6

CROSS-MODERATION METHODS

Robert Adams

Abstract

Aim

The aim of this chapter is to give an account of cross-moderation methods of studying comparability of examination standards – that is, methods based on the scrutiny and judgement of candidates' examination work – in order to describe how methods have evolved over time and to summarise the current understanding of how the methods work.

Definition of comparability

The methods described in this chapter pursue the weak criterion referencing definition of comparability. The criteria exist in the minds of experienced teachers and examiners, and the methods described here rely on their applying those criteria to specially selected samples of candidates' work, as expressed in examination scripts.

Comparability methods

Cross-moderation methods are then simply systematic ways of looking at candidates' work, that ought to look to be of the same standard.

History of use

Comparability approaches based on looking at candidates' work go right back to the beginnings of collaborative work by the then GCE boards – forerunners of today's awarding bodies – in the 1950s, though the first studies to be published date from the 1960s. Much of this chapter is concerned with tracing the history and evolution of the methods since then.

Strengths and weaknesses

The undoubted strength of cross-moderation methods is that they appear 'sensible', that is to say that a lay person would understand how they address the problem, for example, of comparability between boards. From the practitioner's point of view, the methods also closely mimic parts of the standard-setting and grade-awarding procedures.

The weakness of the method is that the findings can never be unequivocal. The nature of examination standards and of human judgements is so nebulous that little definitive can be said of them.

Conclusion

The evolution of cross-moderation methods set out in this chapter represents a great deal of work by the awarding bodies over the 50-plus years that have seen comparability work undertaken. In that time, the nature of examinations has changed, as has the collective understanding of what examination standards actually might be.

1 Background

One of the features of the education scene in England is that there has always been, since public examinations were introduced in the 19th century to widen opportunity in the professions, more than one institution running and certifying them. The question as to whether different certifying authorities' certificates were 'worth the same' or, equivalently, were 'easier or harder' to get has occupied the minds of users, candidates and those authorities themselves since the very beginning.

That that preoccupation has survived to this day is shown by the existence of this book. There is still public concern that the examinations and certificates offered by the current awarding bodies are equivalent. In some contexts, this equivalence is required of different syllabuses in the same subject offered by different awarding bodies: for instance, it is not unreasonable to expect that a GCSE grade A in English is worth the same, or is of the same standard, regardless of which body awarded it.

Similarly, for university entrance, offers may be made to candidates provided that they get 'BBB' at A level in a stipulated subject and (almost) any other two subjects. It is not unreasonable to expect, therefore, that, in some sense, a grade B in French is worth the same as a grade B in chemistry.

Again, it is commonplace nowadays for certain pundits to assert that the examinations are being devalued by dumbing down, and that grades are easier to get now than once they were. It is countered that what candidates have to do for their grades is changing, but that standards are not. Indeed the awarding bodies' *raison d'être* is to make sure that standards in individual subject examinations are maintained year on year.

Over the years the awarding bodies have taken seriously their collective responsibility to try to establish and to demonstrate that each of these legitimate public concerns is being met. The whole broad question of comparability investigations is the subject of this book, and this chapter addresses a particular technique for monitoring comparability, known as cross-moderation.

A note on the structure of this chapter. It starts with a definition of cross-moderation methods and moves on to speak of the entities that can form the subject of such

work. Different approaches under the cross-moderation umbrella are then classified. A major section of the chapter deals with how a cross-moderation study should be designed, followed by a section on techniques of analysing the findings that come from these studies. The problems involved in reconciling the findings from different strands of comparability studies are then described.

1.1 Definitions

The definition of comparability that this technique is predicated on is an extension of the 'weak criterion referencing' definition. If the conditions described in that definition are met, and two awarding bodies' syllabuses are comparable, then the scripts of two candidates who just got a grade A, say, with two different awarding bodies should closely resemble each other in standard. This proposition can be tested (more or less) empirically: if the scripts are found to be ostensibly not of the same quality then one explanation would be that the weak criterion referencing definition of comparability does not hold. Note that this need not be the only explanation of the mismatch, and note that a perfect match of quality does not necessarily mean awarding standards are in line. Important questions lurk under the surface about the relationship of apparently identical exam questions in the contexts of different syllabuses, for example. If one syllabus is a complete sub-set of another, then awkward questions might be raised about what comparability actually means, as well as how it might manifest itself.

The first published review of the then GCE boards' comparability work (Bardell *et al.*, 1978) concluded that 'The boards' current view is that cross-moderation involving the boards' examiners (possibly with outsiders too) is the most fruitful and sensitive of the methods available for the study of comparability' (p. 36). Nearly 30 years on, the awarding bodies might agree or might argue that statistics have become more sensitive, though cross-moderation still has its place. Perhaps a more relaxed view of 'outsiders' in the light of developments in the methods would be taken.

Complete catalogues of the studies undertaken by the boards between 1964 and 1978 can be found in Bardell *et al.* (1978); a successor, Forrest & Shoesmith (1985) carries the catalogue forward to 1983. Perusal of these two invaluable lists will persuade the reader that cross-moderation methods, as defined below, are indeed at the core of comparability research.

At about the same time as this second catalogue was published, the then Schools Council commissioned a rigorous review of cross-moderation, by Johnson & Cohen (1983), which summarised the history and current understanding of the method, suggested improvements, conducted three studies and undertook analysis of findings using generalisability methods.

The boards' own catalogue was brought up to 2000 by the publication by Cambridge Assessment of the third instalment, Bell & Greatorex (2000).

For the purposes of this chapter, the term 'cross-moderation' will be taken to mean any investigative comparability activity that is based on inspection of candidates' examination work. The term may have meaning in other parts of the examinations process, for example, techniques for establishing common assessment standards for schools' coursework; equally, some may quibble that certain script-based activities may not be considered to be cross-moderation, but for this chapter a broad definition will be assumed, so that it will deal with any approach to comparability work that consists of the judgements, by suitably equipped individuals, of the quality of candidates' examination work, scripts, coursework or practical artefacts of one sort or another. Throughout the rest of the chapter, the term 'script' will be used for the sake of conciseness to mean any sort of examination work produced by candidates.

1.2 An outline cross-moderation exercise

The bare definition of the method probably doesn't convey the actuality of what happens. This is covered in some detail in this chapter, but at the outset it might be helpful to present an outline of what all studies have in common.

The definition baldly says that the method depends on judgements of the quality of candidates' examination scripts by people suitably qualified to make such judgements. The vital elements of a study are then:

1. samples of these scripts drawn from those produced by candidates attempting the entities: syllabuses of different awarding bodies, years or subjects
2. scrutineers to do the judging of the quality of the scripts
3. an experimental design to set out which scrutineer looks at which scripts, in what order and for how long
4. attendant staff to make sure that the design is implemented and runs smoothly
5. a systematic means of recording the results of the judgements
6. a means, perhaps simple, of analysing the results
7. a report to communicate any findings of the study.

The detail that fleshes out this bare outline is covered later on in this chapter, but first a description of the evolution of the technique is described.

1.3 Evolution

Reference will be made throughout what follows to a sort of evolution in the methods used in cross-moderation studies over the period in which the awarding bodies have been active in the field. Christie & Forrest (1980) had as part of their intention the design of a comparability study that could be used and reused, transported at will to cover any comparability study. This turned out to be a vain aspiration.

The first cross-moderation study on public record, in Bardell *et al.* (1978), dates from 1964 and concerned two boards' GCE O level examinations in biology, French and Latin. The technique had been tried earlier, and reports can be found included in minutes of Secretaries' (as Chief Executives were then known) meetings. (See Secretaries of the Examining Bodies, 1952, for example.)

Since then, perhaps the main driver of innovation in comparability work has been the fact that findings have never been unequivocal: it has never been possible to say with certainty as a result of a comparability study that 'awarding body X is lenient or severe in its grade standards'. The quest for an experimental protocol that would furnish this sort of evidence has led to the restless urge to innovate that is to be seen down the history of comparability work described in this book.

But that is not the only driver. Repeated mention will be made of the convergence of syllabuses over the years. Plurality has ceased to be celebrated and the trend in England, Wales and Northern Ireland has been toward a centralised, uniform model of the curriculum, culminating in the National Curriculum itself. (Interestingly, though, the trend is being reversed and in the name of providing each child with teaching and learning opportunities that suit him or her as an individual, plurality is being allowed to creep back into schools' provision.)

The first fruits of this convergence were the round of five comparability studies based on the first GCSE examinations in 1988. The number of such studies was determined by the number of GCSE examining boards at that time, four in England with the Welsh and Northern Ireland boards jointly organising a study. All six boards collaborated in producing a coordinated design for the studies and organised the production of the final reports through the newly formed Joint Forum for the GCSE. Typically, the choice of subjects was arbitrary: the core subjects of English, mathematics and physics seemed obvious choices, the reasons for choosing the other two, French and history, are probably lost for ever. Following the next year's GCSE examinations, in 1989, studies were carried out in English literature, CDT Technology, geography, chemistry and music.

