Inter-Subject Comparability: A Review of the Technical Literature

ISC Working Paper 2

December 2015

Ofqual/15/5794

Contents

Suggested citation:

Ofqual (2015b) *Inter-Subject Comparability: A Review of the Technical Literature: ISC Working Paper 2*. Coventry, the Office of Qualifications and Examinations Regulation.

Executive summary

Comparability is the degree to which results from separate exams embody the same standard. When a high level of comparability has been achieved between two exams, they are said to be equally difficult, and their results can be used interchangeably. The more similar the constructs[1] measured by two exams, the greater the potential to apply the same standard, and the greater the expectation that a high level of comparability should be achieved. The more dissimilar the constructs measured by two exams, the harder it becomes to make sense of the idea of applying the same standard, let alone achieving a high level of comparability. Inter-subject comparability, the focus for the present report, is one of the most complex and contentious forms of comparability there is.

In our 2013–16 *Corporate Plan*, we committed to consider with experts the available data, information and evidence on comparability between subjects. The present report is a review of the technical literature on inter-subject comparability. It reviews both outcomes from research and controversies over how these outcomes might be interpreted.

Discussion of inter-subject comparability has been both long-running and highly contentious, with published research dating back to 1928. Over the years, exam boards, qualifications regulators and other experts have made various attempts to investigate and address comparability between subjects, using both statistical and judgemental methods. However, no approach has gained widespread support. Much of this work was carried out between the 1970s and mid-1990s. In recent years, the literature has taken a more philosophical turn. Properly conceptualising inter-subject comparability is now seen as a critical precursor to any attempt to monitor or measure it.

There have been many recent attempts to explore alternative conceptions of comparability. For some commentators, comparability should be defined in terms of the level of performance or attainment demonstrated by a student in an exam; for others, it should be defined as the likelihood of a student gaining a certain grade in an exam. Others maintain that two qualifications can only be compared if they share a common 'linking construct'. Some authors believe that we should attempt to reach consensus over the most appropriate definition of inter-subject comparability; others believe that any of the definitions might be quite legitimate depending on the context of the comparison. Others warn against making such extreme comparisons at all,

[1] The construct of an exam is the characteristic, or attribute, which the exam is designed to measure, for example attainment in biology.

arguing that conceptions of standards are inherently subject-specific; that is, where different A level and GCSE exams are intended to assess different knowledge, skills and understanding, this limits the comparisons that can, and should, ever be made between them.

As many conceptions of comparability as there are, there are even more methods for investigating it. Broadly, though, methods fall into one of two camps – judgemental and statistical. Judgemental approaches require experts to compare the content, assessment or performance demand of a qualification. They allow us to compare the level of student attainment required to gain a certain grade. This attainment-based comparison reflects the 'traditional' view of success in exams, and the common-sense conception of comparability. In practice, judgemental approaches to inter-subject comparability are limited by the shortage of experts who might be considered suitably qualified to make valid comparisons between disparate subjects; if, indeed, any judge could be considered suitably qualified to make such complex and multi-dimensional judgements between attainments in qualitatively different subject areas.

Statistical approaches use data to compare the relative likelihood of students achieving a grade in certain subjects. Many writers believe that statistical approaches to inter-subject comparability are equally flawed, for methodological reasons related to the broad assumptions that underpin them. Perhaps the most heavily debated of these assumptions is the notion that there is a common underlying dimension of ability (howsoever that 'ability' might be defined) which would allow us to compare meaningfully subjects as diverse as chemistry, French, art and history. Statistical approaches, based on cohort-level data, also mask significant differences in apparent difficulty for particular sub-groups of students.

Using a purely statistical approach, studies of subject difficulty at A level and GCSE often show what appear to be consistent differences between subjects; with language and science subjects the most 'difficult' for students to succeed in, particularly at A level. This pattern is one of the few things that not disputed in the comparability literature. However, what such patterns mean, and what, if anything, should be done about them, is quite another thing. For some experts, it is unsurprising that such consistent patterns emerge where the studies involved are all underpinned by the same assumptions. Others believe that such consistent patterns must reflect a genuine difference in subject difficulty – one that simply cannot be ignored because of methodological limitations. We should note that these patterns are not unique to England, with the same trend experienced in a number of other countries around the world. If these patterns really are 'grading errors' then why have so many countries fallen into the same trap as England?

The existing research does not give us a clear steer on whether certain subjects are actually more difficult than others. This is not due to any shortage of empirical data, but rather to a lack of agreement on how to interpret those data. Despite all the work

reviewed, both practical and conceptual, the ‘facts’ of the matter, namely whether or not specific subjects can justifiably be said to be graded more harshly than others, are still far from clear. Although thinking has moved on considerably over time, there is still substantial disagreement about how to define and conceptualise inter-subject comparability, let alone how to go about measuring it.

1. Introduction

Comparability is the degree to which results from separate exams embody the same standard. In theory, this sounds like a fairly straightforward matter. In practice, it is an extremely complex area and can encompass a multitude of definitions, methodologies and contexts (Elliott, 2013).

There are different forms of comparability: we might wish to compare the standard of qualifications in the same subject area, that is from year to year, or between exam boards during the same year; or, as is the focus of the present report, we might wish to compare the standard of qualifications in different subject areas. In England, there are few more controversial issues in education than ensuring the comparability of exams, leading Nuttall (1979) to describe it as the "English disease" (p. 12). Exam boards have always accepted the requirement to demonstrate comparability of qualifications over time and between exam boards. However, the requirement to demonstrate comparability of standards between subjects has remained far more controversial. It is this issue of inter-subject comparability that we focus on in this review of the technical literature.

The concept of inter-subject comparability is highly contentious. It has been strenuously debated by assessment experts over many decades and has been subjected to intense public and political scrutiny. The origins of these concerns have varied, but, at their simplest, they derive from differences in the proportions of students succeeding in the different subjects (QCA, 2008). For all inter-subject comparability's prominence in England, we should note that concerns over inter-subject standards are not uniquely English. The English system is characterised by the wide range of subjects available for study, and the unusually high level of student choice in selecting between subjects, particularly at A level. These circumstances make inter-subject comparability a particular challenge. However, where student choice is a significant feature of the education system elsewhere, similar debates have taken place around the world, although rarely to the same extent (Korobko et al., 2008; Lamprianou, 2009).

Various commentators in England have promoted the extreme position that it is meaningless to relate standards in very different subjects to one another (see Nuttall, 1979; Goldstein and Cresswell, 1996; Coe, 2010). Yet, others have promoted a very different position: that inter-subject comparability is an essential requirement of a system that awards the same grades across subjects, and which, therefore, encourages qualification users to assume that results have a 'common currency' across subjects and can be treated as equivalent. If so, it has been said, then we should try to take all reasonable steps to ensure inter-subject comparability (Coe, 2010). If not, then we make it very clear that grades represent subject-specific attainments and, therefore, cannot be considered equivalent (Nuttall, 1979).

This review of the technical literature aims to distil the highly complex debate about inter-subject comparability into a relatively short yet comprehensive summary. The report is just one contribution to the body of work that will inform our position on inter-subject comparability. It brings together research from key contributors to the literature from England. It focuses predominantly on literature published in the last 20 years, but also touches on some of the earlier debates, as well as drawing on some of the unpublished 'grey literature' from within exam boards. Although this is an international concern, we have based the review primarily on the literature from England, acknowledging that it already incorporates insights from the wider international literature.

1.1 Report structure

In section 2, we examine the historical context of the technical literature on inter-subject standards, before moving on to a discussion of comparability concepts (section 3). In section 4, we look at the two main methods used to investigate comparability, and in section 5 we consider findings from the research. Section 6 concludes the report.

2. Background

The public exams system in England has long been concerned with comparability between subjects. Published research on this issue dates back to at least 1928 (for example, Crofts and Caradog Jones), and has at times proved to be highly controversial (for example, Fitz-Gibbon and Vincent, 1994, vs. Goldstein and Cresswell, 1996). Studies have often built on one another, and the debate has developed significantly over time. Therefore, before we discuss the concepts and methods involved in comparing standards between subjects, we put this discussion into its historical context.

Over the years, exam boards, qualifications regulators and other experts have made various attempts to investigate inter-subject comparability, both quantitatively and qualitatively. Neither investigative approach has gained widespread support. For methodological reasons, related to the broad assumptions that underpin them, statistical approaches to inter-subject comparability have been particularly controversial. This has prompted debate lasting many decades (see Newton, 2012, for a historical overview).

During the early 1970s, Britain witnessed a rising concern with the comparability of standards between subjects in public exams. This led to the development of various statistical techniques for analysing 'the problem' (Forrest, 1971; Nuttall et al., 1974; Kelly, 1975). Although many questions were posed concerning their validity (for example, Nuttall et al., 1974; Kelly, 1976), statistical techniques became widely used, both to investigate inter-subject standards and influence subsequent grading decisions (Forrest and Vickerman, 1982; Jones, 2003, 2004).

The first major study of inter-subject standards using statistical techniques was carried out by Nuttall et al. in 1974. This considered relative subject 'difficulty' by using five different statistical methods, including subject pairs analysis. Results revealed a consistent pattern of subject difficulty across methods and also across exam boards, with sciences and languages apparently harder than other subjects. We explore notions of 'difficulty' or 'grading severity' in more detail later. However, for the studies in this section, we might describe 'difficulty' loosely as the likelihood of a student achieving a lower grade in one subject in comparison with others. In their heavily hedged interpretation of results, Nuttall et al. emphasised that they were presenting issues requiring further dialogue and their findings should not be viewed as conclusive.

Kelly (1975, 1976), working in Scotland, built on the subject pairs methodology to develop her own approach for monitoring inter-subject comparability. She found a similar pattern of difficulty between subjects, but also observed a disparity in apparent subject difficulty experienced by boys and girls. She noted that use of any 'correction factor' to adjust results would, therefore, be challenging. To put it simply,

the idea of raising or lowering a subject's grade boundaries for boys but not for girls would be ethically questionable, to say the least.

During the late 1970s to mid-1980s there were several internal, unpublished studies into inter-subject comparability. For example, researchers from the Joint Matriculation Board (JMB) used subject pairs analysis to monitor the inter-subject comparability of their O level and A level exams. Forrest and Vickerman (1982) reported results from 1972 to 1980: the tendency was for languages, and chemistry, physics and mathematics to appear 'harder' than other subjects, albeit with some disparity.

Around the same time, one of the earliest studies to question seriously the validity of statistical approaches was published. Newbould (1982) argued that student achievement across subject areas was heavily affected by factors that could not be taken into account through simple statistical techniques, particularly differential motivation. In other words, grades do not, and should not, simply reflect the average 'general academic ability' of students. Any attempt to investigate and 'correct' grades would need to take additional factors, such as differential motivation across subject areas, into account.

