A Recent History of Regulatory Perspectives on Inter-Subject Comparability in England

ISC Working Paper 5

December 2015

Ofqual/15/5795

Contents

1. Introduction 3
2. The formal (albeit implicit) position 6
3. Reticence to adopt an explicit position 9
4. Informal inclinations 12
   The School Curriculum and Assessment Authority – Dearing Review (1996) 12
   QCA – independent panel (2002) 14
   QCA – independent committee (2004) 15
   QCA – qualitative research (2008) 18
   QCA – modern foreign languages (2008) 19
5. Overview 21
6. References 24

Suggested citation:

Ofqual (2015e) *A Recent History of Regulatory Perspectives on Inter-Subject Comparability in England: ISC Working Paper 5*. Coventry, the Office of Qualifications and Examinations Regulation.

This report was written by Paul Newton (Research Chair).

1. Introduction

Although we are the first official regulator of qualifications and exams in England, a succession of official bodies has supervised the operation of school exams since 1917, when the Board of Education constituted the Secondary Schools Examinations Council to "ensure a general equivalence of standards between the approved examining bodies" (Bruce, 1969, p. 10), which were to offer the new School Certificate and Higher School Certificate exams (predecessors to today's GCSE and A level exams).

During the winter of 1937 and 1938, the Secondary Schools Examinations Council conducted an investigation into the Higher School Certificate exam, with a particular focus on the use of exam results for the purpose of awarding state and local education authority scholarships for the universities. State scholarships were generally not awarded on the basis of excellence in one subject alone, which meant everything depended on the "principles adopted for weighing one subject against another" (Secondary Schools Examinations Council, 1939, pp. 18–19) for students who scored the highest marks across different subjects. The investigation report reflected at length upon the matter:

> 61. How can standards of performance in two subjects be equated?
>
> From an absolute point of view no final theoretical answer can be given to this question in the light of present knowledge. Some working hypothesis must, however, be found, and there are two main lines of approach in use. The first takes the view that it is a reasonable definition of equality of standard, provided the groups concerned are sufficiently large, to say that the work of a candidate x hundredths of the way down the list in, say, History is of equal merit with that of a candidate occupying a similar position in the Mathematics list. The usual methods of standardisation based on percentile ranks are then applied to give two such candidates the same mark. It is to be noted, however, that with these methods of standardisation, even with a given standard curve, the marks allotted to the top few candidates in a given subject are to a considerable extent arbitrary, and it becomes necessary to fall back on the personal judgment of the examiner as to whether the top candidates in his subject are exceptionally good, rather weaker than usual or about normal, and to allot marks to these candidates based, within limits, on his verdict. Thus, in the rather important upper reaches, the method definitely loses that entirely objective character which may make it seem attractive.
>
> 62. For this reason it may be well to recognise from the first that the ultimate decision is bound to rest on the subjective judgment of examiners.

> It may indeed be argued further than this, that the basis of the method outlined in the last paragraph is not one to command general assent. Probably the abler intellects in our Sixth Forms are attracted more strongly to some subjects than to others. This fact is indeed recognised by those Examining Bodies who use percentile standardisation methods, and the curves for those subjects in which a greater degree of selection may be thought to exist are adjusted accordingly. Apart from this it may or may not be true that what is usually regarded as outstanding literary ability is as common in the population as, say, outstanding mathematical ability. In view of such considerations some Examining Bodies prefer a system in which each panel of examiners selects those candidates who in their opinion have done outstanding work in their subject and the relative merit of performances in various subjects is decided by discussion at a general meeting.
>
> 63. Both systems have their disadvantages. […] For instance, in the case of one Examining Body which instructed its examiners to mark candidates who in their subjects were clearly of scholarship calibre with a double star, the Investigators were not satisfied that the considerable disproportion in the numbers of double stars awarded from subject to subject by any means reflected real differences in the intellectual calibre of the groups of candidates concerned. […] For instance, in a case in which the percentile standardisation method was used, the Investigators felt that some candidates had gained Scholarships through being top or nearly top of what happened to be a comparatively weak group, the discretion of the examiners not having been sufficiently exerted against the effects of a routine standardisation (Secondary Schools Examinations Council, 1939, pp. 35–6).

This passage illustrates a number of important issues related to the conundrum of inter-subject comparability. First, this is not a new problem. It has exercised the minds of officials within exam boards and supervisory bodies from the days of the earliest national exams to the present day. Second, morally worrying adverse consequences arising from a lack of inter-subject comparability are not a new phenomenon either. The threats and impacts have certainly evolved over time, but the stakes have always been high. Third, the observation that there is “no final theoretical answer” to the conundrum rings as true today as it did back in 1939. Fourth, any practical system that is adopted in response to the challenge is likely to have serious disadvantages.