For the awarding bodies, and in particular their research staffs, the convergence has had a weighty influence on the design of cross-moderation studies. In effect, one source of variation among syllabuses has been all but removed, and studies can be designed in the knowledge that awarding bodies' syllabuses will not be so different as to defy comparison of their resulting examination scripts.

The flurry of interest in standards over time in 1995 led the awarding bodies to incorporate a longitudinal element into their next series of studies.

The increasing popularity of modular A level syllabuses during the 1990s led to a number of studies comparing their examination outcomes – examination scripts – with non-modular versions. See, for instance, D'Arcy (1997). The introduction of modular syllabuses across all subjects with Curriculum 2000 led to a suite of studies comparing their examination products with non-modular versions; the introduction of GNVQs

and latterly vocational A levels led to studies that compared them with suitably cognate 'academic' A levels. See, for instance, SCAA (1995) and Guthrie (2003).

Furthermore the evolution of methods is also susceptible to the influence of individual researchers. Much mention will be made of the revolution that was the introduction of paired comparison methods and the associated Rasch analysis of their results. This was entirely down to the bursting onto the research scene like a bright comet of one individual – Alastair Pollitt – upon his appointment to the research team of one of the awarding bodies.

Thus, the evolution of the design of cross-moderation methods, and indeed comparability studies in general, has been driven by a combination of these factors: the quest for methods that would yield authoritative findings; the response to developments in the curriculum and methods of assessment; and the particular resources and people available to the awarding bodies to design and carry out the work.

A particular difficulty that has beset the designers of cross-moderation studies is that it has always been difficult to evaluate successive designs. It has long been acknowledged that it is almost impossible to be certain that any effect identified in a cross-moderation study is the direct result of different grading standards having been applied, and there could be no other check on the effectiveness of the designs used. Evaluation has been more intuitive: a study was held to be successful if it ran smoothly, the participants thought a good job had been done and the report seemed authoritative. For this reason, it is difficult to present the evolution of the methods – save for the milestone events described above – as being logical and evidence based.

1.4 What is compared?

Over the years, the preoccupation of the GCE and GCSE examining boards and groups (lately awarding bodies)[1] has been the study of the comparability of grades in different syllabuses in the same subject. Until the mid-1980s, there was a robust tradition of diversity, freedom of choice and experimentation in the curriculum and in individual subjects within England and Wales, which meant that there was a rich variety of syllabuses with varying characteristics – different approaches, curriculum projects – to suit all tastes. A disadvantage of this richness was that the assertion that a grade B, say, in A level history represented the same standard of attainment in a host of different and deliberately diverse syllabuses was hard to justify. We shall return to this variety later, but back in the 1970s it was an important driver in the attempt to find ways of demonstrating the comparability of grades in different syllabuses in the same subject, usually across examining boards.

Most examining boards in those days offered two, three or more syllabuses in most subjects, and comparability of grades between syllabuses offered by the same examining board was as legitimate a subject for research as was comparability of grades between examining boards. The great bulk of the bibliography of comparability reports addresses these questions of cross-syllabus comparability.

Latterly (see, for example, Edwards & Adams, 2003) the Scottish Qualifications Authority has, at its own request, been included in certain comparability studies, to test the assertion that Scottish Highers are equivalent to GCE A levels. The Scottish Certificate of Education Examination Board was included in a study in O level French in 1969, though no formal report was produced.

The examining boards themselves have seen the potential for expanding the method beyond the comparability of syllabuses in the same subject and, in the past, attempts have been made to extend the methods to judge grade comparability between subjects. There are no published formal reports of the boards' early attempts at such work, though Wood (2006) has resurrected the idea with some experimental work using students (recent A level candidates) as judges, partly prompted by the wave of confidence inspired by the adoption of paired comparison methods. It appears, though, that the success of the approach might be compromised by the reading demands on these students. We shall be returning to the arrival of paired comparison methods; indeed a whole chapter (Chapter 7) is devoted to them. The early tentative attempts were in subjects that were seen as cognate, where it might reasonably be expected that suitably qualified expert scrutineers could make judgements about the quality of scripts in, say, French and German. These early attempts were not seen as particularly successful, in part because of the unconvincing precept that any human agency is capable of making these highly subjective judgements across syllabuses in different subjects, though UCLES (for instance, see Elliott & Greatorex, 2002) have undertaken some between-subject work in their international examinations.

A question of perennial interest, in the world at large as well as to the awarding bodies, is that of grade comparability over time. It is a commonplace of saloon-bar (if not public-bar) theorising that A level grades are not worth half of what they were 20 years ago. *Prima facie* support for this proposition is that so many more students get them now than did then; indeed the theory has progressed so far out of the saloon bar that it has acquired the importance of its own technical-sounding name: grade inflation. It is not clear, though, whether this phenomenon is simply the greater profusion of grades, or the supposed concomitant erosion of standards. Wherever the truth lies, the awarding bodies have attempted on a few occasions, for example, Quinlan (1995), to harness cross-moderation methods to scripts gathered from different examinations in different years, based on the same syllabuses, or even different successive syllabuses. Bell *et al.* (1998) provide another example.

Prompted by a supposed public unease SCAA (predecessor of QCA) and Ofsted[2] themselves undertook a script-based study to compare standards over time in three subjects at A level and GCSE (SCAA/Ofsted, 1996). The conclusion in all three subjects was that they had all changed, and that whether the changes represented diminution or enhancement of standards was a matter of the values of the observer. Some tentative comments were made about standards, but the tone was cautious.

The present author holds an extreme view that comparability of standards over time, even year on year in principle, is a meaningless idea, given the ineffable web of changing social, cultural and economic influences on the context of the whole idea of

education, and in particular, schooling and examination performances. (See Adams, 1994.) At the same time it is recognised – having tried it – that this is a hard line to sell in the saloon bar; it also has to be accepted that many processes – university admissions, for example – have to be predicated on the assumption that standards *are* comparable year on year.

Again, the success of this sort of study over time was regarded as limited and the awarding bodies continued to put their resources into the study of comparability between their syllabuses in the same subject at the same sitting.

Curriculum 2000 saw the incorporation into the mainstream of modular GCE A level schemes as well as the introduction of so-called 'vocational' A levels, which had evolved from GNVQs and relabelled to establish their equivalence to 'academic' A level subjects. This gave at once opportunities for different applications of cross-moderation methods. The studies of 1996 examinations, for example, Gray (1997), concerned comparison of the grade standards of modular and non-modular syllabuses; in 2002, in, for example, Guthrie (2003) the comparison of 'academic' and 'vocational' syllabuses in business studies was incorporated.

Special mention should also be made of various one-off studies that have been based on scrutiny of scripts. GCSE French, for instance, has had, since its inception, a highly particular form of aggregation of the four language elements (reading, writing, speaking and listening), each of which could be offered for examination at a higher or foundation tier[3]. The myriad possibilities of attainment that would lead to a grade C, say, were such that it was impossible to identify a 'just-passing grade C candidate', the type of candidate whose work forms the routine material for cross-moderation studies. Various methods have been devised to deal with such subjects, some of great ingenuity. (See, for instance, Fowles (1989), one of the first series of GCSE studies that had to contend with the points system for aggregation of component results to determine final grades.)

2 Classification of cross-moderation methods

Cross-moderation methods can be characterised and classified in a number of ways, some overlapping, some independent. The following classes can be established by perusal of the historical catalogue of cross-moderation studies.

2.1 Identification, ratification, re-marking or distribution studies

This categorisation expresses the basic design purpose of a cross-moderation study. An *identification* study seeks to identify afresh grade boundaries in a range of candidates' scripts, and then compare those so identified with the actual grade boundaries determined at the operational grade awards conducted by the examining boards. Thus, typically, a carefully selected range of scripts – from the indifferent to the excellent – are set out in a metaphorical or actual row, for each board. Scrutineers move along the row, and scrutinise each script. Periodically they declare that they have come to a grade cut-off. 'Script i is grade G standard; script i + 1 is grade G + 1 standard.' This is repeated over many scrutineers and it is relatively straightforward

to compare these decisions, board by board, with the actual grades as determined by the award. An outstanding example was the study conducted on GCSE CDT Technology, where candidates' work under scrutiny included actual artefacts as well as written papers, which were literally laid out in rows in a hotel basement (more accustomed to hosting wedding receptions) in Cambridge. (See Patrick & McLone, 1990.)

This approach could be (and in some cases has been) used by a single board on a single syllabus to check its grading standards with scrutineers other than those responsible for the operational award. According to our definition, this would still be a cross-moderation study. It was also used to powerful effect by Good & Cresswell (1988) to investigate some of the characteristics of awarding in tiered examinations.

A *ratification* study, by contrast, takes scripts deliberately chosen to be near certain grade boundaries – typically A/B and E/U at A level; A/B, C/D and F/G at GCSE – and judges them according to whether scrutineers from other boards agree that they are of a standard typical of that boundary, in their experience. Most studies are of this form. Ratification studies, which once relied on scrutineers' internalised notions of grade standards, have been transformed into studies based on paired comparisons of scripts, which comparisons don't need internalised grade standards in the scrutineers at all, but rely on intuitive 'snap' judgements – barely articulable – of script quality.