The interest in inter-subject comparability died down for a time, before being reawakened in 1994 by Fitz-Gibbon and Vincent. Their study addressed the question 'Are mathematics and science more difficult than other subjects at A level?' using four methods, which included value added and subject pairs approaches. They concluded that mathematics and science A levels were more 'severely' graded. However, like Kelly (1975, 1976), they found noticeable differences in the statistical patterns between boys and girls.

In early 1996, Sir Ron Dearing's review of 16 to 19 qualifications in the UK was published (Dearing, 1996). This included replications of statistical analyses from Fitz-Gibbon and Vincent, comparing A level performances in different subjects. Again, this suggested that certain subjects – primarily sciences, mathematics, economics, history, French, German and general studies – were 'harder' than others. Dearing suggested that where subjects fall well below the average difficulty there might be a levelling-up of demand. This recommendation was later abandoned in the face of strong opposition from the exam boards, based on technical concerns and practical obstacles (Jones, 2004).

Dearing recommended debate on the validity of conclusions from statistical studies, and numerous papers were subsequently published that highlighted limitations. They strongly challenged methods like subject pairs analysis, both conceptually and technically, questioning what equivalence between subjects meant and whether it could realistically be achieved quantitatively (for example, Goldstein and Cresswell, 1996; Newton, 1997; Jones, 2003). Alton and Pearson (1996) modelled the impact of

adjusting A level and GCSE grades based on statistical methods. Their findings led them to advise that a cautionary approach should be taken.

The most recent phase of research has taken a more philosophical slant, suggesting that discussion has been overly simplified in preceding decades and that there might be more than one legitimate answer to questions of inter-subject comparability (for example, Newton, 2010a, 2010b; Coe, 2010). Developing ideas from earlier studies (for example, Christie and Forrest, 1981), researchers began to recognise a range of definitions of comparability, such that criticisers and defenders of statistical methodologies could both be correct, albeit on their own terms (Newton, 2003; Jones, 2003, 2004; Coe et al., 2008). A central idea formed that the way in which inter-subject comparability is conceptualised and defined is crucial to interpreting and addressing outcomes from comparability monitoring research.

In more recent years, the debate on inter-subject comparability has been driven largely by subject groups. Those with an interest in languages (Myers, 2006; Dearing and King, 2007) and mathematics and sciences (Coe et al., 2008; Royal Society, 2008) have been particularly vocal. These organisations have largely relied upon outcomes from statistical approaches, particularly those conducted by the Centre for Evaluation and Monitoring at Durham University.

The background above gives a brief flavour of the historical debate on inter-subject comparability in England. What is notable is that, whilst thinking has moved on considerably over the decades, we still appear to be no closer to reaching any consensus over the 'facts' of the matter, that is whether or not specific subjects can justifiably be said to be graded more harshly than others (Newton, 2012). Frustration as to this lack of consensus was expressed by Dearing and King (2007):

> This needs to be resolved one way or the other by a definitive study, followed by publication of the conclusions, because the present widely held perception in schools, whether right or wrong, is adversely affecting the continued study of languages through to the GCSE (p. 12).

Newton (2012) argued that this lack of consensus was not due to any lack of data on inter-subject standards, or to the need for any more studies; rather it lies with the lack of agreement about how to make sense of the existing data, that is with the lack of consensus over how standards and comparability ought to be defined.

3. Comparability concepts

Although this literature review focuses on inter-subject comparability, this cannot be viewed in isolation from other forms of comparability. Many of the issues involved when comparing subject standards apply to all other forms of comparability. Although inter-subject comparability resembles other forms of comparability, it is comparability at the most complex and abstract end of the spectrum. In comparing very different entities, comparisons are far more extreme and far less straightforward. Both quantitative and qualitative comparisons are very difficult to characterise and to justify.

At the most basic level, comparability appears to be a deceptively straightforward concept. Newton (2007) defined it as the application of the same standard across different exams. Yet, it rarely implies equivalent exam standards or identical features of performance at common grade boundaries. Even within the same subject over time, students will not know exactly the same facts or have mastered exactly the same skills (Nuttall, 1979; Goldstein and Cresswell, 1996).

Potential dimensions of comparison are manifold. They can apply to: the demand of an assessment; the curriculum content and domain coverage; the performance of students; grading standards; or the predictive potential of exam results (Elliott, 2013). These are the 'attributes' on which a comparison may be based, which in turn form part of the definition or conception of comparability. Studies tend to address different attributes in isolation, but comparing two different qualifications can yield very different results depending on which attribute we select. So, if two qualifications are equivalent in terms of content coverage, it does not follow that they are also equivalent in terms of the proportion of students achieving a particular grade.

As already mentioned, there are also many forms of comparability – comparability of exam standards over time, between different exam boards, between optional routes of a specification and between subjects. Historically, we have required exam boards to "maintain standards across specifications, both within and between awarding organisations and from year to year" (Ofqual, 2011, p. 5), but there is currently no explicit requirement to maintain standards across subjects.

In practice, the primary concern of exam boards, Government and the general public has been on the maintenance of exam standards over time and between exam boards (Baird et al., 2000; Jones, 2004; Coe et al., 2008), rather than between subjects. Given that it would not be possible to address inter-subject comparability without threatening other forms of comparability (Coe et al., 2008), the other less controversial forms of comparability have traditionally taken precedence.

In 2004, an independent review by McGaw et al. found that the qualifications regulator and the exam boards were doing a commendable job in their attempts to

maintain qualifications standards in England. They observed that no other exams system was more highly controlled in its attempt to maintain comparability of exams. They concluded that strategies in England for maintaining standards over time "do as well as possible", whilst acknowledging that "no examination system has found an adequate way to determine whether standards are consistent across subjects" (McGaw et al., 2004, p. ii).

For all its initial appearance of simplicity, the comparability literature is highly complicated: the underlying issues interweave with one another to such an extent that it is very hard to disentangle them. Furthermore, one reason why debates about comparability have often been prolonged and inconclusive is that there is little agreement between those involved about which (or whose) concepts should be used as the foundation of comparability practice (Elliott, 2013). The fact that authors use various methods and techniques to conceptualise the topic, different terminology to describe the same notion, and sometimes the same term to describe different things, makes the literature even more complex to understand (Elliott, 2013). Such disagreements are widely acknowledged:

> Coe (2010) claimed that much debate on the comparability of examination standards is at cross-purposes, since protagonists use the same words to mean different things. Within the educational measurement community we have both variants of this problem: the use of the same term to mean different things and the use of different terms to mean the same thing. […] There seem to be almost as many terms as commentators (Newton, 2010a, p. 289).

In particular, terms such as 'standards' and 'difficulty' often have different meanings conferred on them (Bramley, 2005; Baird, 2007; Coe, 2010). Qualification 'standards' might refer to the content studied as part of a course, the demand of questions in an exam or the performance demonstrated by students, that is what they know and can do (see Baird, 2007). Bramley observed that a "common misconception" of the word 'standard' is the percentage of the population meeting or exceeding a certain level of attainment (Bramley, 2005, p. 252).

Recently, authors have suggested that differences in the use of terminology may lie behind some of the apparent disagreements between experts in their interpretation of research outcomes. An example of this is the disagreement between Fitz-Gibbon and Vincent (1994; 1997) and Goldstein and Cresswell (1996), concerning the claim that mathematics and science A levels were too 'difficult'. Coe (2007) suggested that there may have been no real disagreement between the two parties and they were merely using the word 'difficulty' to mean two quite different things.

3.1 Conceptions of inter-subject comparability

Much of the literature on inter-subject comparability in the past 10 to 15 years has acknowledged the conceptual and definitional challenges involved (Newton, 2012). Properly conceptualising inter-subject comparability is now seen as critical to understanding the issues and moving the debate forward. During much of the 20th century, debates over definitions of comparability often seemed to be confused with debates over comparability methods. Some authors now believe that this was a serious impediment to understanding truly the issues involved:

> An issue that has clouded conceptual analysis of comparability in England, perhaps the principal issue, is the failure to distinguish effectively between definitions of comparability and methods for achieving comparability (or methods for monitoring whether comparability has been achieved) (Newton, 2010a, p. 288).

Elliott (2013, p. 3) described 'comparability definition' as the "rationale and purpose behind the comparison", in contrast to 'comparability method', which is the technique for making a comparison. Elliott (2013), Newton (2010a) and Coe (2007) all argued that definitions and methods do not exist in a one-to-one relationship with one another. Whilst it is true that certain definitions of comparability are well-suited to certain approaches – definitions that invoke notions of 'general academic ability', for instance, seem well-suited to statistical methods – there is no simple relationship between conceptions of and approaches to measuring/achieving comparability.

As the importance of properly conceptualising inter-subject comparability has increasingly been recognised, authors have begun to compare and contrast alternative definitions of comparability. Often, this has taken the form of taxonomies for categorising conceptualisations (for example, Cresswell, 1996; Newton, 2003; Newton, 2010a, 2010b; Coe, 2007, 2008 and 2010). Newton (2010a) noted that some of these frameworks had been constructed to illustrate the illegitimacy of all-but-one definition (for example, Cresswell, 1996), whilst others had been constructed to illustrate the legitimacy of many (for example, Newton, 2010b). No attempt to categorise conceptions of comparability has yet become the dominant model (Newton, 2010a).

The paper by Coe (2010) provides a useful illustration of a framework for categorising conceptions of comparability. He identified three distinct notions of comparability within the assessment literature, arguing that most definitions to date have fallen within the first two of these categories – 'performance comparability' and 'statistical comparability'. He identified and then rejected a third notion – 'conventional comparability' – before describing 'construct comparability', his own

preferred conceptualisation. Each definition is discussed below, drawing heavily upon Coe (2007) and Coe (2010).

3.1.1 Performance comparability

According to the 'performance' view of comparability, the 'standard' of a qualification exists in the levels of skills, knowledge and understanding required to achieve it. These skills and this knowledge are subject-specific. The 'difficulty' of the exam is viewed in terms of the subject-specific performance demands it makes on the student. One exam is more difficult than another if it requires students to demonstrate skills, knowledge and understanding at a higher level for the award of the same grade (Coe, 2007).

Coe (2007, 2010) argued that the 'performance comparability' conception is the one that would most often be assumed by the general reader when the word 'difficult' is used, even if this is not openly stated. Newton (2010a) has argued that the use of the word 'performance' is misleading, because we are actually interested in comparing levels of attainment, rather than features of performance, per se. Two students of the same level of attainment might perform quite differently on two questions that test the same area of knowledge – one getting full marks and one getting no marks – simply because one of the questions is much harder than the other, perhaps because of the context in which it is set, the time allowed, or suchlike. So, we are not interested in comparing levels of performance, per se; we are interested in comparing levels of attainment, which will be expressed differently in performances according to the demands of the task that has been set. In fact, Coe (2010) acknowledged that some versions of 'performance comparability' do state that in order to judge it fairly, we must take into account, as far as is possible, the context in which performances are produced (see Baird et al., 2000; Baird, 2007). For example, an apparently identical essay response might be judged rather differently if it had been produced under different time constraints, with different resources, or as part of a modular or linear exam.