When there is no theoretically correct answer to a policy question, and when any practical system for responding to it is likely to have serious disadvantages, it becomes extremely hard to establish a defensible policy. The following report has been written to illustrate how this challenging policy area has been navigated over the past few decades: by successive supervisory bodies – the 'regulators' that operated prior to our establishment; by prominent figures associated with those bodies; and by committees appointed by those bodies.

2. The formal (albeit implicit) position

Since the first codes of practice for GCSE and A level exams were written, during the early 1990s, successive qualifications regulators have never formally required awarding organisations to achieve comparability of standards across the full range of subject areas. However, beyond this formal position – which might reasonably be interpreted as an implicit recommendation not to take steps to align standards across the full range of subject areas – it is possible to detect informal inclinations towards a variety of different perspectives; as bodies, figures and committees have either expressed or implied somewhat different opinions and expectations.

The formal position on inter-subject comparability has been very consistent over the past few decades. The first code of practice for the GCSE was published by the School Examinations and Assessment Council (1993). It explained that:

> 4. The Chief Executive is responsible to the examining group's governing council for: […] ensuring that any necessary action is taken to maintain parity of standards in each subject from year to year across different syllabuses and with other examining groups (p. 1);
>
> 107. […] The prime objectives are the maintenance of grade/level standards over time and across different syllabuses (p. 22).

The School Curriculum and Assessment Authority took over the regulatory remit in 1993, and in its 1994 *Code of practice* it specified exactly the same requirements.[1] Similarly, when the Qualifications and Curriculum Authority (QCA) published the first joint GCSE/GCE code of practice (1998), very little changed:

> 5. The awarding body will appoint a single named person accountable directly to its governing council for ensuring the quality and standards of its qualifications (the accountable officer). In doing so, the awarding body guarantees to the regulatory body that it ensures: i. all necessary action will be taken to maintain parity of standards in each subject and qualification from year to year, across different syllabuses, and with other awarding bodies (p. 1);
>
> 101. […] The prime objectives are the maintenance of subject grade standards over time and across different syllabuses (p. 21).

[1] It used exactly the same wording, with the exclusion of 'level' from the second paragraph.

The most recent code of practice (Ofqual, 2011) expressed exactly the same requirement:

> 1.5 The awarding organisation will appoint a single named person to be accountable directly to its governing council for ensuring the quality and standards of its qualifications (that is, the accountable officer). In doing so, the awarding organisation guarantees to the regulators that it ensures: i all necessary action will be taken to maintain parity of standards in each subject and qualification from year to year, across different specifications and with other awarding organisations (p. 8);
>
> 6.6 The objectives of the awarding committee are to maintain grade standards over time, and to align grade standards across awarding organisations and across different specifications within a qualification type (p. 38).

The primary requirement upon awarding organisations is, and in recent years always has been, to maintain comparability of subject grade standards over time and across different syllabuses (that is specifications). The way in which these requirements have typically been phrased – in terms of 'subject grade standards […] across different syllabuses' – seems to imply that the second clause, 6.6 above, concerns different syllabuses in the same subject area, but not different syllabuses in different subject areas. Presumably, if the more general interpretation had been intended, then this would have been explicitly clarified at some point, which has never happened.

Having said that, the requirement related to comparability across different specifications (for example, 6.6 above) has sometimes been interpreted by awarding organisation officers as a requirement to establish inter-subject comparability (for example, Jones, 2004). So it is fair to conclude that the formal position is ambiguous.

Formally, then, there is no explicit regulatory requirement to align grade standards across subject areas. Furthermore, by implication, if standards happened to be misaligned across subject areas, then the primary requirement to maintain standards over time would mean that any such misalignment should simply be carried forward. Consequently, the lack of regulatory requirement might reasonably be interpreted as an implicit recommendation – perhaps even an implicit requirement – not to take steps to align standards across the full range of subject areas, to the extent that doing so would counteract steps taken to align standards within subject areas.

Incidentally, it is unclear whether successive qualifications regulators may have presumed that standards are generally aligned across subject areas, such that maintaining standards within subject areas would be presumed also to maintain

standards across subject areas, or whether successive qualifications regulators have simply taken no view on the matter.

Turning from the issue of comparability across the full range of subject areas to the even thornier issue of comparability across cognate, that is closely related subjects, the picture seems even less clear. Successive codes of practice have said nothing directly on the topic. It appears only to have been hinted at indirectly, within codes of practice that have been published since 2003. Following the A level grade awarding crisis of 2002, the code of practice was revised to strengthen requirements concerning the role of the accountable officer in signing off grade boundary recommendations from examiners. At the end of a list of evidence that must be considered, paragraph 128 specified:

> The most complete technical and statistical evidence available, including that generated subsequent to the awarding meeting (for example, information from cognate subjects) (QCA, 2003, p. 31).

