, supported by an independent curriculum body. It too had two externally examined written papers and an internally assessed coursework option, but with a different weighting between the written papers (30% and 50%). The qualification had a quasi-modular structure, meaning that the first paper could be taken early. Syllabus C was billed as the 'tried and tested' modular approach. Students sat at least two module tests from a suite of ten (15% weighting each) – the best two of which counted – plus a tiered terminal unit examination (50% weighting) and an internally assessed coursework component (20% weighting).

So, for GCSE mathematics alone, OCR had to ensure comparability of examination standards at equivalent grade boundaries:

- from one year to the next for Syllabus A (B and C, respectively)
- across the tiers of Syllabus A (B and C, respectively)
- between Syllabuses A, B and C.

Of course, OCR wasn't the only examining board to offer GCSE mathematics. In fact, at the time of writing, all five of the GCSE examining boards offered at least one mathematics syllabus. Not including pilot examinations, AQA offered three (including a statistics option); CCEA offered one; Edexcel offered three (including a statistics option); and WJEC offered two.

Each of the GCSE examining boards had to ensure comparability of examination standards for GCSE mathematics in exactly the same way as OCR. Moreover, all five of the examining boards needed to apply the same standard in the same way.

If this wasn't complex enough, within each GCSE subject area, each examining board is expected to ensure comparability of standards across subject areas. In short, a grade C in mathematics is supposed to be of the same standard as a grade C in astronomy, manufacturing, Welsh as a second language, and so on.

And if this wasn't complex enough, not only is it required that comparability of standards be ensured from one year to the next across all subject areas, but there is also a supposition that – if by extension alone – comparability of standards should be ensured from one decade to the next. Indeed, the widespread use of aggregated GCSE results for monitoring trends in educational standards explicitly requires this. This is not only despite substantial changes that are made to individual syllabuses on a fairly regular basis, but also despite substantial changes that occur less frequently in assessment formats and curriculum organisation.

Beyond GCSE, there is a further expectation of comparability of standards between qualifications at the same level of the National Qualifications Framework (NQF). This is less precisely formulated. However, it is still explicitly stated that, for example, a Level 2 NVQ represents the same standard as a GCSE awarded at grades A* to C; while a Level 1 NVQ represents the same standard as a GCSE awarded at grades D to G.

4 Facilitating the comparability of standards

Over the years, a major driver behind the move towards increased regulation has been concern over the comparability of examination standards: not simply the conspiracy theory that standards might deliberately be manipulated to improve market share, but also the belief that increased regulation could help prevent standards from losing alignment 'accidentally'. Consequently, much of the work of the regulators has focused on the comparability of examination standards.

There are three main dimensions to regulation, as it pertains to comparability:

1. specifying system regulations
2. ensuring adherence to system regulations
3. monitoring the technical quality of examinations and results.

4.1 Specifying system regulations

System regulations are specified at a range of levels, in a number of documents prepared jointly by the regulators (of England, Wales and Northern Ireland), of which the following are central:

1. *The regulatory authorities' accreditation handbook* (2003)
2. *The statutory regulation of external qualifications* (2004)
3. *GCSE, GCE, VCE, GNVQ and AEA code of practice* (2006/7).

The first two of these specify the procedures and criteria by which awarding bodies may gain recognised status, allowing them to offer qualifications within the NQF, and by which individual qualifications may be accredited to the NQF. The third specifies procedures to be followed at all major stages involved in processing the named qualifications. All of these documents can be found on the QCA website (www.qca.org.uk).

Before an awarding body is given recognised status, the regulator will conduct a review of its systems and procedures. This focuses primarily on governance, experience, expertise, financial stability and quality assurance arrangements for developing and delivering qualifications. The review is designed to ensure that the awarding body is capable of offering qualifications within the framework to the required quality standards.

Once an awarding body is recognised it may submit qualifications for accreditation. Qualifications need to satisfy the common criteria (specified within *The statutory regulation of external qualifications*) and, normally, also additional subject criteria (specified separately for examinations in each subject for each qualification type).

The common criteria specify generic regulations for the development of qualifications covering the following areas, as well as additional criteria specific to different qualification types:

- function of qualifications
- content of qualifications
- assessment in qualifications
- determination and reporting of results.

Subject criteria relate to features that are required to be common across all syllabuses developed for the same award (for example, GCSE economics), including:

- aims
- content
- opportunities for developing key skills
- assessment objectives and their weighting
- schemes of assessment
- attainment standards (i.e. grade descriptions for key grade boundaries).