Coe (2007) claimed that, from a performance view, only those exams that give rise to the same kind of performances can be compared. Or, at the very least, there must be a substantial set of skills, knowledge or understandings in common. At first glance, this might suggest that it is not possible to view inter-subject comparability in terms of performance. However, Coe (2007) argued that we might be able to compare certain subject clusters, if a plausible set of common criteria could be identified. Defining these criteria would, of course, be extremely challenging in the context of inter-subject comparability, where such different entities are being compared.

3.1.2 Statistical comparability

Coe (2007) described a second definition, which he termed 'statistical comparability'. Many of the high-profile inter-subject comparability studies from the 1970s to the

1990s were based on statistical methods (Nuttall et al., 1974; Kelly, 1976; Fitz-Gibbon and Vincent, 1994; Dearing, 1996), which some (but not all) authors seemed to interpret on the basis of a purely statistical definition of comparability. The statistical conception maintains that two exams are comparable if a 'typical' student has an equivalent opportunity of attaining a particular grade in each. No consideration of exam demands or the quality of student performance is required. Here, the standard depends on its likelihood of being reached, possibly after taking into account other factors. 'Difficulty' is defined in terms of students' relative chances of success in a qualification:

> The term 'difficult' cannot be taken as meaning necessarily or intrinsically difficult. Rather, subjects are said to be either 'difficult' or 'severely graded' if the grades awarded are generally lower than might have been expected on the basis of adequate statistics (Fitz-Gibbon and Vincent, 1994, p. i).

> If we say that physics is 'harder' than English, for example, we are not saying anything about the relative demands of the two examinations in terms of the kinds of skills and abilities required to succeed in them. Instead, we are simply saying that a 'typical' student has a better chance of getting a particular grade in English than he or she does in physics (Coe, 2010, p. 275).

Of course, there may be many other factors that influence whether physics is 'harder' than English – it may be less interesting, worse taught, less crucial or given less curriculum time than English, for example. These factors are typically not accounted for by statistical methods. More importantly, though, they are simply not relevant from the perspective of a statistical definition. In other words, a stakeholder who adopted a statistical definition of comparability would not want to take into account those additional factors.

Coe (2007) argued that the statistical conception of comparability is the most open-minded on the issue of which subjects can be compared. If comparability is based on the 'chances of success' principle, then there is no reason to exclude any subject from comparison. All that the statistical conception requires is that the same grade should be 'equally reachable' in each subject. This contrasts starkly with the traditional criticism of statistical methods: that results in different subjects must correlate sufficiently to warrant the assumption that they are all measuring some kind of 'general academic ability' (for example, Goldstein and Cresswell, 1996; Newton, 1997). This is presumably because an alternative conception of comparability is implicit within the traditional criticism (see Newton, 2012).

3.1.3 Conventional comparability

Coe (2010) identified a third notion of comparability in the literature, which does not fit into either the performance or statistical conceptualisation. He called this 'conventional comparability' and aligned it with 'sociological' or 'conferred power' definitions (see Cresswell, 1996; Wiliam, 1996b; Baird, Cresswell, and Newton, 2000; Newton, 2005; Baird, 2007).

This approach sees standards as a social convention, defined by the values of a 'community of practice' rather than by any explicit rationale (Wiliam, 1996b). Judgements are made by those who have been empowered to make them and who are (often, although not always) uniquely qualified to make them. In other words, subject standards are comparable if experts tell us that they are and if we accept their conclusions (Newton, 2005). Such judgements will always be somewhat subjective, in much the same way as are the decisions of a judge or jury. Having outlined this definition, Coe (2010) ultimately rejected it as "inadequate" (p. 271), arguing that "it offers nothing in the way of a conceptualisation" (p. 277).

3.1.4 Construct comparability

After summarising the definitions in the literature above, Coe (2007, 2010) proposed an alternative conception, which he called 'construct comparability' – a notion "subsuming both performance and statistical conceptions" (Coe, 2010, p. 280). This definition was based on the idea that two exams may legitimately be compared provided they share a common 'linking construct'. Here, the standard of a particular exam performance is dependent upon the level of the linking construct that it denotes. Any comparison made between two exams is valid only with reference to the linking construct identified: "There is no absolute sense in which one examination is harder than another – it depends on the construct" (Coe, 2010, p. 271).

The idea of a linking construct was introduced as a general model for understanding comparability in Newton (2005), was elaborated on in Newton (2010b), and was applied to the challenge of inter-subject comparability in Newton (2012). The idea of specifying a linking construct as the basis for linking standards across subject areas can be interpreted in a variety of ways.

If, for instance, it were possible to identify a substantial set of skills, knowledge or understandings in common to certain subjects, then this could be specified as the linking construct. Inter-subject comparability could, therefore, be defined – albeit for those subjects only – in terms of a common level of performance/attainment in that subset of common skills, knowledge and understandings. For example, 'independent research ability' might be proposed as a linking construct between psychology and geography A levels, even if it were the primary purpose of neither exam to measure this construct. To link standards across a wider group of subjects, it might be necessary to define inter-subject comparability in terms of a far broader construct, for

example in terms of a common level of 'general academic ability', which provides a useful label for encapsulating the idea of a general ability to do well at school.

Clearly, if we adopt the 'construct comparability' conception, then it is possible to reject the idea that comparability is necessarily a subject-specific notion; instead, we define comparability on the basis of a subject-general entity that is a subset of the various subject-specific entities whose standards are to be linked. Importantly, the linking construct represents only that dimension, or sub-construct, that is assumed to be common across subjects. Even in the case of semi-cognate subject clusters, such as mathematics and chemistry, the majority of subject-specific skills, knowledge and understanding would be unique to each subject, and could not, therefore, be taken into account by the linking construct. Moreover, if the common element were to represent a different proportion of the full set of learning outcomes for each of the subjects being compared, then this might raise further questions concerning the fairness of the comparison.

Coe (2010) accepted that it may seem implausible that a common sub-construct could be found to link standards across a large group of widely different subjects. Furthermore, exams measure educational attainment in relation to highly specific areas of the curriculum, and this might raise ethical questions concerning the social defensibility of linking on a far more restricted basis. However, Coe argued that the assumption of a common sub-construct is surprisingly well supported by the empirical evidence. He found that, for 34 apparently diverse GCSE subjects, a single latent trait explained 83 per cent of the observed variation in performance (Coe, 2008).

Whilst linking constructs may take many forms, scholars have often discussed inter-subject comparability at A level and GCSE in terms of a very broad common construct of 'general aptitude' or 'general academic ability' (for example, Fitz-Gibbon and Vincent, 1997; Newton, 2005; Coe, 2008). Coe (2007) provided the example of a university admissions tutor using grades from across a profile of subjects to surmise a student's suitability for entry (most grades, therefore, coming from subjects other than those taken by the student). He suggested that, for this particular use of results, it would be highly desirable to be able to interpret the same grade from different subjects in terms of the same level of general ability for learning, a concept that is quite similar to the idea of 'general aptitude' or 'general academic ability'. If all the subjects being compared do measure 'general aptitude' – at least to some extent – then their outcomes can be compared, and linked, along this dimension. We might then interpret grades in terms of this linking construct. So, having adopted and applied this conception of comparability, we could conclude that a particular grade in (say) physics indicated the same level of general ability for learning as the same grade in French, or geography, or English.

3.1.5 Alternative frameworks

As noted earlier, Coe's framework for conceptualising comparability is one amongst various, and no one framework is generally accepted within the field. Indeed, it is not even clear that the field is united in accepting the legitimacy of alternative conceptions of comparability, per se. Newton (2010b) proposed an alternative framework, which was essentially an elaboration of the linking construct model. He proposed three major categories of comparability definition:

Phenomenal definition. For two exams to have comparable grading standards, students who score at linked (grade boundary) marks must be the same in terms of the character of their attainments. They must be equally good at knowing, understanding, and being able to do X, where X is a set of knowledge, skills, and understanding that is common to both exam constructs.

Causal definition. For two exams to have comparable grading standards, students who score at linked (grade boundary) marks must be the same in terms of the causes of their attainments. They must have experienced equally the set of causes that are common to attainment in both exams. This notion overcomes some of the problems associated with the phenomenal definition; in particular, the requirement that a subset of attainment can be identified which is common across highly diverse subjects. It seems far more plausible to assume that a cause of attainment could be identified across highly diverse subjects which could be singled out as a possible linking construct (for example, student effort or study time). More plausible still, we might single out a set of causes – a composite linking construct, like general academic ability – as the basis for a causal definition of comparability across subject areas.

Predictive definition. For two exams to have comparable grading standards, students who score at linked (grade boundary) marks must be the same in terms of the extent to which their attainments predict their future success. That is, they must indicate the same potential for success in the future (for example, the same likelihood of gaining a 2:1 or above in a university degree).

In response to analyses that had been produced by the National Foundation for Educational Research, Robert Wood, then Head of Research at the University of London exam board, wrote to the *Times Educational Supplement*, essentially dismissing the concept of inter-subject comparability as a "lunatic idea" (Wood, 1976/1987, p. 42). Implicit in his dismissal was the idea that comparability is irreducibly a performance/attainment concept, and that inter-subject comparability simply cannot be construed in these terms. This way of thinking was also implicit in the conclusion from Newton (1997) that "we should learn to accept and adapt to the

unintelligible enigma of comparability between subjects" (p. 448). Although some[2] comparability scholars would probably still side with Wood, others are no longer so sure. The framework provided by Newton (2010b), for instance, was a direct response to his earlier recommendation, as it identified a variety of ways in which the enigma of inter-subject comparability might be made intelligible.

Although none of the frameworks could (yet) be said to have achieved widespread support amongst comparability scholars, they do at least offer alternative ways of understanding what might otherwise be dismissed as unfathomable.

[2] If not 'many' – it is impossible to get a sense of scale from the literature.

4. Comparability methods

As numerous as the various concepts of comparability are, there are various methods that can be used to investigate or address it. Once again, debates concerning comparability methods are complex and highly contested, with no consensus as to the most valid approaches. Indeed, Newton (2012) argued that these debates had been further complicated by a failure to distinguish between criticism of the method and criticism of the definition; implying that both critics and defenders of alternative methods may have been right, on their own grounds, but without realising that they were arguing from different premises. The literature is, therefore, not simply confusing, but possibly also confused in places.

Methods for investigating inter-subject comparability fall within two broad groups – statistical and judgemental. Although methods and definitions do not relate to each other in a one-to-one manner, certain methods do lend themselves more readily to certain definitions of inter-subject comparability. The literature on the use of these techniques to compare inter-subject standards is lengthy, complex and highly technical. Whilst we try to provide a clear sense of the analytical detail, the discussion below is inevitably over-simplified in places.