This particular requirement was repeated in all subsequent editions of the code of practice. It seems to suggest that standards in cognate subjects ought somehow to be similar, although it falls a long way short of stating this explicitly.

3. Reticence to adopt an explicit position

This apparent reticence to adopt an explicit position on inter-subject comparability merits further discussion. Although we have not been able to identify a definitive explanation, it seems reasonable to assume that the absence of positively stated regulatory requirements over the past few decades has reflected an ongoing unwillingness to commit to any particular position on inter-subject comparability in the absence of a straightforwardly defensible position for a qualifications regulator to adopt. Although the lack of a "final theoretical answer" was evident to the Secondary Schools Examinations Council in 1939, the conundrum became even more enigmatic in the wake of important intellectual and political developments that occurred during the mid-1980s.

Intellectually, an influential movement in the USA had argued against the dominance of 'norm-referenced' assessment and in favour of a new approach, 'criterion-referenced' assessment, which judged students in terms of absolute performance criteria, that is their mastery of well-specified achievement domains. Politically, discontent had been voiced over the stability of O and A level pass rates over time – particularly in the context of an increasing percentage of the national cohort being entered for these exams – and this translated into concern that falling educational standards were being concealed. Government insisted that public exams, like the proposed new 16+ exam, which was to become known as the GCSE, should 'no longer' be norm-referenced, and should be criterion-referenced instead. Schools would be charged with raising educational standards, and public exams should, by virtue of criterion-referencing, be capable of measuring those increases accurately. Although this debate arose specifically in the context of the new 16+ exam, it held implications for standards and comparability at A level, and both levels of exam soon became embroiled in the 'norm- versus criterion-referencing' debate.

Importantly, the political discourse around the mid-1980s concerned the need to 'transition' from norm-referencing to criterion-referencing, or 'criteria-referencing' as it was also known. The implication of this discourse was that public exam standards were to be defined quite differently. As a report from the Schools Council (Orr and Nuttall, 1983) explained:

> The meaning of comparability in a decentralized system is, however, open to interpretation. There appear to be two basic notions of comparability: one is that different examinations should test and reward in a similar manner the attainment of the same specified skills; the other is that for a given level of ability, the expectation of a candidate achieving a specific grade should be constant (across syllabuses or across time). […]

> Comparability between different subjects, in contrast, can rest only on the second notion (p. 20).
>
> The development of national criteria for the new system at 16+ represents a move towards the narrow concept of comparability whereby equivalence of grade standards is defined in terms of mastery of skills and concepts (p. 21).
>
> The Government has nailed its colours to the mast: it attaches preeminent importance to the notion of standards defined in terms of specific levels of attainment. This calls for the formulation of grade-related criteria which would specify in some detail, and in subject-specific terms, the levels of attainment necessary for the award of each grade (p. 23).
>
> However, the move towards criteria-referencing, which elevates within-subject considerations to the dominant position in the grading process, once and for all destroys the notion of subject comparability. The inference that a grade C in history is 'as good as' or 'conveys the same information about general competence as' a grade C in geography is probably unsafe at present, but it would be categorically false under the new approach (p. 23).

From a technical perspective, it seemed, during the 1980s, that the transition to criteria-referencing would render meaningless the very idea of inter-subject comparability. Yet, this 'transition' never occurred, for two reasons (see Newton, 2011). First, it was simply not true that public exam standards were norm-referenced prior to the mid-1980s. The political rhetoric was technically incorrect and the approach to maintaining exam standards was always far more subtle than had been portrayed. Importantly, comparability had always been judged in terms of the attainments of students at grade boundary marks. This was not criteria-referencing, in a strict sense, but it was certainly attainment-referencing. Second, the kind of criteria-referencing envisaged by politicians was simply not compatible with public expectations concerning how a public exam system ought to function.

In short, there was no transition from one definition of exam standards and comparability to another. The definition remained essentially unchanged from the 1970s to the 1990s. However, there were certain technical changes to A level grade awarding procedures during the mid-1980s, and both GCSE and A level exams did adopt some of the associated trappings of criterion-referencing, including the development of grade descriptions and an increased emphasis upon the role of examiner judgement in locating grade boundaries. Both of these changes gave the impression of a substantive transition – and this impression allowed the discourse

surrounding criterion-referencing to survive – although, in fact, there was no substantive transition in what the awarding organisations were attempting to achieve when locating grade boundaries.

Given this awkward and confused intellectual and political background, it seems reasonable to conclude that there would have been no straightforwardly defensible position for the School Curriculum and Assessment Authority to adopt during the early 1990s. The School Curriculum and Assessment Authority was responsible for regulating a system that had notionally transitioned from norm-referencing to criterion-referencing, that is to a system in which the very idea of inter-subject comparability was meaningless. In fact, though, this transition had not actually occurred. So, concern for inter-subject comparability could not simply be ruled out on the technicality that exams were now criterion-referenced. On the other hand, as earlier reports from the Schools Council had demonstrated, it was not at all clear how inter-subject comparability concerns were appropriately dealt with under the 'old' system anyhow (see Christie and Forrest, 1981). If anything, to claim that there was no straightforwardly defensible position for the School Curriculum and Assessment Authority to adopt during the early 1990s is a gross understatement.