Subject criteria are located on the QCA website, while the individual qualification syllabuses, developed by awarding bodies to the satisfaction of the subject criteria, are located on their own websites.

Finally, the code of practice specifies a range of regulations according to which qualifications must be processed, covering the following areas:

- the responsibilities of awarding bodies and awarding body personnel
- the relationship between awarding bodies and centres
- preparing question papers, tasks and mark schemes
- standardising marking: external assessment
- standardising marking and moderation: internal assessment
- awarding, marking review, maintaining an archive and issuing results

- candidates with particular requirements
- malpractice
- enquiries about results and appeals
- access to marked examination scripts.

Clearly, across all of these documents, there is a heavy emphasis upon commonality. And this emphasis is driven, in large part, by our enduring fixation with comparability.

4.2 Monitoring adherence to system regulations

Examining boards are monitored by the regulators against the requirements of the accreditation criteria. Not all qualifications or boards are monitored with equal frequency. The nature and frequency of monitoring is determined taking into account the level of risk posed to the public interest (using criteria such as examination entry size).

Examining boards are monitored by checking published information and by carrying out interviews, questionnaires and surveys. Visits are also made to boards and centres to observe and test their systems, and to evaluate their self-assessment activities. Procedural monitoring is undertaken to ensure compliance with the code of practice, including attending question paper evaluation meetings, standardisation meetings and grade awarding meetings.

If an examining board is identified as not complying fully with the accreditation criteria, it will be required to rectify the non-compliance within a specified period. Where necessary, a regulator has the power to apply sanctions, which might extend to withdrawing accreditation.

4.3 Monitoring the technical quality of examinations and results

The technical quality of examination grades – the principal system output – is defined primarily in terms of the accuracy of inferences that can be drawn from them regarding student attainment (for example, Messick, 1989). Evidence concerning the accuracy of inferences from grades can be categorised under one of three main headings:

- Validity – the extent to which results reflect attainment across the full range of knowledge, skill and understanding that is intended to be assessed by the examination, and nothing more than this.
- Reliability – the likelihood that a different version of the same examination would award students the same results.
- Comparability – the extent to which the same standard is applied across different examinations, or across different versions of the same examination.

The criteria and (particularly) codes of practice, described earlier, help to ensure the validity, reliability and comparability of grades by specifying structures and (particularly) procedures that are associated with high-quality outputs. They require, for instance, that marking consistency be monitored as the marking proceeds, and that feedback be given to examiners once inconsistency is detected (improving reliability). Similarly, they require that Question Paper Evaluation Committees are convened to identify changes needed to bring draft versions of question papers and mark schemes up to scratch (improving validity), and they indicate the range of information that must be considered when establishing grade boundaries (improving comparability).

There is a range of other mechanisms for monitoring the technical quality of examinations and results, including Script Review meetings. During these meetings, which target examinations identified as under scrutiny during a particular year, the question papers and work produced in response to them are evaluated in detail by independent expert subject consultants. During the autumn of 2006, for example, Script Reviews in England involved 70 consultants, 100 question papers and 1000 candidate scripts.

Controls are also put in place to monitor just how technically accurate the examination results turned out to be. As such, these tend to require post hoc investigations, designed to monitor aspects of validity, reliability and comparability. Techniques for monitoring the comparability of examination standards are, of course, the topic of this book. Comparability monitoring has, for many decades, represented a dominant concern of examination researchers in England, with only a very limited amount of research into other key concerns, such as:

- the consistency with which different markers would award grades to the same scripts (reliability evidence)
- the consistency with which different versions of the same examination would result in equivalent grades for the same students (reliability evidence)
- the accuracy of examination results as predictors of future performance (validity evidence)
- the extent to which examination results are affected by factors that ought not to affect them (validity evidence).

Early efforts to monitor comparability were, in the main, undertaken collaboratively by the examining boards (for example, Bardell *et al.*, 1978; Forrest & Shoesmith, 1985), although sometimes also by bodies with more of a regulatory overview (for example, SCAA/Ofsted, 1996). More recently, though, the regulators have funded monitoring exercises on a regular basis, far more systematically than had any of their predecessor agencies.