4.1 Statistical methods

Statistical methods are based on the idea that the standard of a subject can be judged by analysing the number or proportion of students attaining given grades in relation to data related to concurrent or previous measures of attainment (Elliott, 2013). Statistical methods are particularly compatible with statistical conceptions of comparability where 'difficulty' is defined (in one way or another) in terms of students' relative chances of success in different subjects. However, they may also be used to investigate inter-subject comparability from alternative perspectives, for example with reference to a linking construct such as general academic ability (Coe et al., 2008), assuming that certain statistical assumptions are not obviously violated.

We should note that these statistical methods are not the same as simply comparing the grade profile across subjects. Indeed, this may give another picture entirely. In 2013, A level mathematics students may have appeared, on the face of it, to be more likely to achieve higher grades, with 43 per cent of students gaining an A or A* compared with 21 per cent in English and 11 per cent in media studies.[3]

Attempts by organisations to compare and align standards between subjects have almost invariably used statistical methods. In the past, English exam boards have

[3] Figures taken from the Joint Council for Qualifications.

used such comparisons to monitor standards across subjects and inform grading (Nuttall et al., 1974), although these methods have only played a minimal role in the awarding process in recent years (Jones, 2003, 2004). This decline is both due to the increased emphasis on maintaining other forms of comparability – standards over time and between exam boards – as well as growing concerns about the validity of statistical methods.

Statistical methods actually refer to a number of different approaches. Coe et al. (2008) divided these into five groups. The first three – subject pairs analysis, common examinee linear models and latent trait models – can be categorised as 'common examinee methods', as they rely on comparisons of the results achieved by the same student in different exams. The underpinning principles behind these approaches are summarised in the much-used quote by Nuttall et al. (1974):

> We (the three authors) argue as follows: we do not expect an individual candidate to achieve the same grade in every subject that he takes. However, we can see no logical reason why, if a large group of candidates representative of the population took, for example, both English and mathematics, their average grades should not be the same (p. 12).

The most widely used statistical method is subject pairs analysis. This method has been the basis for many of the high-profile (and controversial) studies of inter-subject standards, including those by Fitz-Gibbon and Vincent (1994) and those presented in Dearing (1996). As such, it has received the most criticism of any comparability method. It is also, perhaps, the simplest method conceptually, as outlined by Coe (2007) below:

> If we pick two subjects to compare we can consider all candidates who have taken both. We then simply calculate the difference between the mean grade achieved by those same candidates in each subject. If they typically achieve better grades in one than the other we may say that the former is 'easier', the latter 'harder' […] (p. 334).

Common examinee linear models and latent trait models (such as the Rasch model) are more complex. These methods are not widely used in England but have been applied in Scotland (Kelly's method) and Australia (Average Marks Scaling and the Rasch model). These methods overcome some of the problems with simpler statistical methods, for example compensating for the fact that students taking 'harder' subjects are more likely to take them with other 'harder' subjects, and similarly that 'easy' subjects are more likely to be combined with other 'easy' subjects (Coe et al., 2008).

The other two statistical methods – reference tests and value added methods – are broadly similar to each other. They both rest on comparing the grades awarded by exams in different subject areas for students with the same level of attainment, as indicated by an independent measure: results from a concurrently administered (reference) test; or results from a test, or tests, administered years earlier (value added). So, if the overall performance of two groups of students on a reference test is similar, but the students attain markedly different grades in different subjects, this is taken as an indication that the two subjects are not comparable in terms of their grading standards (Murphy, 2007). Value added methods are already used in England to address other forms of comparability. Prediction matrices, based on prior attainment in Key Stage 2 tests and GCSE exams, are used (respectively) in the GCSE and A level awarding process to improve comparability of similar qualifications, between exam boards and over time.

Both reference tests and value added methods are fairly straightforward and can be publicly explained with relative ease. They also do not require the exams being compared to have any common students, which is an advantage over common examinee methods. However, they depend upon the independent measure having the same strong statistical relationship across the range of exams being compared. Critics argue that neither reference tests nor prior attainment indices have a consistently strong relationship with exam attainment across the full range of examined subjects (McGaw et al., 2004; Murphy, 2007; Tremain, 2008).

Although different statistical methods of investigating inter-subject standards give somewhat different results, some research indicates that all methods produce generally consistent findings on which subjects are most ‘severely’ graded (Coe et al., 2008). For some, this recurring pattern of statistical differences in the grades that the same students achieve in different subjects at GCSE and A level might indicate the existence of a phenomenon which we cannot ignore or dismiss as the result of a problematic or uninterpretable methodology (Coe, 2007, 2008). However, for others, this consistency is only to be expected, given that the statistical methods operate on a similar logic and are underpinned by the same assumptions (Newton, 2007).

As we discuss below, there are many important criticisms of statistical methods for achieving comparability. Many commentators have argued that these are sufficiently compelling to reject the claim that these methods indicate genuine problems (Pollitt, 1996; Goldstein and Cresswell, 1996; Newton, 1997; Jones, 2003; Murphy, 2007). Having said that, Murphy (2007) argued that they can still “provide interesting data, which need to be handled cautiously by those who know both the strengths and weaknesses of the approach” (p. 309).

4.2 Criticisms of statistical methods

Criticisms of statistical methods are many, and critics are often vociferous. The criticisms include both philosophical and technical concerns. Various commentators

have contributed to the debate, including Christie and Forrest (1981), Newbould (1982), Alton and Pearson (1996), Pollitt (1996), Cresswell (1996), Goldstein and Cresswell (1996), Newton (1997), Fitz-Gibbon and Vincent (1997) and Jones (2003), to name a few.

Whilst, at first glance, statistical methods appear simple, many authors agree that "there may be many reasons why candidates achieve different grades in different subjects, apart from the obvious one that some are harder" (Coe, 2007, p. 332). To a greater or lesser degree, all statistical methods are underpinned by similar (and significant) assumptions, which many argue means they cannot yield "unambiguous, unproblematic conclusions" (Jones, 2004, p. 1).

A criticism of purely statistical approaches – which some commentators consider sufficient to undermine the validity of any comparability conclusion drawn from them – is that they ignore the educational content and demand of the specifications and exams being compared. Goldstein and Cresswell (1996) provided the hypothetical example of a comparison between a spelling test and an A level in English. Whilst the two assessments could be made 'comparable' – from a statistical perspective – it would be invalid to say that their standards were 'equal' because they clearly do not measure the same thing. As such, it would not be appropriate to treat their outcomes interchangeably, when used (for example) as the basis for selecting students to a degree course in English. Statistical methods are only ever a proxy measure of exam performance (Baird et al., 2000) and, consequently, many consider them inadequate for monitoring or achieving comparability.

In this section, we outline some of the major criticisms of statistical methods. The list is far from exhaustive, and Coe et al. (2008), Coe (2007) and Murphy (2007) give more detailed criticisms of each specific statistical method in turn. The main arguments against simple statistical approaches can be grouped under the following four headings. Many of these arguments have been specifically directed at subject pairs analysis. However, they are likely to hold true for other methods too, given the shared assumptions underpinning them (Pollitt, 1996; Newton, 1997). The major criticisms are as follows:

1. Statistical approaches make an assumption of unidimensionality, yet evidence from inter-correlations questions this assumption.

2. Performance in exams is affected by many factors apart from 'general academic ability'.

3. Groups of students taking particular combinations of subjects are not representative of the full cohorts who take those subjects.

4. Apparent differences in subject 'difficulty' between different sub-groups of students throw doubt upon the interpretation of overall comparisons.

We should note that the extent to which any of these criticisms are considered threats to valid interpretation of outcomes from statistical methods will vary according to the conception of comparability that we are using. This might, for instance, depend on whether we are using statistical methods to compare relative 'chances of success' in subjects or relative 'levels of attainment'. Coe et al. (2008) stressed that whilst statistical methods are problematic, and can suffer from many of the issues discussed below, there are still some valid interpretations which can be drawn from their results.

4.2.1 Unidimensionality

Many argue that the whole concept of comparing two subjects is only meaningful if there is some sense, or some dimension, in which the subjects are 'the same' (Coe et al., 2008). In other words, the subjects must measure the same thing, or at least have a significant trait in common. The extent to which subjects measure the same thing is the extent to which they are 'unidimensional'. We know that, between subjects, this is simply not the case at an overall level. Therefore, in an important sense: "It is meaningless to say, for example, that 'art is easier than physics'; they are just different." (Coe et al., 2008, p. 116).

This assumption of unidimensionality is, perhaps, the single most significant criticism of statistical methods. In fact, unidimensionality is a philosophical challenge to comparability, which could, therefore, apply to any method whatsoever, as well as to any form of comparability. This is important, because even when comparing standards between exams for the same subject, for example between exam boards, the assumption of unidimensionality is violated to some degree, as alternative specifications assess significantly different aspects of the same subject (which, of course, is a key feature of the qualifications market). The question, then, is how much of a violation of this assumption can be tolerated. In the context of inter-subject comparability, McGaw et al. (2004) described the unidimensionality assumption as a "heroic" one (p. 30). In their critique of Fitz-Gibbon and Vincent's 1994 analysis of inter-subject standards, Goldstein and Cresswell (1996) argued that even when dealing with different exams within a single subject area, it is very difficult to assume unidimensionality. "When, however, the focus of attention is comparability between subjects, the need to assume unidimensionality is clearly a major impediment to its satisfactory definition." (Goldstein and Cresswell, 1996, p. 436).

Newton has also written at length about the concept of unidimensionality (1997, 2005). He proposed that, for statistical comparisons between subjects to be interpreted validly, we must take one of two positions. Either we accept that there is a sufficient degree of unidimensionality – which might be described as an underlying 'linking construct' common to all subjects – or we recognise that exams measure entirely different constructs, and take a norm-referencing approach to linking inter-

subject standards. He observed that one way of viewing a linking construct between subjects might be in terms of general academic ability:

> My assumption of 'general academic ability' is not the invocation of some innate, genetically inherited, intelligence. Instead, I use it to refer to the composite of all those factors that would influence a candidate's performance in examinations—a candidate's general capacity to do well in examinations (Newton, 1997, p. 439).

Thus 'general academic ability' might well reflect genetic factors, but would certainly also reflect environmental ones, including student motivation to succeed, parental support and encouragement, or an individual's attitude to schooling. But, it is still a general concept: "Whatever factors contribute to an individual's examination success in one subject contribute similarly across all subjects." (Newton, 1997, p. 439).

Although the term 'general academic ability' has been widely used in the literature – often in the 'broad' sense that Newton (1997) made explicit – it is perhaps not ideal in this context, because the idea of 'ability' is often (in other contexts) associated primarily with cognitive factors. Some might even read into the term 'ability' shades of innateness that would be even less appropriate. In the absence of a conventional label for the construct of interest, here, the term 'general academic application' might be useful. It is intended to embrace the full range of factors that a student is able to apply to his or her course of learning; in other words, how effective that student is at learning.