4. Informal inclinations

Although there have been no positively stated regulatory requirements concerning inter-subject comparability over the past few decades, different regulatory bodies, figures and appointed committees have inclined in different directions, at different points in time, revealing a variety of implicit and explicit preferences for alternative positions.

The following overview does not necessarily capture each and every foray into the territory of inter-subject comparability, although it does cover much of the ground. It focuses upon five key publications, from 1996, 2002, 2004 and (two from) 2008.

The School Curriculum and Assessment Authority – Dearing Review (1996)

Sir Ron Dearing had been Chair of the School Curriculum and Assessment Authority since its establishment in 1993. In 1995, he was invited to review 16–19 qualifications.[2] His interim report identified six issues in relation to A levels, the first of which concerned standards in different subjects. This emphasis was driven by outcomes from previously commissioned research by the School Curriculum and Assessment Authority, which suggested that A level standards might vary across subjects; specifically, that mathematics and science subjects, some modern foreign languages and general studies were more difficult than other subjects, according to analyses based on the A Level Information System (ALIS) database (Fitz-Gibbon and Vincent, 1994). Additional research was undertaken for Sir Ron, including replications of the earlier analyses.

Although the replications identified broadly the same patterns of results, outcomes from additional analyses revealed that the headline results were not always straightforward to interpret. With different assumptions or controls, somewhat different conclusions might be drawn. The subsidiary report on quality and rigour in A levels (Dearing, 1996a) concluded as follows:

> 33 In conclusion, the available evidence appears to suggest the following:
>
> - there are differences in subject 'difficulty' at A level. To what extent this is a new phenomenon is not known. These differences may simply reflect conceptual differences across the subjects themselves. Physics, for example, is not only different in content from geography, but

[2] Although Sir Ron was Chair of the School Curriculum and Assessment Authority at the time, and although the report was published by the School Curriculum and Assessment Authority, this was technically Sir Ron's independently written report.

> requires different skills and almost certainly requires understanding of inherently more difficult concepts (in terms of the general public's understanding of such concepts);
> - there is evidence to show that some of the more 'difficult' subjects at A level (particularly physics, chemistry, and mathematics) are compensated for by relatively lower entry requirements into Higher Education;
> - that students starting courses in these subjects are amongst the most highly qualified (in terms of UCAS points scores);
> - that those institutions which mainly offer courses related to the more 'difficult' A levels tend to recruit students who have achieved higher A level grades generally;
> - in some subjects, history and French for example, Higher Education appears to make no compensation for relative difficulty at A level. This may be a function of supply and demand;
> - mathematics and science degree courses are more difficult to fill than those in some other subjects;
> - 'bridging' courses are offered on entry into higher education to a much greater degree in science and mathematics related subjects than in others (p.9).

In his full report (1996b), Dearing developed issues identified within the subsidiary report. It included the carefully worded conclusion that "there are apparent differences in subject difficulty at A level" (p. 84, emphasis added). It then proposed that the first issue to be debated was "whether this analysis is valid" (p. 85), adding that this would be for the awarding organisations to consider and to discuss with the regulatory bodies. The report noted four possible courses of action, should validity be acknowledged:

1. Level down the difficulty of the most demanding subjects.
2. Raise the level of the easiest subjects.
3. A combination of both.
4. Publicise the differences, but do not change grading standards.

Sir Ron concluded with two recommendations:

> 10.21 I **recommend** that:

> There should be no reduction in the standards required in any subject.
>
> The awarding bodies should review the evidence prepared for this Report and reach agreed conclusions with the regulatory bodies. Where subjects seem decidedly below the 'average' level of difficulty, there should be a levelling-up of demand, after giving advance warning to institutions. Details of the procedure for bringing about this change should be agreed between the regulatory and awarding bodies and be subject to the approval of the Secretaries of State (pp. 85–86).

In other words, although there was an explicit acknowledgement of continuing debate over the interpretation of investigations into inter-subject comparability, there was also an implication that Sir Ron was reasonably persuaded by the evidence that some A levels were more difficult than others and he wanted the awarding organisations to sort out the problem.[3]

The Department for Education and Employment supported this direction of travel. For instance, in a letter to heads of secondary schools in England (11th February 1997), Under Secretary Michael Richardson explained that the Secretary of State had endorsed the School Curriculum and Assessment Authority's proposals to "continue to develop a methodology to reduce differences in standards between subjects" (p. 2, *The Future Shape of GCE AS/A Levels*).