Re-mark studies are for the moment obsolete. They are characterised by scrutineers' re-marking of scripts selected to be at important grade boundaries, using other boards' mark schemes so that they can absorb the standards as they mark. An example is to be found in an O level study in three subjects (JMB & London, 1966). See also Christie & Forrest (1980).

In a *distribution* study, the boards supply equal numbers of scripts at each grade and scripts are independently judged, producing a new distribution. Johnson & Cohen (1983) used such an approach in their work for the Schools Council, one of the predecessor bodies of the current QCA.

2.2 The basis of judgement

All cross-moderation methods, according to our definition, depend upon scrutineers judging scripts against some standard. Studies can be categorised according to what those judgements are made against. These can come from a variety of sources.

Defined by the study

In the days when diversity in the classroom and syllabuses was celebrated, part of the business of a practical comparability study was to decide exactly how scripts, or for that matter other material, were to be compared. Thus, the first part of a cross-moderation study would be a sort of committee meeting to establish the criteria or dimensions against or along which judgements were to be made. In one case (UCLES, 1981) this process broke down: the assembled scrutineers were unable to find enough ground in the syllabuses common enough to make judgements of script

quality in the time available. Similarly, a 1976 study (Bardell, 1977) reports that not only were independent scrutineers recruited, but they too were unable to agree upon the terms upon which scripts should be judged. Nevertheless, the study went ahead. In other cases, though, the method appears to have functioned adequately. See, for instance, Francis & Lloyd (1979). This report is also instructive in what it reveals about the discretion allowed scrutineers in designing the study as they went along. They had pretty well free rein; nowadays studies are more tightly designed and typically scrutineers are more or less told what to do rather than asked what they think they should do.

Internalised standards

Given the convergence of syllabuses at both 16+, with the advent of GCSE with its normalising national criteria for syllabuses, the National Curriculum, with its even greater control on subject content, and at 18+ with A level common cores and then restrictive subject criteria, the need for an identification of criteria phase has largely passed. Syllabuses are now so similar that the importance of at least one source of variation has been much reduced in comparing syllabuses.

It thus became ever more reasonable to ask experienced senior examiners to look at scripts from other boards' examinations and ask them to make judgements of quality. Typically, an examiner from Board *A* could be shown a script from Board *B* and asked to say whether he or she was not surprised to see this script as representing borderline grade C, say; or, if he or she was surprised, whether it was because the script was too good or too bad for that grade boundary. It transpired that these judgements could be made quite quickly, and so large bodies of data could be assembled in reasonable time.

A phase of syllabus scrutiny was nevertheless retained, partly as a worthwhile experiment in its own right, to judge the relative demands of syllabuses as a separate comparability strand, and partly as an exercise in familiarising scrutineers with the various syllabuses under study. In this context, 'syllabus' is understood to mean question papers and mark schemes, as well as the description of the cognitive content that the examination will address. The studies of GCSE subjects in 1989 dropped this process, reasoning that the syllabuses were so similar, because of the constraints of the National Criteria for GCSE subjects, that a review would be unnecessary. Hindsight suggested, though, that it was still a good idea and most studies since 1989 have included such a review.

Comparison with grade descriptions

One view of the first of these approaches, described earlier, is that the scrutineers begin the study by deriving a set of grade descriptions for the subject in question, and then go on to compare scripts with those descriptions. When the vogue for grade descriptions as part of syllabuses arrived in the 1980s, this had been done for them, and scrutineers were spared the task of deriving them. Some studies made use of them. In particular, when work on grade descriptions for GCE O level subjects

started in 1980, a study (Forrest & Williams, 1983) was designed and undertaken particularly to exploit them.

The general consensus, though, about grade descriptions, whether as a basis for comparability work or for awarding of grades, is that they can never be precise enough to describe the differences in scripts that are one mark apart; grade awarding ultimately boiling down to judging such scripts as 'in' or 'out'. Further, the impulse to describe in ever greater detail the characteristics of a typical script at some grade has been found to lead to an atomisation of the subject as expressed through examination papers. Further still, the fact of the matter is that one of the cornerstones of the British public examination system is the principle of compensation: that a candidate can make up for a less than competent performance in one part of an examination by a very good performance elsewhere. Grade descriptions don't sit too comfortably with this principle.

Another difficulty is that if a grade description is meant to be a description of a 'typical, mid-grade C' for example, then in its own terms, it is no use for determining a 'just-grade-C' performance; and if it purports to be a description of a 'just-grade-C' performance then it is inconceivable that it should differ from the description of a 'top-grade-D' performance.

3 The design of cross-moderation studies

3.1 Organisation and location

The emphasis in what follows is on designing a ratification study, since these are by far the most frequent sort of studies organised. The detail will be more or less applicable to other sorts of study, but the principles will be found to be applicable in those sorts of study too.

At the outset, let it be said that a cross-moderation exercise cannot be over designed. The role of scrutineer at a cross-moderation exercise is not one of life's more pleasing byways, and one of the ways in which the organisers can palliate the experience for scrutineers is by making sure that the event runs as smoothly as possible. It may be that an 'objective' scientific experiment is being conducted, but the instruments – that is, the scrutineers – are distinctly human, and will respond to being treated as humanely as possible. Also, of course, good organisation will mean that as much data as possible can be obtained during the study. A strict timetable for the whole exercise should be devised, with clear breaks for refreshment and meals. An example from the study in GCE geography (Edwards & Adams, 2003) is shown in Appendix 1.

To get the necessary amount of data, especially if two, three or even four grade boundaries are to be addressed in a study, it is necessary for it to last for about two days. Any longer than that and scrutineer fatigue would become a serious matter, and it might be doubted whether the quality of judgements could be sustained. For this reason, if for no other, the tendency from the earliest studies has been for cross-moderation exercises to be held in comfortable hotels, so that scrutineers get the

feeling of being pampered in exchange for the mind-numbing activity that the exercise requires of them.

Questions about who the scrutineers are, what scripts are available and what is done with them are addressed below, but, in general, it may be said here that scrutineers respond well to knowing exactly where they should be at any particular time, exactly what they should be doing and having arrangements made for them to do it as comfortably as possible. All this needs planning and arranging with the hotel in advance. The hotel also has a part to play in ensuring that refreshments are available at the times set out in the programme for the study, and that, for instance, everyone can be served with lunch in the time available.

Typically, a study will entail scrutineers sitting in small groups to work on scripts; these groups should be located in separate rooms with enough space for comfort. The necessary scripts should be set out in advance of the working session. Some designs require that scripts be passed around the scrutineers and it may be that the organising body will need to have staff on hand for each scrutineer group to assist in the management of scripts.

3.2 What are the entities to be studied?

It may seem an obvious remark, but it should be established unambiguously at the outset of the design of a study exactly what entities – usually syllabuses – are to be the focus of the study. There will then follow an identification of what materials – scripts in our shorthand – will form the raw material of the cross-moderation. This will be an essentially practical decision. Most syllabuses will have a coursework component and if this is to be included in the study, it will be necessary to make sure that enough coursework is available. And, of course, some coursework doesn't come neatly in folders. There may be artefacts of greater (e.g. furniture) or lesser (e.g. Yorkshire puddings) durability, musical compositions, gymnastic or dramatic performances: in each case it will be necessary to decide exactly which components are to be used. Where there is difficulty in finding the complete work of candidates, for example, in studies involving modular syllabuses, it is often necessary to use 'synthetic candidates', with scripts of different individuals chosen at the necessary component or unit grade boundary.

Again, to reduce sources of variation as much as possible, attempts should be made to choose scripts where, when there is a choice among questions, the same questions, or sections or even papers, have been chosen by the selected candidates.

3.3 Who are the scrutineers?

In general, the terms of the experimental design will make clear who is to make up the team of scrutineers. The majority of studies over the years have relied on the experience of awarding bodies' senior examiners in the subject in question. The intention of any study will be to generate as many judgements as possible during the time available, bearing in mind the limits to what can be expected of individual scrutineers. In this case, then, the more senior examiners that can be assembled the

better, though for most comparability studies about three suitably experienced (Chief) Examiners from each awarding body seems a reasonable number. The number is also limited by cost, though with the diminution in the number of awarding bodies this pressure has been eased.

For reasons to do with the experimental design, where scrutineers can be thought of as 'representing' their parent awarding body, it seems sensible to recruit the same number from each. There appears to be a 'home-and-away' effect in judging, with scrutineers reacting differently to scripts from their 'own' syllabus as they do to scripts from others'. This is evened out by having the same number of scrutineers from each awarding body. The existence of this effect is well established and has been referred to in, for instance, Forrest & Shoesmith (1985). For this reason, many studies are designed so that no scrutineer ever looks at scripts from his or her own awarding body.

Inevitably, scrutineers may be prevented at the last minute from taking part in a study, or may even have to leave a residential study half way through. Every effort should be made to have reserves available, though it is recognised that this is more easily said than done.