A particular concern, when rationalising inter-subject comparability in subject-general terms, is that we are unable to account for the demand of skills, knowledge and understanding which are unique to individual subjects or subject types. If we compare English and chemistry, for example, how much of the essence of each subject is left by the time we have found a common linking construct? Perhaps not much. To be fair, even Newton (1997) questioned whether it was plausible to assume a sufficient degree of unidimensionality to underpin inter-subject comparability monitoring – either practically or theoretically – and ultimately concluded that the best approach may be simply to accept the enigma of inter-subject comparability.

Over the years, many commentators have questioned whether it is legitimate to compare subjects that are so different. Some have suggested that we should be content with ensuring that clusters of cognate subjects are comparably graded (see Coe et al., 2008). This approach seems to be taken in some education systems where students select from groups of similar subjects. Yet, this idea of comparability within clusters negates one of the more common complaints about inter-subject comparability from certain sectors of the education community: namely that particular subject clusters, for example sciences and languages, are perceived to be harder

than other clusters of subjects, and that action should be taken to address this imbalance.

Returning to our earlier point about methodological criticisms varying in importance depending on how we interpret and use results, some authors have simply dismissed the relevance of philosophical criticism altogether. In response to Goldstein and Creswell's (1996) critique of their 1994 paper, Fitz-Gibbon and Vincent (1997) insisted that they had not intended to take a philosophical stance, but merely worked out "quantitatively what most students and teachers simply know qualitatively: on average, students do not stand to get as high a grade in mathematics as in, for example, English" (p. 292). They claimed that the statistics were correct, but acknowledged that they were amenable to a different interpretation of 'difficulty' based on the relative chances of success of students in different subjects.

Coe et al. (2008) also argued that viewing differences in grades in terms of students' chances of success allowed us to be more open on the question of which subjects can be compared: "This conception makes no requirement for different subjects to be related in any way, only that a particular level of achievement should be equally rare in each." (Coe et al., 2008, p. 121). This is some way from the traditional view of success in exams, which is generally presumed to be defined in terms of absolute levels of attainment, that is what a student has achieved in terms of his or her knowledge, skills and understanding.

4.2.2 Factors other than 'general academic ability'

Another criticism – that is probably best viewed as the 'other side of the unidimensionality coin' – is that subjects are specifically intended to measure far more than just general academic ability. Again, this response goes some way towards upholding the traditional view of success in exams, based on subject-specific attainment rather than subject-general aptitude. It notes that statistical methods do not model the impact of certain factors that, most stakeholders might presume, ought to be 'rewarded' by exam grades. Conceivably, if the impact of these factors happened to differ systematically across subjects, yet inversely to the impact of general academic ability, the effects might even cancel out. Or, to put it another way, if it were possible to measure these other factors alongside general academic ability, then even statistical methods might lead to the conclusion that subjects were comparable after all. At the very least, the failure to model factors that most stakeholders would consider integral to the meaning of success in exam subjects is sufficient to cast major doubt upon the validity of conclusions from statistical methods, according to many commentators (see Pollitt, 1996; Alton and Pearson, 1996; Goldstein and Cresswell, 1996; Newton, 1997; Newbould, 1982; Coe, 2007; Sparkes, 2000).

The kinds of factors at stake here would include: interest in the subject; the quality of teaching or amount of teaching time; motivation to achieve a particular qualification;

and so on. Subjects like art and media studies often appear 'leniently' graded, according to outcomes from statistical methods. Yet, if art or media studies students happened (as a group) to be more motivated (in art or media studies) than business studies or chemistry students (in business studies or chemistry), then maybe this would be enough to offset any general difference in general academic ability? Without the ability to measure these factors accurately or comprehensively, we have no way of knowing for sure. But the argument is fair enough in principle. Similarly, general studies often appears to be 'harshly' graded, according to outcomes from statistical methods. Yet, if general studies students happened (as a group) to spend fewer hours studying (general studies), then perhaps their ostensibly lower grades would be deserved?

Unless it can be assumed or demonstrated that these factors are equal across subject areas – or unless it can be argued that differences in these factors are irrelevant – difficulty cannot be judged simply by comparing results statistically, at least according to a more traditional view of success in exams (Coe, 2007). Subjects that appear to be 'misaligned' using simple statistics might, in fact, be aligned if these other factors were allowed for. For this reason, Pollitt (1996) warned that: "A subject pairs analysis is simplistic and dangerously misleading." (p. 4).

Unfortunately, many if not most of these factors are extraordinarily difficult to build into statistical analyses. Where research has taken place, however, it appears their impact can be significant. Using a combination of subject pairs analysis and student preference data, Newbould (1982) found a strong correlation between student preference (taken as a proxy for motivation) and 'ease' of a subject. In other words, subjects that would be considered to be 'easier' in terms of statistical analyses are also the subjects which appear to 'motivate' students more. Similar conclusions have been reached by Massey (1981), Rutter (1994) and Sparkes (2000). A more recent study by Korobko et al. (2008) used an item response theory model to analyse data from the Netherlands. This found that student preferences did affect performance in certain subjects. For example, language-orientated students were more likely to perform better in Dutch than students more focused on science subjects. This pattern was reversed when the researchers considered mathematics-orientated students.

Because of these issues, some experts argue that, in the context of subject pairs analysis in particular, terms such as 'leniency' and 'severity' should be used very carefully, as statistical differences between subjects may not reflect genuine differences in subject standards (see Jones, 2004; Newton, 2007). However, Coe (2010) argued that the impact of these factors would not be of interest or relevance to all users of qualifications. For example, university admissions tutors may be less interested in what grades indicate about subject-specific attainment, and more interested in what grades say about general academic ability. If so, then perhaps the impacts from factors other than 'general academic ability' ought to be factored out of exam grades after all?

4.2.3 Unrepresentativeness

A technical criticism of particular significance to techniques like subject pairs analysis is that groups of students with particular combinations of subjects – who form the basis for statistical comparisons – tend not to be representative of the full cohorts of students taking each of those subjects (Coe, 2007). For example, A level students taking both physics and media studies are not likely to be representative of those taking either subject. This argument was also raised by Goldstein and Cresswell (1996) and Newton (1997).

In some contexts, this criticism may have less force than in others. For instance, at GCSE, almost all students take both mathematics and English. However, for comparisons like the physics-media studies example, where the common pool of students is far smaller, the criticism is likely to have more force. Under these circumstances, the inferences we can draw from statistical comparisons are correspondingly limited (Coe, 2007).

Newton (1997) considered the issue of representativeness from a more philosophical perspective, which raised questions concerning the nature of the population in terms of which representativeness ought to be judged. Is it, for instance, the subset of A level students who actually took physics (or media studies) that year? Or is it the full set of A level students who could, had they wished to, have taken physics (or media studies) that year? Or is it the full cohort of 18-year-olds? Issues of representativeness are hard to conceptualise, let alone resolve.

The threat of unrepresentativeness is far higher in education systems where students are expected to choose a small number of preferred subjects, that is in systems where students specialise early rather than studying a common broad curriculum. The greater the specialisation, the greater the threat of unrepresentativeness, and the less robust any conclusion from statistical approaches (McGaw et al., 2004; Lamprianou, 2009).

4.2.4 Sub-group differences

Various studies have found that when we repeat statistical analysis, conducting it separately for distinct sub-groups of students, we do not always get the same results as we did for the whole population (Kelly, 1976; Pollitt, 1996; Newton, 1997; Newbould, 1982). For example, history may appear to be ‘harder’ than mathematics for males, whilst the reverse may be true for females (Pollitt, 1996). Differences are often most salient for comparisons by gender, but also exist between students from different school types and between those with different subject choices (Kelly, 1976; Fitz-Gibbon and Vincent, 1994; Rutter, 1994; Sparkes, 2000). Findings for Scottish Highers and A levels led Sparkes (2000) to conclude that the sub-group variations between subjects for these qualifications were so significant that “subject ‘difficulty’, measured in purely statistical terms, is unhelpful” (p. 188).

The sub-group issue raises problems for the application of any overall adjustment to grade boundaries. If we adjust subject grading standards at the overall level, we may risk making subjects less comparable for certain groups of the population (Alton and Pearson, 1996). Nuttall et al. (1974) suggested that the only way to overcome this issue would be to adjust grades for different sub-groups separately; thus, for example, achieving inter-subject comparability for each gender. Newton (1997) argued that, even if this were politically and publically acceptable, it would still not be enough as there were so many other factors which led to variations in results, including type of school, tier of entry, and third or fourth subject.

A broader implication of this criticism is that any verdict about subject difficulty depends on the demography of the students who happen to take the subjects being compared. In other words, the definition of 'standards' is population-dependent. As such, if the features of the entry change, then so would apparent 'difficulties' (Alton and Pearson, 1996; Pollitt, 1996; Newton, 1997; Sparkes, 2000).

4.3 Judgemental methods

Judgemental methods are based on the view that we can judge exam standards by considering candidate responses to exam tasks. Unlike statistical methods, they take into account the educational content of specifications and exams. They rely upon human judgement to detect and compare 'the standard' by empowering subject matter experts, often senior examiners, to scrutinise assessment materials and performances (Wiliam, 1996a, 1996b; Adams, 2007; Coe et al., 2008). The undoubted strength of judgemental methods is that they appear 'sensible', making comparability more accessible to the layperson (Adams, 2007). Bramley (2011) explained that:

> When investigating comparability of assessments, or of qualifications, we have focused mainly on comparing them on the basis of: i) the perceived demands (of the syllabus and assessment material); and ii) the perceived quality of examinees' 'work'. Both 'perceived demand' and 'perceived quality' might be thought of as higher-order attributes that are built up from lower-order ones. The definition of these attributes suggests that they be investigated by methods that use the judgement of experts (quoted in Elliott, 2013, p. 10).

Judgemental methods have routinely been used to compare exam standards across exam boards and over time, for within-subject comparisons. We and the exam boards, as did our predecessor bodies, have undertaken rolling programmes of comparability research along these lines for years. However, it is rare for judgemental methods to be used to compare standards across subjects, because it is very hard to agree on what basis such comparisons might be made. Coe et al. (2008) noted that,

if the standard of an exam is inherent in the skills, knowledge and understanding demonstrated in observed performances – judged to be worthy of specific grades – then we would require generic, cross-curricular performance criteria, against which to compare two or more subjects. Yet, the development of plausible performance criteria can be challenging even within subjects, let alone between them.

Despite this, inter-subject comparability research studies based on judgemental methods have been undertaken, and they have claimed some success, albeit only within fairly cognate subject clusters. Research includes a study by QCA (2008), where comparisons were made between subjects within a range of subject clusters, including: GCSE geography and history; GCSE sciences; A level biology, psychology and sociology; and A level history, English literature and media studies.