Since it was set up in 1997, the QCA has undertaken, or commissioned, investigations into a range of comparability contexts, comparing standards over time, across boards,

across subjects and even between qualification types (for example, QCA, 2001; QCA, 2003; QCA, 2006a). In particular, the QCA, in collaboration with DELLS and CCEA, runs a programme known as the Standards Review. This programme was established to monitor standards in five-yearly cycles in each major GCSE and A level subject area (originally comparing standards across examinations separated by a five-year period, although the reviews no longer operate on a strict five-yearly basis in all subjects). A description of the review programme can be found in QCA (2006b). The QCA also occasionally initiates Special Investigations, which investigate specific subjects for which there is some reason (such as an unusual pattern of results) to suspect a comparability problem. Although conclusions from these studies are often somewhat tentative, where they have been persuasive they have been used to guide awarding decisions in subsequent years, to bring into alignment standards that appeared awry.

Finally, the QCA has occasionally invited international experts to evaluate whether key quality assurance systems, such as those described earlier, have functioned adequately. For example, in 2001, an independent panel of advisers was invited by the QCA to review the adequacy of the system designed to maintain GCE A level standards. The invitation was extended in the context of a year-on-year improvement in examination results that had raised questions in some minds about whether rising performance standards reflected declining demands in the examinations. The panel was fairly positive concerning the adequacy of the system, despite being quite explicit concerning the problem they saw in conducting such a review. As they stated in the first sentence of their first conclusion: 'There is no scientific way to determine in retrospect whether standards have been maintained' (Baker *et al*., 2002). This would seem to be a useful starting point for the remainder of this book.

5 Conclusion

The comparability of examination standards is, and has been for many decades, a topic for hot debate in England, both within academic and professional circles and within society more generally. This enduring fixation with comparability is perhaps not surprising given the multiple high-stakes uses to which examination results are put – particularly in relation to selection and accountability – and the fact that nearly all students' life chances are affected by them. The significance of the debate is heightened by the fact that there exist multiple examining boards, offering essentially parallel qualifications within a regulated marketplace. It is given further energy by the particular features of the system that make comparability so hard to investigate, for example: the relative 'authenticity' of examinations, including the combination of different forms of assessment; the use of multiple tiers of entry for GCSE; the relatively small numbers of subjects studied for A level; subject choice, leading to examination populations with vastly differing characteristics, and to some examinations having only small numbers of entries; the almost complete absence of pre-testing; and so on. It is in this highly charged and highly confusing context that the following chapters have been written. As explained earlier, they don't claim to solve the complex enigma of comparability, but they do aim to present a state-of-the-art account of the many techniques that have been used to investigate it over the years.

References

Admissions to Higher Education Steering Group. (2004). *Fair admissions to higher education: Recommendations for good practice*. Nottingham: Department for Education and Skills.

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

Baker, E., Sutherland, S., & McGaw, B. (2002). *Maintaining GCE A level standards: The findings of an independent panel of experts*. London: Qualifications and Curriculum Authority.

Bardell, G.S., Forrest, G.M., & Shoesmith, D.J. (1978). *Comparability in GCE: A review of the boards' studies, 1964–1977*. Manchester: Joint Matriculation Board on behalf of the GCE Examining Boards.

Black, P., & Wiliam, D. (2006). The reliability of assessments. In J. Gardner (Ed.), *Assessment and learning* (pp. 119-131). London: Sage Publications.

Christie, T., & Forrest, G.M. (1981). *Defining public examination standards.* Schools Council Research Studies. London: Macmillan Education.

Claessen, M.J.A. (2005). *Provision of GCE A level subjects.* Statistics report series No. 2. Cambridge: Cambridge Assessment.

Cresswell, M.J. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein & T. Lewis (Eds.), *Assessment: Problems, developments and statistical issues* (pp. 57-84). Chichester: John Wiley and Sons.

Curtis, P. (2003, August 14). QCA comes under fire over 'easier' A levels. *The Guardian*.

Daugherty, R. (1995). *National curriculum assessment: A review of policy 1987–1994*. London: The Falmer Press.

Dearing, R. (1996). *Review of qualifications for 16–19 year olds*. London: School Curriculum and Assessment Authority.

Department for Education and Skills. (2006). *Participation in education, training and employment by 16–18 year olds in England: 2004 and 2005.* SFR21/2006. London: Department for Education and Skills.

Department of Education and Science/Welsh Office. (1988). *National curriculum task group on assessment and testing: A report*. London: Department of Education and Science and the Welsh Office.

Fitz-Gibbon, C.T., & Vincent, L. (1997). Difficulties regarding subject difficulties: Developing reasonable explanations for observable data. *Oxford Review of Education*, *23*, 291–298.