QCA – independent panel (2002)

In 2001, an "independent panel of experts" was invited by the QCA to review the adequacy of the quality assurance systems that were designed to maintain A level standards, under the context that exam achievements had improved steadily over time, raising doubts in some minds about the maintenance of exam standards. The members of the independent panel were all senior academics, two of them having had distinguished careers as assessment researchers:

- Professor Eva Baker, Co-Director of the National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles;

[3] In fact, even Sir Ron acknowledged that it was problematic to interpret conclusions from quantitative analyses of inter-subject comparability for subjects like music and art, "since talent in them is not necessarily closely related to achievement in the generality of GCSE subjects" (Dearing, 1996b, p. 84).

- Dr Barry McGaw, Deputy Director for Education at the Organisation for Economic Co-operation and Development (OECD);
- Lord Sutherland of Houndwood, Vice-Chancellor of the University of Edinburgh.

The independent panel reported in January 2002 (Baker et al.). The report began by claiming that there was "no scientific way in which to determine in retrospect whether standards have been maintained" (p. 4). It went on to discuss challenges such as the feasibility of political expectations of the exam system and the lack of public understanding of its limitations. The panel observed that:

> The question has also been raised about comparability of standards across subjects. To this concern, we can point to the processes that the awarding bodies and QCA use to assure quality. We doubt, however, that there is a method to document comparability of standards (meaning performance) across different subject areas. For one thing, the disciplines themselves differ in terms of the types of prior knowledge, analytical skills and requirements for expression. Second, the various subjects draw to them students with very different interests and capacities (p. 14).

Although the panel claimed that there was little value in attempting to compare the performances of students in different subjects to discern differences in standards, it did recommend comparing syllabuses, exams and mark schemes across subjects to compare "the task demands they require of the student" (p. 14). The precise intention of this suggestion seems, in retrospect, a little unclear. It was expressed slightly differently in a subsequent recommendation:

> Conduct qualitative analyses, in two subjects, of a series of examinations and resulting scripts detailing content and cognitive requirements. Judge comparability within and between subjects of the demands of the examinations and the standards of performance expected of students (p. 24).

QCA – independent committee (2004)

In 2003, an independent committee was established to consider the maintenance of A level exam standards, following a recommendation from the final report of the Tomlinson inquiry into the 2002 Curriculum 2000 crisis (Tomlinson, 2002). Tomlinson, incidentally, seems to have assumed that there was an obligation upon awarding organisations to establish inter-subject comparability:

> 22. The design implications arising from this have been significant in the recent developments of the A level system. Confident matching of students' performance to a fixed standard requires not just accurate and consistent marking but also the ability to define and maintain standards of grading from year to year and between subjects, and the means to judge the performance of individual candidates against that standard (p. 11).

Again, members of the independent committee were all senior figures in education, two of them having had distinguished careers as assessment researchers:

- Dr Barry McGaw, Director for Education at the OECD;
- Professor Caroline Gipps, Deputy Vice-Chancellor of Kingston University;
- Mr Robert Godber, former Headteacher, Wath upon Dearne Comprehensive School, Rotherham.

Their remit was to focus on a full range of comparability concerns, and on the QCA's regulatory role in maintaining standards. Their conclusions were generally favourable, suggesting that both awarding organisations and the QCA had robust systems in place (McGaw et al., 2004). However, they highlighted unrealistic public expectations concerning the possibility of maintaining exam standards over long periods in time. And, on inter-subject comparability, they noted that:

> No examination system has found an adequate way to determine whether standards are constant across subjects.
>
> Neither QCA nor the awarding bodies have any strategy for determining whether examination or performance standards are constant across subjects. The only examination systems that we know of which seek to achieve this kind of comparability are those in the Australian States and Territories. They achieve this by making the rather heroic assumption that all examinations are measuring essentially a common dimension and then express all results on a common scale. We described this approach briefly in the section 'Reviews of standards across subjects' (p.20) where we also explained why it could not be applied to A level examinations even if it were thought desirable (pp. 30–31).

Incidentally, the independent committee observed that, following recommendations from the Baker et al. (2002) report, the QCA had initiated a number of investigations into inter-subject comparability within cognate subject areas, based on the judgement of subject experts.

Although the independent committee's report, like the earlier one, was intended to be entirely independent, it appears to have been received favourably by the QCA:

> Ken Boston, chief executive of the QCA said: 'I am pleased by the committee's conclusions about the examination system in terms of the extent to which it is both carefully managed and robust. I am pleased to note the conclusions covering QCAs [sic] success in managing at least as well as any country does some challenging requirements of the examination system where there are high expectations' (Curtis, 2004).