Coe et al. (2008) noted that there were essentially two types of judgemental methods used to determine comparability, the first involving absolute judgement against an explicit standard and the second involving relative judgement against other scripts:

Judgement against an explicit standard. In this case, the standard required for the award of a specific grade is explicit. Experts judge whether candidate performances in different exams meet that standard. These approaches are sometimes termed 'cross moderation' methods. They are sometimes hindered by the difficulty of creating suitable descriptors, or performance criteria. Additionally, the requirement to define and exemplify criteria explicitly can lead to a fragmented view of performance. This makes it hard to accommodate 'compensation', whereby poor performance in one area can be made up for by excellent performance in another (Adams, 2007).

Judgement against other scripts. This approach avoids the problem of having to define explicitly a grade standard. By using 'paired comparison' methods, experts are provided with a pair of scripts and asked to judge which is 'better' in terms of performance quality. The results of several judges drawing their conclusions about multiple pairs can be combined using Rasch analysis, which locates all the scripts on a single scale of 'quality' (Bramley, 2007). Once the scripts from two exams have been located on the same scale, it is a relatively straightforward matter to 'carry forward' the grade standard from one exam to the next (when linking standards), or to compare grade standards that have previously been established on both exams (when monitoring comparability).

4.4 Criticisms of judgemental methods

Whilst, in the context of inter-subject comparability, judgemental methods have received far less criticism than statistical approaches – simply because they have been used far less frequently – they are certainly not without their limitations. They rely on the ability of experts to make inherently complex, and necessarily ill-defined, subjective decisions. Some researchers have warned that conclusions from judgemental studies will always be highly equivocal: exam standards and human

judgements are so nebulous in nature that little definitive can be said of them (Adams, 2007).

The task of making detailed, multi-faceted comparisons across exams can be extremely difficult, even when making comparisons within the same subject. Baird et al. (2000) discussed studies where experts had been asked to compare standards between exams in the same subject area, over time, where specifications had changed considerably. They noted that: “many of them have been forced to conclude that the task is, essentially, impossible” (p. 216). And, of course, in the case of inter-subject comparability, the task demands are far more complex.

Although judgemental methods could, in theory, be used to monitor comparability across subjects, it is hard to imagine on what basis judges would make their decisions. Cross-curricular performance criteria for inter-subject comparisons would most likely end up either too ambiguous to be usable or too constricted to reflect the core of any subject (Wiliam, 1996a; Coe, 2010). Although a technique based on ‘judgement against other scripts’ might not require the construction of explicit cross-curricular performance criteria, judges would still need to compare scripts from different subjects on the basis of something like those criteria, even if they were unable to articulate the basis for their decisions explicitly. The task, even if it could be defined, would clearly be exceptionally complex, lengthy and hindered by the availability of people who could be considered somehow ‘qualified’ to make such judgements.

Consequently, another major limitation is that studies of inter-subject standards: “have been hampered by the shortage of experts who are suitably qualified to make valid comparisons between disparate subject areas” (Jones, 2004, p. 1). The kind of expertise required for such comparisons is hard to define, partly because the kind of inter-subject comparability that it is presumed capable of monitoring is also hard to define, and would typically not be defined. Some researchers question whether anyone could be expected to be appropriately qualified to make such judgements (Goldstein and Cresswell, 1996).

> Identifying people competent enough to judge with authority the relative standard of attainment of diverse subjects, let alone make judgements which are publicly accepted, would prove difficult in most cases and impossible in some. (Jones et al., 2011, p15)

Nonetheless, the 2008 QCA studies suggested that this might be possible, if only in the case of cognate or semi-cognate subject clusters.

Coe et al. (2008) identified a number of additional issues in the use of judgemental methods, including those summarised below.

Crediting responses in the context of different levels of demand

Research into comparability across exam tiers has found that, when the same students respond to questions from different tiers, their responses to the easier questions are judged more favourably (Good and Cresswell, 1988). Comparing a good answer to an easy question and a less polished answer that the same student might have given to a harder question, examiners tend to give more credit to the former (Cresswell, 2000). Contextual effects such as this would inevitably impact upon inter-subject comparisons, although exactly how is hard to say.

Crediting responses in the context of different response formats

Coe et al. (2008) suggested that this issue would have affected the QCA inter-subject comparability studies (QCA, 2008). Judges in the QCA studies found that GCSE geography was dominated by short-answer questions, which focused on very specific items of knowledge, compared with the more open-ended essay question in history, which necessitated significant intellectual and communication skills to construct a good answer. Coe et al. (2008) questioned how a common conception of difficulty could be applied to both types of response formats, and queried whether, in this instance, the judges would have been capable of quantifying levels of difficulty on a common scale.

Crediting responses in the context of different assessment structures

The structure of the assessment for a qualification can make a huge difference to its difficulty. Coe et al. (2008) provided an example of a comparison between a specification with a single terminal exam and one with modular exams contributing to the overall grade. The content of the two syllabuses might be exactly the same; even the exam questions might be the same. Yet, by breaking the assessment into smaller units, and by allowing exams to be retaken, the modular specification effectively makes the same ‘standard’ much easier to reach.

Crediting responses in the context of different opportunities to learn

One example of this issue relates to the relative attainments of students in different subjects when they commence their course of study. In England, it was often the case that students studied French from the age of 11, or even earlier, but started German later. Coe (2010) queried, in this context, “if we find that comparable candidates typically gain higher grades in French than they do in German, is this evidence that the latter is harder, or is it consistent with the view that the standards in the two are the same, but that, as we would expect, performance in German is typically lower?” (p. 272).

Of course, many contextual factors simply cannot be known to the individual who is judging student performance; for example, the quality of teaching, which represents

another example of crediting responses in the context of different opportunities to learn. Even if they are known, it is hard to see how they could ever adequately be taken into account as one component of already hugely complex judgements (Baird, 2007; Coe, 2010). Again, though, it is not even clear whether (let alone how) each of these factors ought to be taken into account.

5. Differences in standards between subjects

Over the years, a large number of studies have been undertaken in an attempt to measure differences in 'difficulty' between subjects. Below, we review findings from such studies and comment on the patterns shown. The evidence from statistical analyses in the UK shows a high level of consistency in estimations of subject 'difficulty' (Coe et al., 2008). Having said that, the methods used have been based on the same underpinning logic and assumptions, and are subject to the limitations discussed in the previous section. This has implications for how we interpret the data. Indeed, many of the authors of the studies below stressed that their findings must be treated with great care and should not be viewed as in any way definitive (Nuttall, 1974; Alton and Pearson, 1996).

Some authors have suggested how the concepts of 'standard' and 'difficulty' should be defined and interpreted with reference to their studies (Fitz-Gibbon and Vincent, 1994; Sparkes, 2000; Jones, 2003). Most commonly, alternative interpretations have not been discussed explicitly, and even intended interpretations have remained implicit (Newton, 2012). As such, the following findings are important, and significant in their consistency, but their 'proper' interpretation is not at all clear. Consequently, the following sections continue to apply scare quotes to terms like 'difficulty', and the reader is advised to exercise similar restraint in judging the findings.

5.1 Studies pre-1990

In 1974, Nuttall et al. published research that attempted to compare standards at CSE and O level, in England, using five different methods. Their results consistently showed that sciences and modern foreign languages were more 'difficult' than other subjects. For CSE qualifications, physics and chemistry were the most 'difficult'. At O level, the most 'difficult' subjects were chemistry, French and physics. History and mathematics were mid-range, whilst art and English language had the most 'lenient' grading standards.

During the mid-1970s, Kelly developed a more sophisticated version of the subject pairs methodology, and obtained a similar ordering of subject 'difficulty' across four years of Scottish Highers. She found languages, chemistry and physics to be consistently more 'difficult', with biology, mathematics, history and geography around average. Again, English and art were below average 'difficulty', followed by subjects such as home management. Kelly noted that there were variations in these patterns by different student sub-group. Physics, chemistry and mathematics were much more 'difficult' for girls, whilst the opposite was true for modern foreign languages and English (Kelly, 1976).

Similar trends were reported by Forrest and Vickerman (1982) at A level:

> The modern languages as a group appear consistently below the central diagonal line […] along with Physics, Chemistry and the mathematical subjects. Thus the traditional perception that these subjects are among the more 'difficult' in the curriculum is confirmed (Forrest and Vickerman, 1982, p. 45).

5.2 Studies during the 1990s

The 1990s were marked by two high-profile inter-subject comparability studies (Fitz-Gibbon and Vincent, 1994; Dearing, 1996), followed by unpublished research from Alton and Pearson (1996). All three studies used at least four statistical methods to compare A level subjects, and all were highly consistent in the resulting findings, with science and language subjects consistently found to be the most 'difficult'.

In their 2008 report, Coe et al. combined data from all three of these studies to examine patterns of 'difficulty' between subjects. Chemistry and general studies emerged as the two 'hardest' subjects, followed by physics and mathematics. History and economics were also above average 'difficulty'. English, business studies and many social sciences appeared to be graded more 'leniently', with physical education, art and home economics the 'easiest'.

Fitz-Gibbon and Vincent (1994) and Alton and Pearson (1996) both reported notable differences in apparent 'difficulty' between males and females. Alton and Pearson also raised other practical and theoretical concerns about the use of such models, noting that inter-subject standards were not consistent over time and were quite different across the grade range.

5.3 Studies from 2000 to present day

Sparkes (2000) used common examinee methods to compare subject 'difficulty' for Scottish Highers and A levels. He reported students of equal 'ability' getting lower grades in languages and science subjects than in other subjects. The 'easiest' subjects included those with roots in vocational education (crafts, design, home economics) and the arts (art, music, drama). However, Sparkes found student sub-group variations between these subjects to be so significant he concluded that, "subject 'difficulty', measured in purely statistical terms, is unhelpful" (p. 188).

An internal AQA paper by Jones (2003) reported on a subject pairs analysis for 64 A level subjects taken between 1972 and 1999. This supported the general trend found above, with science subjects, mathematics and languages generally the most 'difficult' subjects, social sciences and humanities mid-range, and art and design subjects, business studies and geography the 'easiest'. However, Jones did find notable fluctuations in relative inter-subject standards over time. This inconsistency was less pronounced in subjects with large entries. He concluded that the degree of

consistency of a subject's apparent 'leniency' or 'severity' over time depended, to some extent, on the size of its entry.

In 2008, Coe et al. analysed data from exams awarded in 2006, applying five different statistical methods to 33 A level subjects and 34 GCSE subjects. They reported that consistency in findings across different methods was high – albeit with a few notable anomalies – and that the results echoed the main themes reported in previous studies of inter-subject standards. When the results from all five methods were averaged, general studies emerged as the most 'difficult' A level, followed by physics and chemistry. Again, French, German, history and mathematics were also relatively 'difficult', along with music. Economics, psychology, law and English literature fell around average, with the 'easiest' subjects being film studies, photography, drama and art. At A level, the discrepancy between the 'easiest' and most 'difficult' subjects was just over 1.5 grades.