The story of cross-moderation methods has been, as outlined in the section on their evolution, one of experimentation interspersed with refinement. One innovation, perhaps born of a desire to try something different on the part of the collected examining boards' research community, or perhaps in response to pressure from regulators, or perhaps again to add legitimacy to the enterprise, was to replace experienced Chief Examiners, with their internalised standards, by independent subject experts, who could perhaps be relied upon to make judgements based on what they saw in scripts, unsullied by any examination experience. In seeking suitably qualified subject specialists not associated with the examining process, Local Authorities (LAs), schools and universities have been consulted, their ranks being identified as a well-populated source of the necessary experts. Three studies in GCSE subjects (Jones, 1993) tried this approach, though many variants on using independent judges had been tried over the years (Jones & Lotwick, 1979; Massey, 1979; and see previous comment about Bardell, 1977). Experience in the more recent studies was mixed, but in general the studies were not regarded as great successes. For one thing, the judgements seemed very difficult for the experts, who took a very long time over the task, and seemed to want to pore over examination papers and syllabuses, perhaps in a vain attempt to turn themselves into the sort of examination veteran that we were trying to avoid. Often, the criteria upon which the independent judges were to make their judgements were articulated in a preliminary part of the study, but often, too, they were left to the judges themselves.

The introduction of paired comparison methods in cross-moderation studies, which rely upon snap judgements of pairs of scripts, one being identified as 'better' than the other, by as many judges as possible, has raised exciting prospects of using all sorts and conditions of person as a judge. For instance, Wood (2006) piloted the use of students to make judgements about comparability between subjects – recent A

level candidates who did the two subjects in question at A level – to judge pairs of scripts, one from each of the subjects. Experimental work is also taking place in using these methods for grade awarding year on year, in which circumstance any number of individuals can be called upon to make the comparative judgements, and can do so in the comfort of their own homes without having to travel to a residential meeting for the purpose. A full discussion of these pair-based methods is the subject of Chapter 7 of this book.

3.4 Which scripts should be used?

The principle of what scripts to look at can be easily expressed. In a ratification study, the most commonly used design to compare standards of two syllabuses at grade C is to assemble the complete examination work of a sample of candidates who just attained grade C in each of the two syllabus' examinations (i.e. attained the minimum aggregate mark for the award of grade C), and then to get scrutineers to make judgements about their quality. If the work of those representing Board A's syllabus is consistently judged to be better than that of those representing Board B, then an effect has been identified. One explanation for this effect may lie in differences in grade standards between the two boards. (A brief discussion of one reason why we can only say 'may' here is given later in the chapter, when a central conundrum regarding standards is described.)

The principle of compensation, referred to above, means that the complete work of candidates attaining the same total mark may, and indeed probably does, look very different in terms of the questions the candidates have attempted and the relative successes and lack of success on those questions. Thus, scripts representing work of any given grade may look very different from each other. Further, where an examination has several components, there will be markedly different profiles of performance across those components by candidates who have the same grade. Further still, candidates may be allowed to choose different components, sections or questions.

Because of concerns expressed by scrutineers taking part in the 1988 GCSE studies, that they found it difficult to judge work with markedly different profiles of performance across components, the custom has evolved of trying to control some of this variation in ostensibly similar grade performances by specifying quite closely the actual marks that components and individual scripts should bear. It seems reasonable to specify that a candidate should have as uneccentric a profile across components as possible: in other words, candidates whose scripts are to be selected for inclusion in a study should have a balanced performance across components. 'Balanced' may be defined in a number of ways: a useful statistical definition is that a candidate's performance is balanced if each component score has the same z-score in the components' mark distributions. For each score in the distribution, the z-score is defined to be the deviation of the score from the mean of the distribution (including the sign, plus or minus) divided by the standard deviation. In most cases this will be equivalent to the scores being the same percentiles of their score distributions. A

description of the effects on examiners' judgements of using balanced scripts is found in Scharaschkin & Baird (2000).

There is a difficulty here in that such balanced performances are quite rare: typically, candidates produce lop-sided performances, showing strength in parts of the examination and relative weakness elsewhere, so we are apparently led to consider comparability judged on the basis of abnormal candidates. Nonetheless, it is generally reckoned that this restriction is a price worth paying to remove a source of variation in scripts representing the same awarding body, and to make the scrutineers' task less difficult.

Balance is also a useful way of establishing the equivalence of scripts where an ephemeral or unwieldy component has to be left out of the study. Thus, if a practical artefact is to be ignored, making sure that the components that are to be included in the study are equivalent can be achieved by using a strict definition of balance in those components' scores. The concept of balance can also be used to select scripts where components of different weights have, for logistical reasons, to be left out of consideration in a cross-moderation study.

If possible, the principle of balance should be extended into the individual script itself: it will make comparisons of scripts easier if the performance on individual questions within a script is fairly well balanced too. It is accepted that this is difficult to attain in practice, though extremes are easy to detect. For example, a grade C candidate at GCSE who has one completely correct answer in his or her script, and nothing else, is presumably so atypical in most subjects that comparing the performance with any other is difficult. Scripts with rubric infringements – that is where candidates have not followed the instructions concerning which questions to answer and which may be regarded as quixotic if marked in accordance with the rubric (e.g. marking the first five questions on the script where five were required but where the candidate had attempted six) and hardly representative of attainment at the grade boundary – should be avoided at all costs.

3.5 How many scripts?

All decisions about the scale of a cross-moderation exercise are based upon the fact that scrutineer time is a scarce resource, if only because there is a limit to the amount of judging that the average human mind is capable of doing at one meeting. Thus, the number of judgements that can be expected and the resulting quantity of data that a study will yield are not limitless.

Further, whatever the judgements of scripts are based on, comparisons of pairs of scripts or comparison with a standard, there is virtue in having the same script or scripts judged many times, so that aside from the main analysis of the outcomes, subsidiary analyses can be undertaken to shed light on the consistency of judgements and the idiosyncrasy of judges. Greatorex *et al.* (2003) give a discussion of this sort of analysis and its consequences.

Also, it is to be recognised that some awarding bodies are smaller, in terms of numbers of candidates, than others, and in all but the core subjects there may be difficulty in finding enough balanced performances to meet the experimental design. There will be other practical considerations to be taken into account when selecting scripts or artefacts: in the case of the latter, organisers of a study may well have access to only a limited amount of work; in the case of scripts, it makes good sense not to choose scripts marked by an examiner who had to be adjusted during the customary post-marking check on the standards of markers, for instance.

To balance these factors, the practice has evolved to its present state of using about five scripts per boundary per syllabus as being representative while allowing repeated judgements to be made, though in the past widely different numbers have been used. It seems to be recognised everywhere that all syllabuses are equally important in comparability terms, and that the total entry for a syllabus has no bearing on its importance: grade standards exist in some absolute sense and not in a sense weighted by entries for different syllabuses.

3.6 Component or complete work?

In so far as grades are *awarded* for subjects and not for components, it is clearly desirable that, if possible, cross-moderation judgements of grade standards are made of the complete work of candidates. Against this ideal are to be set the practical realities of making the judgements, which may have to be based on, say, two examination scripts and a coursework folder. But again, it might be argued that grade awarding decisions are made at the level of examination component, particularly at A level, with its post-Curriculum 2000 unitised schemes. If cross-moderation is to resemble grade awarding, then it is preferable that judgements should be made at component level too. In former times, when candidates sat all the examination components at the same session, it was possible to assemble the complete examination work of candidates for confirmation of grade standards, even if the mechanics of the awarding process were carried out by components. Sometimes, of course, a component such as practical work will have been dropped from the study on pragmatic grounds.

There are logical difficulties in comparing components, especially A level unit tests, which may cover different content; may, in general terms, be of different difficulty or levels of demand; and which may be of different weighting in contributing to the final A level grade. Such considerations should be taken account of in designing the cross-moderation exercise.

3.7 Which grade boundaries?

This again is a balancing act. On the one hand, if fewer boundaries are covered, there will be more judgements and more data about those that are; on the other hand, to set up the study, and assemble the materials and scrutineers gives a rare chance to get information about as many boundaries as possible. In general, each case should be treated on its merits.

At GCSE, the grade boundaries regarded as important are the C/D boundary, which is still held by lay persons to represent a 'pass' performance, the A/B boundary and the F/G boundary. In most GCSE subjects grade C is available on two tiers of the examination, and it may be regarded as of paramount importance that grade C is equivalent on either tier. Grade A, the boundary which is determined by awarders' judgement rather than arithmetically, is also important. Also determined by judgement is grade F, possibly because it was held to correspond to the former CSE grade 4 standard. The F/G boundary is for this reason often included in a comparability study, though experience suggests that the evidence of candidates' attainment at this level is sparse, and judgements of quality are difficult to make.

At GCE A level, only two grade boundaries are currently determined at grade award by judgement – A/B and E/U. The latter still counts as the 'pass–fail' boundary, and as the numbers of grade A awards continue to grow, and universities ask for grade A passes correspondingly more often, comparability of grade A awards is increasingly important. In former times, all grade boundaries – if you go back far enough – were made by judgement, and this is reflected in the boundaries chosen over the years for study.