Yet, the report was received with scepticism by some newspapers, as the opening sentence from an article in the Telegraph illustrates:

> It is impossible to tell whether A-level standards have been maintained, according to a report published yesterday by the Qualifications and Curriculum Authority, the 'guardian of the nation's education standards' (Clare, 2004).

Indeed, in the concluding sentence of the same article, Tim Collins, then opposition Tory education spokesman, was quoted as saying:

> It is disgraceful for genuine concerns about exam standards to be dismissed as pointless by the very organisation supposed to be upholding them.

Although most of the article had focused on the maintenance of standards over long periods of time, the same dissatisfaction might well have been expressed in relation to the committee's treatment of inter-subject standards. A subsequent report by Civitas (de Waal and Cowen, 2007) expressed the same kind of scepticism in response to the committee's warnings concerning unrealistic comparability expectations.

The sceptical reception of this report demonstrates that it can be hard for a qualifications regulator to be associated with a position which is easily caricatured as overly academic or indecisive on matters of public concern. Having said that, the QCA continued its tradition of collaborating with academic experts on matters of comparability, subsequently publishing *Techniques for monitoring the comparability of examination standards* (Newton et al., 2007). Robert Coe, from the Centre for Evaluation and Monitoring at Durham University (and a prominent critical voice on

standards over time, and between subjects) was commissioned to write chapter 9 of this book, on common examinee methods, which focused upon techniques for monitoring inter-subject comparability. He situated his analysis in the context of alternative interpretations and uses of exam results, that is, alternative definitions of inter-subject comparability.

QCA – qualitative research (2008)

Although the QCA reports occasionally mentioned perceptions of subject difficulty amongst teachers and students, notably as a possible contributory factor underlying trends in the uptake of subjects over time (for example, QCA, 2005a, 2005b, 2005c), they rarely commented on the basis for any such perceptions. An important exception to this was the programme of research that had been initiated by the Baker et al. review, which culminated in February 2008 with the publication of six reports:

1. overall report (QCA, 2008a);
2. geography versus history – GCSE, AS and A level (QCA, 2008b);
3. sciences – GCSE, AS and A level (QCA, 2008c);
4. biology, psychology and sociology – A level only (QCA, 2008d);
5. English literature, history and media studies – A level only (QCA, 2008e);
6. forms used for analysis of exam materials and student work (QCA, 2008f).

Each study involved an evaluation of the demands implied by the syllabus materials and a comparison of students' work. The overall report was fairly critical of quantitative approaches to investigating inter-subject comparability, arguing that they required "a particular interpretation of examination results, one in which they are all essentially measures of aptitude or general ability" (QCA, 2008a, p. 8).

The research reports presented results from a qualitative approach that "almost inevitably" (QCA, 2008a, p. 9) ruled out comparisons across unrelated subjects, for example French versus chemistry. The student work reviews – which were fundamental to the question of inter-subject comparability – were reported with certain caveats, including the observation that:

> The judgements that reviewers were being asked to make were very complex in any case, much more so than in any comparability study looking only at standards within a subject (QCA, 2008a, p. 41).

Consequently, only cases where perceived differences in standards were large were reported. Even here, reviewers were often unclear as to whether such differences

were really attributable to exam standards, often questioning their own ability to account for differences in exam difficulty across subjects.

The research was released by the QCA with quotes from Ken Boston, its Chief Executive, which included:

> Our report on inter subject comparability shows that the level of demand is broadly comparable across the subjects considered in the studies. This work, which uses a pioneering technique, will be the beginning of a much broader programme of work on inter-subject comparability that the new regulator, Ofqual, will take forward.

Schools Minister Jim Knight was reported to have reaffirmed that the "level of demand is broadly comparable across the subjects considered in the studies", whilst acknowledging that comparing subjects was "complex" (metrowebukmetro, 2008).

Ken Boston was also reported to have said that:

> Ensuring comparability between examinations, awarding bodies and across subjects is vital to maintaining standards (London Evening Standard, 2008).

QCA – modern foreign languages (2008)

Also published in February 2008 was a separate report on standards in modern foreign languages, which followed the publication of the *Languages Review* by Ron Dearing, now Lord Dearing, and the expression of concerns over the difficulty of language exams by the Association for Language Learning (see Dearing and King, 2007; QCA, 2008g). The *Languages Review* indicated that Lord Dearing remained unpersuaded that languages were appropriately graded:

> In our further consultation we have found strong confirmation of the view that the award of grades [for languages] is more demanding than for most other subjects. 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 (Dearing, 2007, p. 12).

In response to these concerns, the QCA argued that quantitative approaches to investigating inter-subject comparability, which the Association for Language Learning had relied upon to make its case, were based upon an aptitude-based conception of comparability which "holds that two examinations may be seen as comparable if students of a certain 'calibre' have an equal chance of achieving a particular grade in any examination" (QCA, 2008g, p. 5). The QCA report admitted that the alternative, the attainment-based conception of comparability, was, "less straightforward when comparing standards across subjects […] since it requires us to think in terms of general kinds of knowledge, skills and understanding" (QCA, 2008g, p. 4), yet it was still the, "QCA view of comparability" (QCA, 2008g, p. 4).