Coe et al. (2008) found that, at GCSE, there was still a tendency for science subjects to appear more 'difficult', although this was less marked than at A level. Instead, the most 'difficult' subjects were short course IT, statistics and languages. Mathematics and music were around average, with English subjects below average, and art and other performance subjects found to be the 'easiest'. Most methods suggested that the discrepancy between the 'easiest' and most 'difficult' subjects was around 1 grade (excluding short course IT).

After analysing their own data alongside the findings from previous studies, Coe et al. (2008) concluded that the evidence on differential sub-group 'difficulty' was not consistent. Whilst, at A level, sub-group variation did not appear to be significant, at GCSE:

> The differences in difficulty for the two sexes seem to be large enough to challenge the notion of a single unidimensional construct underlying all these different subjects, and to undermine the notion of 'difficulty' as applying to the subject as a whole (p. 109).

5.4 Subject-specific studies

In recent years, a number of small-scale (unpublished) studies have been conducted within exam boards, and focused on specific subjects, in response to public concerns about comparability (see He and Eason, 2007; Jones, 2004; Tremain, 2008; Malpass, 2011). Using statistical techniques, these studies have found that established subjects are not always more 'difficult' than recently introduced ones. For instance, A level accounting and GCSE law seemed to be relatively 'difficult' compared with the average. The studies also replicated Alton and Pearson's (1996) finding concerning the existence of clear differences in relative subject 'difficulty' by

grade. One subject may seem to be 'harder' than average in terms of gaining an A grade, but 'easier' in terms of achieving a C grade (He and Eason, 2007).

5.5 Judgemental studies

Judgemental studies of inter-subject comparability are rare. They are difficult to specify, organise and recruit the necessary experts for. They are also time-consuming and expensive. As such, we only have limited evidence from judgemental inter-subject comparability studies to consider here.

Jones (2004) presented the results of a small-scale investigation, which involved senior examiners judging the quality of A level exam scripts for law and psychology respectively (statistically 'lenient' subjects) and German and physics respectively (statistically 'harsh' subjects) at subject-pairs-analysis-adjusted grade boundary marks (A and E boundaries only). In other words, law and psychology examiners scrutinised scripts that were of higher quality than would otherwise be found at the A and E grade boundaries; whilst German and physics examiners scrutinised scripts which were of lower quality than would otherwise be found at the A and E grade boundaries. The examiners were not told what marks or grades had been awarded to these scripts and were ignorant of the purpose of the study. They were asked simply to indicate the grade that each of the scripts deserved.[4] The purpose of the study was, therefore, to see whether senior examiners considered scripts at the subject-pairs-analysis-adjusted grade boundaries worthy of the grades that would have been awarded had the statistical analysis of inter-subject comparability driven grade awarding. The German and physics examiners tended to respond in a manner that was most consistent with upholding the original standards. In other words, there was some indication that the statistically-adjusted grade boundaries might not have been acceptable to the senior examiners who were responsible for upholding those subject standards; this is useful to know, although perhaps not too surprising. This tendency was less evident for the law and psychology examiners, although the trends were less pronounced and harder to interpret.

In 2008, the QCA reported the results of four investigations conducted into the standards of selected cognate (or semi-cognate) subjects at GCSE, AS and A level. Each study asked subject experts, chosen for their experience of teaching more than one of the relevant subjects, to evaluate the demand of the specification for each subject, as well as to compare examples of students' work at different grades. The study included a comparison of:

- 1a – geography and history (GCSE, AS, A level);

[4] Technically, they indicated 'sub-grades' (B+, B, B-, and so on) for greater precision.

- 1b – biology, chemistry, physics (GCSE, AS, A level);
- 2a – biology, psychology, sociology (A level);
- 2b – English literature, history, media studies (A level).

As well as comparing the content and demand of specifications, studies involving the latter two subject clusters also included a comparative judgement element. Here, a study of grade standards was carried out using Thurstone's paired comparison method to compare the quality of performances observed within exam scripts across the subjects. The following conclusions were reached:

> The review of performance standards at AS found all three subjects very well aligned across the grade range. At A2 there was some evidence that the performance of sociology candidates was not as impressive as that of candidates in either psychology or biology, although the nature of the work used in the study makes it hard to gauge how much weight to place on this finding (2a; QCA, 2008, p. 5).

> The review of performance standards showed there was very little difference between the subjects at the grade A boundary at AS, while at the grade E boundary the performance of candidates in media studies was considered to be slightly less secure than that of candidates in English, with history candidates in between. At A2, at both boundaries, the media studies candidates were considered to be less impressive than the English candidates, with the history candidates in between (2b; QCA, 2008, p. 5).

5.6 International patterns of subject difficulty

It is interesting to note that the general patterns of subject 'difficulty' reported above are not unique to England. Pollitt (1996) reported that similar patterns had been noted in other systems around the world – Scotland, South Africa, New Zealand and Australia, to name a few. We might then ask why – if these common international patterns really do represent grading errors – so many systems have managed to err in the same way. Is this really plausible? Or do these common patterns suggest, instead, that these countries are probably all doing something right, as far as inter-subject grading standards are concerned?

There might, in fact, be plausible explanations for an international phenomenon like this, and similar patterns of subject choice across countries is one possible explanation; more specifically, the possibility that certain subjects (for example, sciences) might tend to attract generally higher attaining students, whilst other subjects (for example, arts subjects) might tend to attract generally lower attaining

students. When combined with a technical desire for similar levels of discrimination across subjects, these patterns of subject choice could go a long way towards explaining the findings commonly observed. That is, student clustering might tend to skew results in certain subjects towards the top of the grade distribution, whilst tending to skew results in other subjects towards the bottom. However, the technical desire for subjects to discriminate similarly well might then act to pull these skewed distributions back towards the average. This kind of effect could cause subjects taken by generally higher attaining students to experience 'grade deflation', whilst causing subjects taken by generally lower attaining students to experience 'grade inflation'. This explanation is quite similar to the contest model of grade awarding, described by Christie and Forrest (1981). It is presented here simply as an example of the kind of impact that could lead to consistent patterns, without proposing that this is the explanation for such patterns.

Pollitt's own explanation posited the existence of psychosocial phenomena that are only partially common across international borders. He justified this by noting that many of the countries in question were similar to England in their culture and education systems. His own study (using subject pairs analysis) compared patterns observed in the UK with patterns observed in a far eastern country where students studied UK A levels but were "culturally quite different" (Pollitt, 1996, p. 1). He found that there was a broad similarity between the pattern of subject difficulty in the far eastern country and in the UK. Again, physics and general studies appeared to be the most 'difficult' subjects, and English the 'easiest'. However, there were also some exceptions. In particular, mathematics appeared to be relatively 'easy' and business studies relatively 'difficult' in the far eastern country. He concluded that:

> The only way to explain these oddities is by assuming that there are significant differences between East and West in subject selection and hence in subject specific ability and motivation. A subject pairs analysis is simplistic and dangerously misleading (Pollitt, 1996, p. 4).

5.7 Closing comment

The empirical patterns described above – particularly those arising from statistical methods – constitute the only aspect of inter-subject comparability that is not in dispute. Consistently, we find that languages and certain science subjects are the most 'difficult' for students, particularly at A level. However, what those patterns mean is quite another thing. For some, the consistency comes as no surprise, because the statistical methods in use are all underpinned by the same logic and assumptions and will, therefore, tend towards similar outcomes. For others, the consistency underscores the reality of an important phenomenon, one that we simply cannot ignore, even in the face of severe methodological limitations.

6. Conclusion

If we were to take outcomes from statistical methods at face value, then we would conclude that there were major differences in the difficulty of A level and GCSE subjects. The patterns are clear and consistent: science subjects and languages, in particular, seem to be considerably harder than other subjects. Unfortunately, for all the work, both practical and conceptual, that has been undertaken on inter-subject comparability over the decades, we still appear to be no closer to reaching unison over the 'facts' of the matter, that is whether or not certain subjects can justifiably be said to be graded more harshly than others. This is mainly because, although thinking has advanced considerably over time, we still see huge disagreements concerning how best to define and conceptualise inter-subject comparability, let alone how best to monitor it, let alone how best to respond to monitoring outcomes. To the extent that leading scholars from academia and the professions cannot agree over matters of principle, there is currently no hope of responding to inter-subject comparability monitoring outcomes in a way that would satisfy everyone.

From one perspective, the very idea of comparing grade standards across subjects is meaningless. It is not possible to make sense of the idea that a grade B in French is of the same standard as a grade B in chemistry; that is, it is not possible to define inter-subject comparability, period. Clearly, from this perspective, no method could be devised to monitor inter-subject comparability, as the very idea has no meaning.

From another perspective, the idea of comparing grade standards across subjects is not necessarily meaningless, although neither is it at all straightforward. In fact, inter-subject comparability can be given a variety of different meanings, although each meaning will be at least a little 'fuzzy' around the edges and, therefore, disputable. Some researchers prefer to think of inter-subject comparability statistically, reducing comparability to the intuitive idea of students having the same 'chances of success' across subjects. Others prefer to rationalise it more substantively, with reference to linking constructs such as those described below.

Attainment-related linking constructs – which capture elements of knowledge, skills and understanding that are presumed to be common across subjects – can only take us so far when it comes to inter-subject comparability. They may have some plausibility within cognate subject clusters. However, they do not provide a plausible basis for establishing inter-subject comparability across all subjects, because the nature of attainment varies so widely across subject areas. Fortunately, alternative (arguably, more plausible) linking constructs can be envisaged, each providing a different take on the meaning of inter-subject comparability, as the following examples illustrate.

One possible meaning is that students with the same level of 'general intelligence' who study different subjects ought, on average, to end up with the same distribution

of grades. The fuzzy principle upon which this definition might be based is that some students are 'smarter' than others, generally speaking, and should, therefore, be expected to achieve more and to be rewarded with higher grades. In fact, it would be fairly straightforward to monitor inter-subject comparability according to this definition, by using a general intelligence test as a reference instrument. Outcomes could be interpreted fairly directly in terms of differences in grading standards, and could be acted upon straightforwardly. However, by definition, anything beyond general intelligence that students put into the job of learning would not be rewarded with higher grades. If, for instance, art students happened to be more 'gritty' (that is resolute and determined) than business studies students, then this would (intentionally) be prevented from influencing their respective grade distributions.[5]

Another possible meaning is that students with the same level of 'general academic application' who study different subjects ought, on average, to end up with the same distribution of grades.[6] The fuzzy principle upon which this definition might be based is that some students 'apply themselves' to the job of learning better than others, generally speaking, and should, therefore, be expected to achieve more and to be rewarded with higher grades. Students might apply all sorts of things to the job of learning, for instance more study time, more working memory capacity, more concentration, more grit, and so on. In practice, this might be harder to measure than general intelligence. However, some might argue that this common factor is exactly what statistical methods capitalise upon, in order to generate the empirical patterns of results which have been fairly consistently observed. If so, then this could be an argument for taking outcomes from statistical methods more or less at face value. Once again, by definition, anything beyond what students generally apply to the job of learning would not be rewarded with higher grades. If, for instance, biology teachers happened to put more into the job of teaching than PE teachers, then this would (intentionally) be prevented from influencing the grade distributions of biology students and PE students, respectively.[7]

---

[5] Obviously, all other things being equal, the grittier art student would still achieve a higher grade in art than the less gritty art student. Yet, by using a general intelligence test to link standards across subjects, the grittier art cohort would not be rewarded with a higher distribution of grades than the less gritty business studies cohort, again all other things being equal.