In unitised schemes for A level subjects, it is virtually impossible, since awarding bodies do not have the warehouse space to keep all scripts indefinitely, to assemble the complete examination work of any candidates, so there is a good reason to carry out cross-moderation exercises at the level of the individual units at any grade boundary. Since the aggregation of six (in a typical A level scheme) bare grade A performances in the units is bound to result in a bare grade A for the subject, this is quite acceptable and does at least mean that the tasks of judgement are manageable: each end-of-unit test generally generating a single examination script. But where units do not correspond particularly between awarding bodies' syllabuses, synthetic candidates could be assembled across the units to provide a sort of whole grade A, for instance, performance.

Compare this with GCSE, where awarding is carried out by component, but where Indicator 2, for which the aggregate grade boundary is arrived at via the average percentages of candidates attaining each component grade boundary, intervenes in the aggregation of components to give subject grade boundaries (see also Chapter 3). It is then by no means the case that a collection of bare grade A component performances will result in a bare grade A in the aggregate mark scale. This raises a whole discussion about where grade standards actually reside: the aggregation of notional component grade boundaries into a grade boundary in the aggregate mark scale will depend in its effect in part on the degree of correlation among the components. So the relationship between component and aggregate grade boundaries is determined in part by the characteristics of the particular candidates taking the examination. This seems somehow to offend natural justice, and the idea that candidates' performances should be judged against standards that are in some sense independent of the actual group of students doing the examination. Put bluntly, it is reasonable to assert that the grade one candidate gets should not depend upon whether or not another person entered for the examination.

A neat way round this problem is to calculate component grade boundaries in such a way as to make sure that they add up to the aggregate boundaries. This can be achieved using equal z-score methods, so that the component grade boundaries thus obtained can be represented as the average performance on the component of candidates who just got a grade C, for example, *for the subject*.

3.8 Randomisation

It is a matter of good experimental practice to randomise or balance every possible feature of the exercise. It is usual to group scrutineers together into two or three groups to address one grade boundary each, perhaps. The allocation of scrutineers to groups and the allocation of scripts to groups should, if necessary, be randomised. For the prevention of any conceivable bias, the order in which scripts are presented should be randomised, and a randomised sequence recorded in advance of the exercise, whether it is pairs of scripts or individual scripts. Where pairs are being judged, it is preferable that a new pair be taken for each comparison, rather than a scrutineer hanging on to one script and changing the second. Appendix 2 shows the randomised sequence of paired comparisons from a recent study, in which the scrutineer doesn't ever look at scripts from his or her own boards' examination.

On a broader front, it is possible to conceive of an incomplete, though balanced, design in which scrutineers only look at some of the other awarding bodies' scripts. (Or, indeed, only some awarding bodies are represented at all.) Received wisdom opposes these refinements, and a discussion of why this should be is to be found in Forrest & Shoesmith (1985, p. 43).

Experience suggests that among scrutineers a group dynamic can arise, with certain individuals competing to do as many judgements as they can. This may be harmless, but it might actually influence the quality of the judging. A way of addressing this question is to stage a completely replicated cross-moderation exercise, with every detail of the design identical, except for a different crew of scrutineers. The extent to which the findings agree would give valuable insight into the merits of the method. Replication would address all sorts of other questions too, including how much confidence we could have in the findings and taking any action as a result.

Such an experiment was conducted in 2004 (Jones & Meadows, 2005), and the results were encouraging, two independent versions of the same study producing similar results. There is scope for more such studies to be undertaken.

The one area where randomisation is not possible, or at least tends not to be practised, is in the selection of the scripts that are going to represent the awarding body at the cross-moderation exercise. If balanced scripts are sought, there may be precious few to choose from, and detailed inspection of scripts chosen to be balanced and unremarkable will mean that they will be far from typical and far from random. It is suggested that, nevertheless, comparability of grade standards can be realistically approached via these non-randomly selected scripts. To deny this – to assert that two awarding bodies may be comparable in the cases of balanced scripts,

but not comparable in cases of unbalanced performances – would take some dexterous rhetoric.

Note, though, that the selection of scripts is based on the same premise as awarding: two scripts with the same mark on them are *de facto* equivalent, however unlikely that may seem in particular cases. How often have phrases like 'But this 62 is better than that 63' rung out over awarding meetings? This question, which also has a marked effect on grade awarding, is fully discussed in Cresswell (1996).

3.9 Real scripts or copies?

The orthodoxy of comparability work has always been that it is essential to work with real scripts rather than copies. The step of copying might introduce variation in the scripts that was not there in the originals. Where single scripts were passed around a group of scrutineers, this was no great handicap, but developments in paired comparison methods, that ideally require the same script to be in more than one place at the same time, are tending to make the insistence on original scripts a luxury that cannot be sustained. An associated question is whether the original examiners' marking of scripts should be left visible or whether it should somehow be removed. Whereas it is thought that scrutineers may be influenced by seeing the marks awarded to a script, only a little progress has been made to identify an efficient, cheap and quick way to remove the marking, which is always in red ink, using filtered photocopying to remove the marking.

In particular, the display of scripts on screen, either because they were composed there or because they had been scanned, raises such a host of possibilities that the advantages outweigh the purity of using real scripts. For instance, for paired comparisons, randomised pairs can be set up in advance and images loaded onto disk. These can then be shown on split screens to facilitate comparison. Moreover, the comparisons can be done by judges in the comfort of their own homes; any number of judgements can be collected over a reasonable timescale, and any number of scrutineers can be used.

Parallel developments are taking place in awarding methods, where scales can be aligned from successive years' examinations using paired comparisons of scripts from the two examinations. There are fewer limits to the sort of person who can be asked to judge pairs: the ownership of awarding can be thus extended to all sorts of interested parties. Indeed, ultimately, the marking of scripts could be abolished and awarding carried out simply by paired comparison methods. A full discussion of these exciting possibilities is given in Pollitt & Elliott (2003).

3.10 Timed judgements?

In studies where scrutineers were judging the quality of scripts against their own internalised grade standards it was common to set a time for scrutineers' interactions with individual scripts or sets of scripts. See, for example, Adams, *et al*. (1990). Each group of scrutineers was presided over by a member of the awarding bodies' research staff who timed each judgement and rang a bell after, for example, ten minutes. This

was the signal for scrutineers to record their judgement and move on to the next script. After initial scepticism, scrutineers were able to keep pace with this regime.

With the introduction of paired comparisons, it was generally conceded that because of the vagueness of how the judgements were to be made, it was impossible to restrict scrutineers to a fixed time, and they had to be allowed to proceed at their own pace. This in turn meant that detailed schedules of pairs of scripts could not be drawn up in advance, because it could never be assumed that a particular script was available when a scrutineer needed it: it might be in use by someone else.

So randomised sequences of pairs of awarding bodies' scripts were prepared and stewards in the rooms where the groups of scrutineers operated were employed to provide scripts from those available at any one time, according to the sequence. Again, this operation was carefully planned, to the extent that dummy sessions were held to make sure that the logistic processes were feasible.

The liberal possibilities of the open-to-all paired comparison methods make all this concern seem irrelevant, if scrutineers up and down the land can switch on their computers and do a few paired comparisons before dinner each evening, though the pairs would be presented in a carefully designed sequence.

3.11 Feedback

One of the spin-off benefits of residential comparability study sessions is that senior examiners of different awarding bodies' examinations congregate and spend a lot of time looking in detail at others' syllabuses and scripts. They often comment that the opportunities to gather these insights makes the whole experience – even doing the judging – worthwhile. Also, social parts of the programme – breaks and meals – give opportunity for a free exchange of views and ideas. It was suggested earlier that a comfortable hotel should be used, so that scrutineers can feel that they are being pampered.

Another way of giving scrutineers a degree of ownership of the study, so that they have a stake in its success, is to give as much feedback as possible during the event. For instance, partial results can be compiled and presented to scrutineers in a plenary session. Their views on the design and conduct of the study should be sought, perhaps formally by means of a questionnaire, and received attentively.

Sometimes this process has gone as far as to ask scrutineers exactly how they made the judgements that they were called upon to make. Given that no instructions are given on this in detail, and the point is made that the judgements should be snap judgements based on a sort of instinct, this seems a bit illegitimate. For one thing, you may not like what they say: 'I always judge against anyone who can't spell 'receive'; and for another, it may be that respondents will feel obliged to say something, and will dream up some principle just for the sake of having an answer. In no case should these responses be included in the final report of the study or used

in the preparation of scrutineers for another study, if it is intended to stick to the original principle that the judging of pairs of scripts should be instinctive.

We have seen, in section 2.2, that in earlier days, agreeing exactly how to judge scripts that arose from what might have been markedly different syllabuses formed a large part of the actual studies. The convergence of syllabuses has made this phase of work unnecessary, and has indeed led to the need for agreeing the basis of judgement to wither and die.

Although the basis of judgement no longer has to be overtly agreed, it is a central feature of cross-moderation work about which little is known. It is debatable whether the question should be followed up at all. Great stress is laid on the fact that the judgements are rapid, instinctive opinions, and it may well be that they are best left at that. On the other hand, it can be argued that the more that is known about the whole process the better it will get.