Forrest, G.M., & Shoesmith, D.J. (1985). *A second review of GCE comparability studies*. Manchester: Joint Matriculation Board on behalf of the GCE Examining Boards.

Goldstein, H. (2000). Discussion (of the measurement of standards, by David Bartholomew). In H. Goldstein & A. Heath (Eds.), *Educational standards* (pp. 138-149). Oxford: Oxford University Press for The British Academy.

Goldstein, H. (2001). Using pupil performance data for judging schools and teachers: Scope and limitations. *British Educational Research Journal*, *27*, 433–442.

House of Commons, Education and Skills Committee. (2003). *A level standards.* Third report of session 2002–03. HC 153. London: The Stationery Office Limited.

Kelly, A. (1976). *The comparability of examining standards in Scottish Certificate of Education Ordinary and Higher grade examinations.* Dalkeith: Scottish Certificate of Education Examination Board.

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer Verlag.

Linn, R.L. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, *6*(1), 83–102.

Massey, A., Green, S., Dexter, T., & Hamnett, L. (2003). *Comparability of national tests over time: Key stage test standards between 1996 and 2001*. London: Qualifications and Curriculum Authority.

McGaw, B., Gipps, C., & Godber, R. (2004). *Examination standards: Report of the independent committee to QCA*. London: Qualifications and Curriculum Authority.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan Publishing Company.

Mislevy, R.J. (1992). *Linking educational assessments: Concepts, issues, methods, prospects*. Princeton: Educational Testing Service.

Murphy, R.J.L. (1978). Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology*, *48*, 196–200.

Murphy, R.J.L. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology*, *52*, 58–63.

Murphy, R.J.L., & Broadfoot, P. (Eds.). (1995). *Effective assessment and the improvement of education: A tribute to Desmond Nuttall*. London: Falmer Press.

Newton, P.E. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, *22*, 405–20.

Newton, P.E. (2005a). Threats to the professional understanding of assessment error. *Journal of Education Policy*, *20*, 457–483.

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

Nuyen, N.A. (1986). Equating achievement across subjects: Is it possible? The Queensland experience. *Studies in Educational Evaluation*, *12*, 245–250.

Partis, M.T. (1997). *Scaling of tertiary entrance marks in Western Australia.* Osbourne Park, Western Australia: Western Australia Curriculum Council.

Please, N.W. (1971). Estimation of the proportion of examination candidates who are wrongly graded. *British Journal of Mathematical and Statistical Psychology*, *24*, 230–238.

Pricewaterhousecoopers. (2005). *The market for qualifications in the UK*. London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (2001). *Five year review of standards: A level psychology.* London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (2003). *Report on comparability between GCE and International Baccalaureate examinations*. London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (2005a). *Monitoring of vocational qualifications (2004).* London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (2005b). *A review of GCE and GCSE coursework arrangements.* London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (2006a). *Review of standards in A level computing and ICT: 1998–2004*. London: Qualifications and Curriculum Authority.

Qualifications and Curriculum Authority. (2006b). *QCA's review of standards: Description of the programme.* London: Qualifications and Curriculum Authority.

School Curriculum and Assessment Authority/Office for Standards in Education. (1996). *Standards in public examinations 1975 to 1995: A report on English, mathematics*

and chemistry examinations over time. London: School Curriculum and Assessment Authority.

Shorrocks-Taylor, D. (1999). *National testing: Past, present and future.* Leicester: The British Psychological Society.

Statistics Commission. (2005). *Measuring standards in English primary schools: Report by the Statistics Commission on an article by Peter Tymms*. London: Statistics Commission.

Stobart, G., Elwood, J., & Quinlan, M. (1992). Gender bias in examinations: How equal are the opportunities? *British Educational Research Journal*, *18,* 261–276.

Vidal Rodeiro, C.L. (2005). *Provision of GCE A level subjects*. Statistics report series No. 1. Cambridge: Cambridge Assessment.

Warmington, P., & Murphy, R. (2004). Could do better? Media depictions of UK assessment results. *Journal of Education Policy*, *19,* 285–300.

Wiliam, D. (1996). Standards in examinations: A matter of trust? *The Curriculum Journal*, *7*, 293–307.

Willmott, A.S., & Nuttall, D.L. (1975). *The reliability of examinations at 16+.* Schools Council Publications. Basingstoke: Macmillan Education.

Wolf, A. (2000). A comparative perspective on educational standards. In H. Goldstein & A. Heath (Eds.), *Educational standards* (pp. 9-30). Oxford: Oxford University Press for The British Academy.