[6] 'General academic application' is recommended here, as a less contentious term for what may have been labelled 'general academic ability' in earlier studies.

[7] Furthermore, systematic differences in how students applied themselves across subjects would also not be recognised, as the definition (and the method) is based on the idea of 'general academic application'. In other words, if geography happened to dispose students to work harder and for longer

Yet, another possible meaning is that everything which gets applied to the job of learning – whether attributable to students, their teachers, their resources, or to the subject itself – ought, in appropriate measure, to be recognised in exam grades. The fuzzy principle upon which this definition might be based is that everything which influences learning in a subject ought, in the same measure, to influence subject grades. Intuitively, this might seem to be the fairest of all definitions. However, it is not at all clear how it could be operationalised via a comparability monitoring method. It would simply not be possible to create yardsticks for measuring each and every causal factor, probably not even in principle. As fair as it might seem, in theory, even this 'all causes' definition is disputable. For instance, should attainment at the commencement of a course of learning be considered a legitimate causal determinant? If so, then, for example A level English students would inevitably end up with far higher grades than A level psychology students, simply because students commence A level English from a far higher baseline of knowledge, skills and understanding in their chosen subject area. These issues are all related to the basic set of questions that underpins all statistical research into comparability of standards across subject areas: what is it legitimate to 'control' for, what is it not legitimate to 'control' for, and (most importantly) why?

To conclude succinctly: as far as inter-subject comparability is concerned, there are no straightforward answers; in fact, there is still no clear consensus concerning the kinds of answers that need to be provided.

---

than German, this would (intentionally) be prevented from influencing the grade distributions of geography students and German students, respectively.

7. References

Adams, R. (2007) *Cross-moderation methods*. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms P. (Eds.) *Techniques for monitoring the comparability of examination standards* (pp. 212–245). London, QCA.

Alton, A. and Pearson, S. (1996) *Statistical Approaches to Inter-Subject Comparability*. Report for the Joint Forum of the GCSE and GCE. Unpublished.

Baird, J. (2007) *Alternative conceptions of comparability*. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards* (pp. 124–156). London, QCA.

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

Bramley, T. (2005) *Accessibility, easiness and standards*. Educational Research, 47 (2), pp. 251–261.

Bramley, T. (2007) *Paired comparison methods*. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards* (pp. 246–294). London, QCA.

Bramley, T. (2011) *Developing methods to address comparability issues.* Paper presented at the Perspectives from Cambridge Assessment Seminar: *Comparability of examination standards* on 6th April 2011. London, Cambridge Assessment Network.

Christie, T. and Forrest, G.M. (1981) *Defining Public Examination Standards*. London, Schools Council Research Studies/Macmillan Education.

Coe, R. (2007) *Common Examinee Methods*. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards* (pp. 331–367). London, QCA.

Coe, R. (2008) *Comparability of GCSE examinations in different subjects: an application of the Rasch model*. Oxford, Oxford Review of Education, 34 (5), pp. 609–636.

Coe, R. (2010) *Understanding comparability of examination standards*. Research Papers in Education, 25 (3), pp. 271–284.

Coe, R., Searle, J., Barmby, P., Jones, K. and Higgins, S. (2008*) Relative difficulty of examinations in different subjects.* Durham, Centre for Evaluation and Monitoring, Durham University.

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

Cresswell, M.J. (2000) *The Role of Public Examinations in Defining and Monitoring Standards*. In Goldstein, H. and Heath, A. (Eds.) *Educational Standards* (pp. 69–120). Oxford, Oxford University Press.

Crofts, J.M. and Caradog Jones, D. (1928) *Secondary School Examination Statistics*. London, Longmans.

Dearing, R. (1996b) *Review of Qualifications for 16–19 Year Olds: Full Report*. London, School Curriculum and Assessment Authority.

Dearing, R. and King, L. (2007) *Languages Review*. Nottingham, Department for Education and Skills Publications.

Elliott, G. (2013) *A guide to comparability terminology and methods*. Cambridge, Cambridge Assessment.

Fitz-Gibbon, C.T. and Vincent, L. (1994) *Candidates' Performance in Public Examinations in Mathematics and Science*. London, School Curriculum and Assessment Authority.

Fitz-Gibbon, C.T. and Vincent, L. (1997) *Difficulties Regarding Subject Difficulties: developing reasonable explanations for observable data*. Oxford, Oxford Review of Education, 23, pp. 291–298.

Forrest, G.M. and Vickerman, C. (1982) *Standards in GCE: subject pairs comparisons, 1972–80*. Manchester, Joint Matriculation Board.

Forrest, G.M. (1971) *Standards in subjects at the Ordinary level of the GCE, June 1970*. Manchester, Joint Matriculation Board.

Goldstein, H. and Cresswell, M.J. (1996) *The Comparability of Different Subjects in Public Examinations: A Theoretical and Practical Critique*. Oxford, Oxford Review of Education, 22 (4), pp. 435–442.

Good, F.J. and Cresswell, M.J. (1988) *Grade Awarding Judgements in Differentiated Examinations*. British Educational Research Journal, 14 (3), pp. 263–281.

He, Q. and Eason, S. (2007) *A comparison study on the 2006 GCE outcomes of accounting and archaeology.* Report for AQA. Unpublished.

Jones, B.E. (2003) *Subject pairs over time: A review of the evidence and the issues.* Report for AQA. Unpublished.

Jones, B.E. (2004) *Inter-subject standards: An investigation into the level of agreement between qualitative and quantitative evidence in four apparently discrepant subjects.* Paper presented at the annual conference of the International Association for Educational Assessment on 13th to 18th June in Philadelphia, USA.

Jones, B.E., Philips, D. and van Krieken, R. (2011) *INTER-SUBJECT STANDARDS: AN INSOLUBLE PROBLEM?* Manchester, AQA.

Kelly, A. (1975) *THE RELATIVE STANDARDS OF SUBJECT EXAMINATIONS*. Research Intelligence, 1, pp. 34–38.

Kelly, A. (1976) *A study of the comparability of external examinations in different subjects*. Research in Education, 16, pp. 37–63.

Korobko, O., Glas, C., Bosker, R. and Luyten, J. (2008) *Comparing the Difficulty of Examination Subjects with Item Response Theory*. Journal of Educational Measurement, 45 (2), pp. 139–157.

Lamprianou, I. (2009) *Comparability of examination standards between subjects: an international perspective*. Oxford Review of Education, 35 (2), pp. 205–226.

Malpass, D. (2011) *Is GCSE law too severe? A comparability study.* Report for AQA. Unpublished.

Massey, A.J. (1981) *Comparing standards between AL English and other subjects*. Cambridge, Oxford and Cambridge Schools Examination Board.

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

Murphy, R. (2007) *Common test methods*. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms P. (Eds.) *Techniques for monitoring the comparability of examination standards* (pp. 301–323). London, QCA.

Myers, H. (2006) *The 'severe grading' of MFL grades at GCSE and A level*. London, Association for Language Learning.

Newbould, C.A. (1982) *Subject Preferences, Sex Differences and Comparability of Standards*. British Educational Research Journal, 8 (2), pp. 141–146.

Newton, P. (1997) *Measuring comparability of standards between subjects: why our statistical techniques do not make the grade*. British Educational Research Journal, 23 (4), pp. 433–451.

Newton, P. (2003) *Contrasting definitions of comparability*. Paper presented at the QCA Standards and Comparability Seminar on 3th to 4th April in Milton Keynes.

Newton, P. (2005) *Examination standards and the limits of linking*. Assessment in Education: Principles, Policy & Practice, 12 (2), pp. 105–123.

Newton, P. (2007) *Contextualising the comparability of examination standards*. In Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms, P. (Eds.) *Techniques for monitoring the comparability of examination standards* (pp. 9–42). London, QCA.

Newton, P. (2010a) *Contrasting conceptions of comparability*. Research Papers in Education, 25 (3), pp. 285–292.

Newton, P. (2010b) *Thinking about Linking*. Measurement: Interdisciplinary Research and Perspectives, 8 (1), pp. 38–56.

Newton, P. (2012) *Making sense of decades of debate on inter-subject comparability in England*. Assessment in Education: Principles, Policy & Practice, 19 (2), pp. 251–273.

Nuttall, D.L. (1979) *The myth of comparability*. Journal of the National Association of Inspectors and Advisers, 11, pp. 16–18.

Nuttall, D.L., Backhouse, J.K. and Willmott, A.S. (1974) *Comparability of Standards Between Subjects (Examinations bulletins / Schools Council).* London, Methuen Educational and Evans Bros.

Ofqual (2011) *GCSE, GCE, Principal Learning and Project Code of Practice*. Coventry, the Office of Qualifications and Examinations Regulation.

Pollitt, A. (1996) *The 'difficulty' of A level subjects*. Report for the University of Cambridge Local Examinations Syndicate. Unpublished.

QCA (2008) *Inter-subject comparability studies*. London, QCA.

Royal Society (2008) *Science and mathematics education, 14 –19. A 'state of the nation' report on the participation and attainment of 14–19 year olds in science and mathematics in the UK, 1996–2007.* London, the Royal Society.

Rutter, P. (1994) *The effect of studying A-level mathematics on performance in A-level physics*. Physics Education, 29 (1), pp. 8–13.

Sparkes, B. (2000) *Subject Comparisons – a Scottish Perspective*. Oxford, Oxford Review of Education, 26 (2), pp. 175–189.

Tremain, K. (2008) *An investigation into the comparability of GCE biology with other GCE subjects.* Report for AQA. Unpublished.

Wiliam, D. (1996a) *Meaning and consequences in standard setting*. Assessment in Education: Principles, Policy & Practice, 3 (3), pp. 287–307.

Wiliam, D. (1996b) *Standards in examinations: a matter of trust?* The Curriculum Journal, 7 (3), pp. 293–306.

Wood, R. (1976/1987) *Your chemistry equals my French*. Letter to the Times Educational Supplement on 30th July. Reprinted in Wood, R. (1987) *Measurement and Assessment in Education and Psychology: Collected Papers 1967–1987* (pp. 40–44). Lewes, the Falmer Press.

We wish to make our publications widely accessible. Please contact us at publications@ofqual.gov.uk if you have any specific accessibility requirements.

This publication is available at www.gov.uk/ofqual.

Any enquiries regarding this publication should be sent to us at:

Office of Qualifications and Examinations Regulation

Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346