3.12 Scheduling and character of judgements

The sample cross-moderation study programme displayed as Appendix 1 shows an efficient way of organising scripts and scrutineers. They are divided into three groups and each group addresses one grade boundary. Then all change over and a new grade boundary is addressed by each group. In this way, all scripts are in use at any one time and all scrutineers are occupied all the time.

In former times, judgements of scripts – one at a time – were made against standards that had either been articulated in advance, or against the scrutineers' internalised standards that were identical to those mobilised at grade award meetings. Appendix 3 shows the results of part of such a study, Abbott *et al.* (1989), in GCSE English with the judgement of an individual scrutineer of an individual script being shown as '–', '0' or '+'. It is clear from Appendix 3 that the design of the study meant that no scrutineer looked at scripts from his or her own awarding body.

Parallel to the evolution of cross-moderation methods, and strongly influencing that evolution, has been the convergence of syllabuses. At one time, teachers in the UK enjoyed a great deal of freedom in designing and choosing their syllabuses. The introduction of examinations based on the National Curriculum in 1994, with its centrally determined programmes of study in a range of subjects, forced Key Stage 4 (GCSE) syllabuses into a common, or at least more common, mould. Similar developments in the regulation of GCE syllabuses have had a similar effect on AS and A level syllabuses.

While adherents of plurality might find this regrettable, organisers of comparability studies can be quietly grateful that one source of variation among syllabuses – that of content and treatment – has been largely removed. Similarly, the extent of candidates' freedom to choose among questions has been substantially reduced.

This is not to say that syllabuses are identical, simply a lot more similar than they were in the 1970s. Syllabus variety indeed gives rise to a fundamental question about comparability in general and the judging of scripts in particular.

The point is made by reference to a specific if slightly hypothetical example: suppose two awarding bodies each offer a syllabus in GCSE history. (The details are a little unrealistic to make the point clearly.) Syllabus I is a complete sub-set of Syllabus II, that is, everything that is in Syllabus I is in Syllabus II, but Syllabus II has some material that is all its own.

In one year's examination, an identical question appears on the French Revolution (which is, it goes without saying, part of the common material). Is this question easier, harder or of the same difficulty in the two examinations? Convincing answers can be constructed for all three possibilities.

1. It is harder for Syllabus II candidates, because the candidates must choose from a wider bank of knowledge to answer it for the Syllabus II examination.

2. It is easier for Syllabus II candidates, because having done more history, candidates will be better at the subject, so Syllabus II candidates will benefit from this wider experience.

3. It is of the same difficulty, because a question is a question: if you meet the success criteria then this is absolute and cannot depend on anything else.

(This, incidentally is similar to the question that faced the examining boards in 1986/7 when the first Advanced Supplementary – forerunners of the current Advanced Subsidiary – examinations were introduced. The political stance was that AS examinations were of the same standard as A levels in the subject but had half the volume. The practical question was if an AS examination comprises half the components of its corresponding A level examination, should the same component grade boundaries be used for the two awards? Significantly, the collective wisdom of the boards' research officers couldn't agree on the answer to this. It was referred to their superiors, the then GCE Secretaries, whose collective wisdom also couldn't reach agreement. The issue was settled by the then regulators, who ruled that the same grade boundaries should be used, though on what basis this conclusion was reached is not recorded.)

Here's the same question in another context: in two religious studies syllabuses, one embraces three world religions, the other four. One year, candidates are asked to explore wedding rites in three world religions. Is this easier for the first group or the second? The same convincing answers can be adduced.

1. It must be easier to answer if you've only done three religions, because you won't get the details of any of them confused with the fourth extraneous one.

2. It must be easier if you've done four religions, because you can choose the three you know best, and, moreover, by studying four you'll have greater all-round religious know-how.

3. The question is the same, therefore the answers will be directly comparable.

This matter has a direct bearing on how scripts are to be judged, yet no-one is absolutely clear as to what that bearing may be. And this is on top of the difficulty of comparing the merit of an easy task done well and a more difficult task done moderately.

These cases epitomise this central conundrum about comparability and indeed, for that matter, awarding: the judgement of tasks arising from different contexts. Experience suggests that there is no simple answer, and answers may be different for different subjects and for different candidates. In practice, it has had to be tacitly assumed that the same answer to the same question is of equal worth, whatever the rest of the syllabus or examination may look like.

3.13 Preparatory work

In spite of, or perhaps because of this central conundrum, it is to be regarded as good practice in preparing scrutineers for the cross-moderation study by making sure that they are familiar with the syllabuses and examination materials that they will be scrutinising. In the early years, when the basis upon which comparisons could be made was a serious preliminary, this work arose naturally. In latter years, with the convergence of syllabuses that has already been described, this was formalised into a phase of comparability that was concerned with the judgement of cognitive demand of syllabuses and examination materials. This is covered in detail in Chapter 5 of this book.

There is an optimistic school of thought (see, for example, Gray, 1997) that holds that this preparation is not only worthwhile in its own right, but also prepares scrutineers to 'lay off' their judgements in the light of what they have seen of the relative demands of the syllabuses. Sceptical colleagues find this degree of mental gymnastic nimbleness difficult to imagine. In some ways, the two attitudes are restatements of the two sides of the central conundrum: the importance of context in judging standards.

So the central conundrum remains unresolved. Comparability methods are after all quite approximate and crude, and it is unwise to expect too much of them by way of precise findings.

4 Analysis of findings

Until the paired comparison revolution, the data generated by cross-moderation studies tended to be simple enumerations of a limited number of judgements (see, for example, Kingdon *et al*. 1984). There was a clear tendency to avoid sophisticated statistical analysis because of the unsophisticated nature of the devices that produced the data, that is, extremely tenuous human judgement.

With the first round of similar-syllabus studies, a simple means of scoring was devised. Scrutineers were all experienced Chief Examiners-and-awarders who were selected because they had the notional grade standards internalised. They were then to mobilise their minds to judge scripts carefully selected to be *just* grade C, say. The question posed over each script was quite simple: are you surprised to see this script here posing as a just grade C script? If you are not surprised record a zero. If you *are* surprised, record a '+' if it's too good to be masquerading as a low grade C; or record a '–' if you think it too poor to be even a low grade C. Analysis of these data was then mainly enumeration. An example of the results of this activity has been shown in Appendix 3.

The principle of the analysis was simple: if the scripts from one awarding body all got a lot of + scores, then it was concluded that there was an effect, one explanation of which could be a difference in the standard of grade C for that awarding body. It's difficult to identify any other possible explanation for such an effect, but all sorts of possibilities could be envisaged concerning the outward aspect of the scripts from different awarding bodies.

Incidentally, part of the design of these studies, carried over into paired comparison ones, is that in recent years' studies, no scrutineer looks at scripts – either singly or for a paired comparison – from the awarding body that he or she 'represents'. There is a 'home and away' effect that runs as a consistent seam through all this work. It was observed in several early studies and is discussed in Forrest & Shoesmith (1985, p. 35). This fact alone might serve to remind practitioners of the tenuous nature of the data that studies yield.

A number of statistical techniques were used to analyse the data arising from these studies. Simple χ^2 tests of the frequencies of judgements were mobilised to detect 'significant' differences among syllabuses; the '+' and '–' judgements were turned into 1 and –1 values and a one-way analysis of variance conducted on the results. Kolmogorov-Smirnoff tests of similarity of distributions were used. All were informally criticised as being too elaborate given the nature of the data, and for requiring assumptions to be made that clearly didn't hold. Adams (1995) did, however, show by simulation that the test statistic arising from the analysis of variance did have the predicted F distribution, which lent some legitimacy to the use of that technique.

An interesting sideline of these studies is to look at how ready scrutineers are to stick their necks out and make '+' or '–' judgements, rather than playing safe and recording '0'. Extremes are to be found in a GCSE science study (Cron & Houston, 1990) where one scrutineer was so certain of his or her fine-tuned sense of standards that he or she was able to give a '+' or '–' to every single script; another was so uncertain that he or she recorded '0' for every script. This generated a brief spurt of interest in the 'width of zero' question, though it is nowhere formally aired. The 1989 study in GCSE English literature (Fearnley, 1990) used a five-point scale for judging script quality but no mention in the report is made explicitly about the 'width of zero'.

An important point might be made here that applies to all statistical analyses of these sorts of data, no matter how simple. Statistical procedures are designed to find effects, and find effects they will. It is as well to bear in mind that in a typical cross-moderation exercise, there are countless effects *not* identified. It is inevitable, though, that those that are will be seized upon and interpreted. If classical significance tests are used, then, at the 5% level of significance, by definition, about 5% of comparisons will yield an effect, *even where none is present*: that is precisely what 5% significance means.

The paired comparison revolution came with a built-in analytical method, and a sophisticated statistical model of the data. The Rasch model could be used to place all scripts on a single scale. The technique was first used in a suite of GCSE studies published in 1997 but based on the 1996 examinations. A useful summary is given in Adams (2000). Chapter 7 deals with this approach fully.