It is important to notice echoes of the distinction that Orr and Nuttall had drawn, in 1983, between different ways of defining comparability. In the words of the Schools Council report, the 2008 QCA report had observed that standards, nowadays, were defined in terms of the "attainment of the same specified skills" (Orr and Nuttall, 1983, p. 20). In 2008, this attainment-based conception of comparability was not the extreme version to which Government had aspired 15 years earlier, which required the strict application of grade-related criteria, but it seemed far closer to this alternative than to an aptitude-based one.[4]

[4] The report also noted a variety of technical challenges, if one was to adopt an aptitude-based conception.

5. Overview

Over the past few decades, the regulatory bodies in England have made relatively few statements concerning inter-subject comparability. However, beyond their formal position – the implicit recommendation/requirement not to take steps to align standards across the full range of subject areas – different regulatory bodies, individuals and appointed committees have inclined in different directions, at different points in time, revealing a variety of more informal preferences for alternative positions.

Lord Dearing, Chair of the School Curriculum and Assessment Authority during the mid-1990s, was prepared to acknowledge that there were legitimate concerns over the interpretation of quantitative analyses, particularly as voiced by awarding organisation researchers. The School Curriculum and Assessment Authority committed to debate these issues with the awarding organisations, and debate ensued (for example, Goldstein and Cresswell, 1996; Fitz-Gibbon and Vincent, 1997; Jones, 1997; Newton, 1997). However, there was also a strong sense of suspicion in the 1996 Dearing Review that exam standards were not comparable between subjects. Incidentally, Lord Dearing appeared to be no less suspicious a decade later, when the *Languages Review* was published.

As the debate that had been initiated by the Dearing Review subsided, towards the end of the 1990s and the beginning of the 2000s and without clear resolution, a somewhat different zeitgeist appeared to be crystallising, in sympathy with conclusions from independent expert groups, which had been appointed by the QCA. This perspective was inclined to be far less convinced by conclusions based upon quantitative analyses.

Both the independent panel (2002) and the independent committee (2004) were very sceptical of the idea of inter-subject comparability, per se. Many, like opposition Tory education spokesman Tim Collins, may have presumed that the QCA endorsed its reports and its general scepticism concerning the possibility of straightforwardly ‘solving’ the problem of inter-subject comparability. The QCA certainly accepted recommendations from the independent panel to explore the potential of qualitative approaches and to question conclusions from quantitative ones. The new focus upon qualitative approaches necessarily restricted comparability conclusions to groups of cognate subjects, rather than to the full range of subjects.

Interestingly, in the period between these two independent reports, John Dunford, the General Secretary of the Secondary Heads Association, initiated a high-profile public debate on inter-subject comparability, following the 2003 A level results day. He associated the rising A level pass rate with rising entries for non-traditional subjects –

such as business studies, law, media, film and TV studies, and psychology – which he claimed were easier than traditional subjects. He described the lack of comparability between subjects as a "hidden scandal", which provoked a swift response from both Government and the QCA. Schools Minister David Miliband insisted that, "every A-level subject meets rigorous standards", and the Department for Education and Skills published tricky-sounding A level psychology questions to prove it (Woodward, 2003). Ken Boston insisted that:

> There is no such thing as an 'easy' A-level.
>
> The A-level is a world class qualification and standards are maintained year after year in all subjects. […] Some students will have a greater aptitude for certain subjects than others, but the standard remains the same (Curtis, 2003).

In other words, Ken Boston seemed far from agnostic about inter-subject comparability. Quite the reverse; he seemed to be insisting that standards were comparable.

When the QCA's qualitative research concluded, in 2008, its reports were still very critical of quantitative approaches. This was formalised in the claim that conclusions from quantitative approaches were based on a particular conception of comparability – one that the QCA did not recognise. (This perspective was even more explicit in the modern foreign languages report.) Although the research reports were fairly circumspect, acknowledging that definitive conclusions could not be drawn from the research, QCA press releases did quote Ken Boston as claiming that levels of demand were "broadly comparable" across subjects.

To summarise, beyond the formal position, it is possible to detect a variety of implicit and explicit preferences for alternative positions on inter-subject comparability. At different times, these differing perspectives have appeared to:

- support the idea that standards may be pitched at different levels in different subjects (as voiced by Lord Dearing);
- deny the idea that standards may be pitched at different levels in different subjects (as voiced by Ken Boston);
- reject the possibility of being able to judge whether or not standards are pitched at different levels in different subjects (as voiced by the independent expert groups).