A useful by-product of the Rasch modelling of paired comparisons is that the patterns of judgements of individual scrutineers can be analysed to detect any who might be eccentric, that is, give judgements that seem to conflict with those of the majority of judges more often than mere chance would suggest.

Over the years a number of investigators have tried to examine the stability or replicability of statistical findings of these analyses. For instance (Adams, 1995; 1999) simulated large numbers of data sets to see if statistical predictions and distributions were realised. As suggested above, these analyses partly legitimised the use of analysis of variance techniques in these contexts, and also the use of χ^2 tests of the frequencies of judgements by an awarding body.

In Jones & Meadows (2005), as we have already seen, an attempt was made to replicate a study of GCSE religious studies. Identical scripts were used in an identical programme in the same hotel but using different scrutineers. The results were encouraging in that the main profile of script parameters was roughly reproduced. Certainly, on the basis of that study, it could never be claimed, as some feared, that the outcomes of cross-moderation studies were entirely random!

5 Reconciling different strands

In the previous section, some attention was given to the preparatory work that scrutineers undertook, ostensibly to familiarise themselves with the various syllabuses, question papers and mark schemes whose scripts would eventually be judged. This has become formalised into a separate form of comparability study, described in Chapter 5.

It is tempting to relate the two strands, and to ask if any systematic effect emerges in the grade standards of the awarding bodies corresponding to whether their syllabuses are seen as more or less demanding. This does have some interest, though a neat instance of complete correspondence has never been found. A problem is that awarding intervenes between the demand of the syllabus and the grade standard. Grades can be relatively easy to get in a very hard syllabus, if the grade boundaries

are set low enough. Likewise a grade A in a 'less demanding' GCSE syllabus can be made very difficult to attain if the grade boundary is set high enough. That, after all, is what grade awarding is all about.

Two strands that might be thought to show effects, if any, are present in the cross-moderation strand and the statistical strand, that uses multilevel statistical models to account for as much variation between schools and pupils as possible. These models are fully discussed in Chapter 10, which also addresses the question of consistency of findings.

It is disappointing to find that such approaches, asking the same question but approaching it in entirely different ways, tend to yield different if not contradictory conclusions. For instance, in While & Fowles (2000) substantially contradictory results were found by the cross-moderation and statistical modelling strands.

6 Conclusion

The story of cross-moderation comparability work goes right back to the 1950s. This was the beginning of the period when diversity, innovation and experiment were the hallmarks of the UK education systems and the examination bodies that served them. It is to the then GCE examining boards that the credit must go for taking comparability so seriously. It is the present author's contention that few research findings have made the examination administrator's life easier; generally they have made it more difficult. And yet the boards and their successor organisations have continued to spend a lot of money to fund this work: comparability in particular but examinations research in general.

An earlier section has described the evolution of the methods used and the factors that weighed upon them, but now the current state of evolution is that we have arrived at a sophisticated method of carrying out script-based comparability studies, with a correspondingly sophisticated method of analysis.

The vagaries of the notion of standards, though, always leaves a slight doubt over the meaning of any experimental findings. As yet, for instance, there is no answer to the central conundrum posed earlier in this chapter. This explains the reluctance of the awarding bodies to take decisive action based upon the results of comparability studies. The code of practice for examinations (QCA, 2006) stipulates that awarding committees should have any pertinent comparability study reports available, but no advice is adduced concerning how they might be used.

It is unlikely that a definitive answer can ever be found: the arguments are so well trodden, and the ground so often gone over, that if an answer were discoverable, it would have been elicited by now. But nonetheless, promising lines of development are emerging. The paired comparison methods are still in their infancy in the realm of awarding and comparability, and the potential for progress seems to be huge. An exciting aspect of these developments is the fact that it opens the ranks of potential scrutineers to many sorts of interested persons hitherto excluded.

Endnotes

1 The term appropriate for the historical context will be used where it makes sense. Similarly, the word 'syllabus' is used throughout this chapter, despite the recent practice of calling them 'specifications'.

2 The Office for Standards in Education, responsible for inspections of schools and colleges in England.

3 A tier of an examination is a scheme of assessment that is aimed at a range of grades rather than at all grades. Foundation tier GCSE typically gives candidates access to grades C to G; higher tier gives access to A* to E.

References

Abbott, M.K., McLone, R.R., & Patrick, H. (1989). *GCSE inter-group comparability study 1988: English*. Organised by the Midland Examining Group on behalf of the Inter-Group Research Committee for the GCSE and the Joint Council for the GCSE.

Adams, R.M. (1994, April). *Standards over time*. Paper presented at the Standing Research Advisory Committee of the GCE boards symposium on year-on-year standards, Meriden, Warwickshire.

Adams, R.M. (1995, October). *Analysing the results of cross-moderation studies*. Paper presented at a seminar on comparability, held jointly by the Standing Research Advisory Committee of the GCE boards and the Inter-Group Research Committee of the GCSE groups, London.

Adams, R.M. (1999, November). *The Rasch model and paired comparisons data: Some observations*. Paper presented at a seminar held by the Research Committee of the Joint Council for General Qualifications, Manchester. Reported in B.E. Jones (Ed.), (2000), *A review of the methodologies of recent comparability studies*. Report on a seminar for boards' staff hosted by the Assessment and Qualifications Alliance, Manchester.

Adams, R.M. (2000). *Comparability studies in GCSE English, mathematics and science 1998: A summary*. Unpublished report for the Research Committee of the Joint Council for General Qualifications, Cardiff, Welsh Joint Education Committee.

Adams, R.M., Phillips, E.J., & Walker, N.A. (1990). *GCSE inter-group comparability study 1989: Music*. Organised by the Welsh Joint Education Committee and the Northern Ireland Schools Examinations Council on behalf of the Inter-Group Research Committee for the GCSE and the Joint Council for the GCSE.

Bardell, G.S. (1977). *Report of the inter-board cross-moderation study in 1976 Advanced level pure mathematics*. Cardiff: Welsh Joint Education Committee.

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.

Bell, J.F., Bramley, T., & Raikes, N. (1998). Investigating A level mathematics standards over time. *British Journal of Curriculum and Assessment, 8*(2), 7–11.

Bell, J.F., & Greatorex, J. (2000). *A review of research into levels, profiles and comparability*. London: Qualifications and Curriculum Authority.

Christie, T., & Forrest, G.M. (1980). *Standards at GCE A-level: 1963 and 1973.* Schools Council Research Studies. London: Macmillan Education.

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.

Cron, N., & Houston, J. (1990). *GCSE inter-group comparability study 1989: Chemistry*. Organised by the Southern Examining Group on behalf of the Inter-Group Research Committee for the GCSE and the Joint Council for the GCSE.

D'Arcy, J. (Ed.). (1997). *Comparability studies between modular and non-modular syllabuses in GCE Advanced level biology, English literature and mathematics in the 1996 summer examinations*. Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.

Edwards, E., & Adams, R. (2003). *A comparability study in GCE Advanced level geography including the Scottish Advanced Higher grade examination. A study based on the summer 2002 examination.* Organised by the Welsh Joint Education Committee on behalf of the Joint Council for General Qualifications

Elliott, G., & Greatorex, J. (2002). A fair comparison? The evolution of methods of comparability in national assessment. *Educational Studies*, *28,* 253–264.

Fearnley, A.J. (1990). *General Certificate of Secondary Education. A comparability study in English literature. A study based on the work of candidates in the summer 1989 examinations*. Organised by the Northern Examining Association on behalf of the Inter-Group Research Committee for the GCSE and the Joint Council for the GCSE.

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.

Forrest, G.M., & Williams, C.A. (1983). *Report on the inter-board study in physics (Ordinary) 1980*. Manchester: Joint Matriculation Board.

Fowles, D. (1989). *GCSE inter-group comparability study 1988: French*. Organised by the Northern Examining Association on behalf of the Inter-Group Research Committee for the GCSE and the Joint Forum for the GCSE.

Francis, J.C., & Lloyd, J.G. (1979). *Report on the inter-board cross-moderation study in history at Advanced level 1979*. Aldershot: The Associated Examining Board.

Good, F.J., & Cresswell, M.J. (1988). *Grading the GCSE*. London: Secondary Examinations Council.

Gray, E., (1997). *A comparability study in A level biology*. Organised by the Oxford and Cambridge Examinations and Assessment Council on behalf of the Joint Forum for the GCSE and GCE.

Greatorex, J., Hamnett, L., & Bell, J.F. (2003). *A comparability study in GCE A level chemistry including the Scottish Advanced Higher grade. A review of the examination requirements and a report on the cross-moderation exercise. A study based on the summer 2002 examinations.* Organised by The Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for Oxford Cambridge and RSA Examinations on behalf of the Joint Council for General Qualifications.

Guthrie, K. (2003). *A comparability study in GCE business studies, units 4, 5 and 6 VCE business, units 4, 5 and 6. A review of the examination requirements and a report on the cross-moderation exercise. A study based on the summer 2002 examination.* Organised by Edexcel on behalf of the Joint Council for General Qualifications.