It is, therefore, not at all surprising that the formal position has sometimes been interpreted (by senior leaders in education and by awarding organisation officials alike) as an expectation that comparability should be engineered across the full range of subject areas.

The variety of informal inclinations identified above helps to illustrate the absence of a ‘right answer’ to the inter-subject comparability question, and the limited plausibility of alternative ‘working hypotheses’, which have been advanced over the years. It hints at the possibility that there may still be no straightforwardly defensible position for us to adopt.

6. References

Baker, E., McGaw, B. and Sutherland, S. (2002) *Maintaining GCE A level standards: the findings of an independent panel of experts*. London, QCA.

Bruce, G. (1969) *Secondary School Examinations: Facts and Commentary*. London, Pergamon Press.

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

Clare, J. (2004) *Public 'expects too much' from exams*. The Telegraph, 4th December. Available at: www.telegraph.co.uk/news/uknews/1478185/Public-expects-too-much-from-exams.html (accessed June 2015).

Curtis, P. (2003) *QCA comes under fire over 'easier' A-levels*. The Guardian, 14th August. Available at: www.theguardian.com/education/2003/aug/14/schools.alevels20035 (accessed March 2015).

Curtis, P. (2004) *English exams 'still best in the world'*. The Guardian, 3rd December. Available at: www.theguardian.com/education/2004/dec/03/schools.gcses2004 (accessed June 2015).

de Waal, A. and Cowen, N. (2007) *The Results Generation*. London, CIVITAS: The Institute For The Study Of Civil Society. Available at: www.civitas.org.uk/pdf/resultsgeneration.pdf (accessed June 2015).

Dearing, R. (1996a) *Review of Qualifications for 16–19 Year Olds: Quality and Rigour in A Level Examinations*. London, School Curriculum and Assessment Authority.

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.

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.

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, pp. 435–442.

Jones, B.E. (1997) *Comparing Examination Standards: is a purely statistical approach adequate?* Assessment in Education: Principles, Policy & Practice, 4 (2), pp. 249–263.

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.

London Evening Standard (2008) *Some GCSEs 'easier than others'*. The London Evening Standard, 22nd February. Available at: www.standard.co.uk/newsheadlines/some-gcses-easier-than-others-6692792.html (accessed March 2015).

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

metrowebukmetro (2008) *Some GCSEs 'easier than others'*. The Metro, 22nd February. Available at: http://metro.co.uk/2008/02/22/some-gcses-easier-than-others-5167 (accessed March 2015).

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. 433–449.

Newton, P. (2011) *A level pass rates and the enduring myth of norm-referencing*. Research Matters, Special Issue, 2, pp. 20–26.

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

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

Orr, L. and Nuttall, D. (1983) *Determining Standards in the Proposed Single System of Examining at 16+, Comparability in Examinations Occasional Paper 2*. London, Schools Council.

QCA (1998) *GCSE and GCE A/AS Code of Practice*. London, QCA.

QCA (2003) *GCSE, GCSE in vocational subjects, GCE, VCE and GNVQ code of practice 2002/3.* London, QCA.

QCA (2005a) *Geography: 2004/5 annual report on curriculum and assessment*. London, QCA.

QCA (2005b) *Mathematics: 2004/5 annual report on curriculum and assessment*. London, QCA.

QCA (2005c) *Modern foreign languages: 2004/5 annual report on curriculum and assessment*. London, QCA.

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

QCA (2008b) *Inter-subject comparability studies: Study 1a: GCSE, AS and A level geography and history.* London, QCA.

QCA (2008c) *Inter-subject comparability studies: Study 1b: GCSE, AS and A level sciences.* London, QCA.

QCA (2008d) *Inter-subject comparability studies: Study 2a: A level biology, psychology and sociology*. London, QCA.

QCA (2008e) *Inter-subject comparability studies: Study 2b: A level English literature, history and media studies.* London, QCA.

QCA (2008f) *Inter-subject comparability studies: Forms used for analysis of examination materials and candidate work.* London, QCA.

QCA (2008g) *Grade standards in GCSE modern foreign languages*. London, QCA.

Secondary Schools Examinations Council (1939) *The Higher School Certificate Examination.* Being the report of the panel of investigators appointed by the Secondary Schools Examinations Council to enquire into the eight approved Higher School Certificate exams held in the summer of 1937. London, HMSO.

School Examinations and Assessment Council (1993) *GCSE Mandatory Code of Practice*. London, School Examinations and Assessment Council.

Tomlinson, M. (2002) *INQUIRY INTO A LEVEL STANDARDS. Final Report.* London, Department for Education and Skills.

Woodward, W. (2003) *Heads hit out at easy courses as A-level passes rise again*. The Guardian, 14th August. Available at: www.theguardian.com/uk/2003/aug/14/politics.alevels2003 (accessed March 2015).

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