Joint Matriculation Board & University of London University Entrance and School Examinations Council. (1966). *O-level Latin, French and biology (1964)*. Occasional Publication 24. Manchester: Joint Matriculation Board.

Johnson, S., & Cohen, L. (1983). *Investigating grade comparability through cross-moderation*. London: Schools Council.

Jones, B.E. (1993). *GCSE inter-group cross-moderation studies 1992. Summary report on studies undertaken on the summer 1992 examinations in English, mathematics and science.* Inter-Group Research Committee for the GCSE.

Jones, B.E., & Meadows, M. (2005). *A replicated comparability study in GCSE religious studies.* Manchester: Assessment and Qualifications Alliance.

Jones, M.J., & Lotwick, W.R. (1979). *Report of the inter-board cross-moderation exercise in biology at the Ordinary level, 1978*. Cardiff: Welsh Joint Education Committee.

Kingdon, J.M., Wilmut, J., Davidson, J., & Atkins, S.B. (1984). *Report of the inter-board comparability study of grading standards in Advanced level English*. London: University of London School Examinations Board on behalf of the GCE Examining Boards.

Massey, A.J. (1979). *Comparing standards in English language: A report of the cross-moderation study based on the 1978 Ordinary level examinations of the nine GCE boards*. Bristol: Southern Universities' Joint Board and Test Development and Research Unit.

Patrick, H., & McLone, R.R. (1990). *GCSE inter-group comparability study 1989: CDT Technology*. Organised by the Midland Examining Group on behalf of the Inter-Group Research Committee and the Joint Forum for the GCSE.

Pollitt, A., & Elliott, G. (2003). *Monitoring and investigating comparability: A proper role for human judgement*. Research and Evaluation Division, University of Cambridge Local Examinations Syndicate.

Qualifications and Curriculum Authority. (2006). *GCSE, GCE, GNVQ and AEA code of practice, 2006/7*. London: Qualifications and Curriculum Authority.

Quinlan, M. (1995). *A comparability study in Advanced level mathematics. A study based on the summer 1994 and 1989 examinations*. University of London Examinations and Assessment Council on behalf of the Standing Research Advisory Committee of the GCE Boards.

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.

School Curriculum and Assessment Authority. (1995). *Report of a comparability exercise into GCE and GNVQ business*. London: School Curriculum and Assessment 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.

Secretaries of the Examining Bodies. (1952). *Minutes of a meeting held in Bristol, March 17th and 18th 1952.* Unpublished minutes, in Secretaries of Examining Boards 1948–1960. Cambridge Assessment Archive, PP/TSW 3/5.

University of Cambridge Local Examinations Syndicate. (1981). *Report of an inter-board cross-moderation exercise in geography at Advanced level in 1978*. Cambridge: University of Cambridge Local Examinations Syndicate and Test Development and Research Unit.

While, D., & Fowles, D. (2000). *A comparability study in GCSE mathematics. Statistical analysis of results by board. A study based on the work of candidates in the summer 1998 examinations.* Organised by the Assessment and Qualifications Alliance (Northern Examinations and Assessment Board) on behalf of the Joint Forum for the GCSE and GCE.

Wood, A. (2006). *What makes GCE A level mathematics questions difficult? The development of the preferred alternative construct elicitation (PACE) methodology for enabling students to make and give reasons for demand judgements: The findings from a pilot study and an outline programme of work arising from the pilot study*. Unpublished Masters in Research Methods dissertation. University of London Institute of Education.

Appendix 1 Programme for a cross-moderation residential meeting

Welsh Joint Education Committee, Cyd-Bwyllgor Addysg Cymru

GCE A level geography 2002

Comparability Study, Residential Meeting

Holiday Inn Hotel, Cardiff City Centre, 28–29 November 2002

Programme

Wednesday 27 November (early evening)

Participants arrive, check in to rooms and meet for dinner at 8.00pm

Thursday 28 November

9.00am	Welcome, Introduction and Induction (Brecon Two – Boardroom)
9.45am	First Working Session – Group 1 (9) (A Boundary) – Syndicate One Group 2 (9) (E Boundary) – Rhossili Suite
10.45am	Coffee
11.00am	Second Working Session
12.00 noon	Break
12.15pm	Third Working Session
1.15pm	Lunch
2.15pm	Fourth Working Session
3.15pm	Tea
3.30pm	Fifth Working Session
4.30pm	Break
4.45pm	Sixth Working Session – Group 1 (E Boundary) – Rhossili Suite Group 2 (A Boundary) – Syndicate One
5.45pm	Close of first day
8.00pm	Dinner

Friday 29 November

8.45am	Seventh Working Session – Group 1 (E Boundary) – Rhossili Suite Group 2 (A Boundary) – Syndicate One
9.45am	Break
10.00am	Eighth Working Session
11.00am	Coffee
11.15am	Ninth Working Session
12.15pm	Break
12.30pm	Tenth Working Session
1.30pm	Lunch
2.15pm	Plenary Session (Brecon Two – Boardroom)
3.00pm	Departure

Appendix 2 Scrutineer's record card. Showing successive pairs of awarding bodies' scripts to be compared

A=AQA C=NICCEA E=EDEXCEL O= OCR S=SQA W=WJEC

A/B boundary

Scrutineer W1

No	B1	#	B2	#
121	C		E	
122	A		S	
123	C		O	
124	E		S	
125	A		O	
126	C		S	
127	E		O	
128	A		C	
129	O		S	
130	A		E	
131	C		O	
132	A		E	
133	C		S	
134	E		O	
135	A		S	
136	C		E	
137	A		O	
138	E		S	
139	A		C	
140	O		S	
141	C		E	
142	A		S	
143	C		O	
144	E		S	
145	A		O	
146	C		S	
147	E		O	
148	A		C	
149	O		S	
150	A		E	
151	C		O	
152	A		E	
153	C		S	
154	E		O	
155	A		S	
156	C		E	
157	A		O	
158	E		S	
159	A		C	
160	O		S	

No	B1	#	B2	#
161	C		E	
162	A		S	
163	C		O	
164	E		S	
165	A		O	
166	C		S	
167	E		O	
168	A		C	
169	O		S	
170	A		E	
171	C		O	
172	A		E	
173	C		S	
174	E		O	
175	A		S	
176	C		E	
177	A		O	
178	E		S	
179	A		C	
180	O		S	
181	C		E	
182	A		S	
183	C		O	
184	E		S	
185	A		O	
186	C		S	
187	E		O	
188	A		C	
189	O		S	
190	A		E	
191	C		O	
192	A		E	
193	C		S	
194	E		O	
195	A		S	
196	C		E	
197	A		O	
198	E		S	
199	A		C	
200	O		S	

No	B1	#	B2	#
201	C		E	
202	A		S	
203	C		O	
204	E		S	
205	A		O	
206	C		S	
207	E		O	
208	A		C	
209	O		S	
210	A		E	
211	C		O	
212	A		E	
213	C		S	
214	E		O	
215	A		S	
216	C		E	
217	A		O	
218	E		S	
219	A		C	
220	O		S	
221	C		E	
222	A		S	
223	C		O	
224	E		S	
225	A		O	
226	C		S	
227	E		O	
228	A		C	
229	O		S	
230	A		E	
231	C		O	
232	A		E	
233	C		S	
234	E		O	
235	A		S	
236	C		E	
237	A		O	
238	E		S	
239	A		C	
240	O		S	

Appendix 3 Results from part of a cross-moderation study in GCSE English

Data matrix for the C/D boundary
4 scrutinising teams (I–IV), 18 scrutineers (a–r), 6 boards and syllabuses (A–F), 20 scripts per syllabus
+ above borderline, 0 on borderline, – below borderline

Team	S	A	B	C	D	E	F
		12345678911111111112 01234567890	12345678911111111112 01234567890	12345678911111111112 01234567890	12345678911111111112 01234567890	12345678911111111112 01234567890	12345678911111111112 01234567890
I	g	–0–+0	000—		+0++–	–+0–0	0+0+0
	j	0—+0	–++0+	+——0		——0–	0000–
	m	0–0+0	0+0++	+0—0	++0+0		00+–0
	p	0–00–	0+——	+0000	+00+–	+000–	
	b		–+000	+–0+0	+0++–	+–0–0	0+0—
II	e	00—0		+000+	0+00	00+00	00+00
	h	0–0–	–0+0–		–++0–	++0—	–+0–0
	k	—00–	0+——	0–0+0		+———	–00–+
	n	0———	–00—	++0++	0++00		–00–+
	q	00–0–	–0—0	+00++	0+++0	00+00	
III	c		+–00–	00000	0—+0	+0000	000+–
	f	–000+		+0+00	000+0	+0000	–000–
	i	000–0	+00—		0—++	+0——	——0–
	l	—00–	+00—	–0+0–		——00	–+00–
IV	o	0——0	–0–00	++–00	–000–		0+++0
	r	0——0	——0–	00——	–0000	—00–	
	a		0–000	0+–00	–00++	000+0	000+0
	d	+0–00		+00–0	0–00+	—000	0+++0