Special Issue 2: Comparability

October 2011

Research Matters

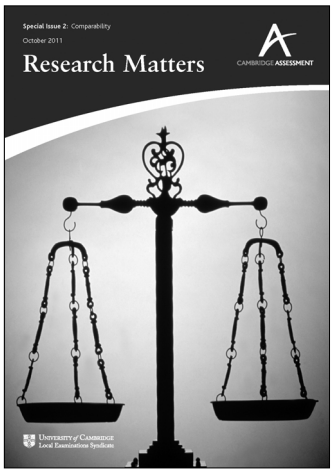
CAMBRIDGE ASSESSMENT

University of Cambridge
Local Examinations Syndicate

CAMBRIDGE ASSESSMENT

Citation

Articles in this publication should be cited as: Elliott, G. (2011). A guide to comparability terminology and methods. *Research Matters: A Cambridge Assessment Publication*, Special Issue 2, 9–19.



If you would like to comment on any of the articles in this issue, please contact Tom Bramley.
Email:
researchprogrammes@cambridgeassessment.org.uk

This issue and previous issues of *Research Matters* are available on our website:
http://www.cambridgeassessment.org.uk/ca/Our_Services/Research

Research Matters:
Special Issue 2: Comparability

A CAMBRIDGE ASSESSMENT PUBLICATION

Foreword

Comparability is an area beset by assumptions, trammelled by methodological dispute, and regarded, by some, as a bankrupt pursuit. For the public, authentication of claims that the standards, in the 'same' qualification in different years, and in different qualifications which claim equivalence, is vital. But this need for confidence frequently is accompanied by ill-grounded assumptions. It may feel self-evident that standards should, in all circumstances, be maintained over time, but standards do NOT have to be maintained over time, with precision, in *all* assessments. They may have to be in the case of qualifications such as A levels. Here, different candidates, who have gained their qualifications in different years, are competing for university places in the same year. In these admissions processes, an 'A' grade from different sessions is treated as if it means the same thing – so it needs to mean the same thing. Contrast STEP assessment – a maths admissions test. STEP is a purely norm-referenced assessment. The standard is not exactly the same each year. But in each year it remains a highly effective instrument for rank ordering candidates – and thus helps universities identify the top performing pupils. But the consequence of this is that a pupil cannot use a score gained in 2009 in 2011 admissions processes – if applying in 2011, they need to take the test again (not a bad idea in fact…have they retained their mathematical understanding?). Yet another example calls many assumptions into question: there are instances of standards in qualifications deliberately being varied, over time, in order to achieve change in education systems. Here, comparability can be used not to secure equivalence, but to understand the magnitude of change. Is comparability work always concerned principally with standards? No; in some cases comparability can be focused on qualitative comparison of the *focus, structure or purpose of different qualifications*. So the assumptions in comparability need clarification.

So too the methods. Methods are hotly contested and, again, beset by assumptions. Some studies have not considered the quality of work produced by candidates. Others consider candidate work but take so long and are so expensive that the results of the analysis come too late to inform action, and draw resource away from more pressing research. Yet others yield interesting results, but appear compromised by assumptions that teaching has not improved, that successive cohorts of candidates have remained equally motivated, and so on. Perhaps comparability should indeed be seen as an analogue of the pursuit of the Holy Grail – ultimately fruitless in attaining the assumed goal, but the real worth is in striving to attain it. In other words, the fact that comparability work is unlikely to yield flawless conclusions of ultimate precision does not mean that we should not strive ceaselessly to improve our techniques.

This edition emphasises the need to be clear about definitions of comparability, to specify precise objectives, to be discriminating in respect of choice of method, and to understand the utility and limitations of findings. Our conclusion is that comparability is NOT a bankrupt activity. It is complex, demanding (both theoretically and practically), and frequently produces indicative rather than definitive findings. But it remains a vital part of both management of, and research on, qualifications and assessments. For as long as we and others make claims about equivalence, we need to strive to put in place robust, practical, and cost-efficient processes for understanding and authenticating those claims. Dark art or science? If science involves review, critique and refinement of method, alongside constant reconceptualisation and redefinition, then yes, it's science.

Tim Oates *Group Director, Assessment Research & Development*

Editorial

Tom Bramley *Assistant Director, Research Division, Assessment Research & Development*

In 2007 QCA published the book 'Techniques for monitoring the comparability of examination standards', the purpose of which was not only to provide a review of the comparability research carried out since a previous review in 1985, but also to describe and evaluate in more depth the different methodological approaches used to investigate comparability. Since that publication, the profile of comparability research has remained high with the formation of the qualifications and examinations regulator Ofqual, which in 2010 began a programme to compare the demand of qualifications and assessments internationally, focusing initially at pre-university level.

Given the central importance of comparability and standards to all aspects of Cambridge Assessment's work across its three business streams (OCR, CIE and ESOL), in 2008 Cambridge Assessment set up a Comparability Programme with three full time dedicated members of staff and an associated governance group in order to contribute to, and maintain an overview of, the wide range of comparability work carried out across the Group.

In this Special Issue of *Research Matters* we present some of Cambridge Assessment's recent thinking about comparability. In the opening article, Gill Elliott, leader of the Comparability Programme, gives an historical overview of comparability concerns showing how they have been expressed in different political and educational contexts in England over the last 100 years.

It has become increasingly clear that comparability research is bedevilled by a lack of clarity over the meaning of its most basic concepts and inconsistent use of terminology. The second article, also by Gill Elliott, identifies and defines some widely used terms and shows how different methods of investigating comparability can be related to different definitions.

A topic of perennial interest is the inexorable rise over the last 25 years in the percentage of students passing, or achieving A grades in, A level examinations. In the third article, Paul Newton, Director of the Cambridge Assessment Network, tries to find evidence to support the popular (mis)-conception that A levels used to be norm-referenced but became criterion-referenced, and that this change was responsible for the rising pass rate.

Another topic of recurring interest is whether, within a qualification type (e.g. GCSE or A level), subjects differ in difficulty. It always seems to have been easier to calculate indices of relative subject difficulty than to explain exactly what they mean. A recent approach has been to use the techniques of Item Response Theory, treating different exam subjects like different questions (items) on a test. In the fourth article I discuss whether this analogy works.

It is an unavoidable fact of comparability research that often there is a need to compare things that are in many ways very different, such as vocational and academic qualifications. A sensible basis for comparison needs to be found, and in the fifth article Jackie Greatorex discusses one such basis – 'returns to qualifications' – that has so far been relatively rarely used by researchers in awarding bodies. The appendices to her article (pp.39–41) include two glossaries, one of qualification types and one of assessment terms, which readers unfamiliar with the acronyms and jargon of assessment in England may find useful.

In the world of language testing, comparability is perhaps best conceived as an aspect of validity – that is, comparability of inferences that are justified about the communicative competence of individuals with certificates from different language testing organisations. In order to bring some coherence to a potentially conflicting area, the Common European Framework of Reference for Languages (CEFR), published in 2001, was devised, with consequent political and commercial pressure for language testing organisations to map their own tests to the proficiency levels in this framework. In the sixth article Neil Jones, Assistant Director of Research & Validation for Cambridge ESOL, discusses some of the conceptual issues involved in linking tests to the CEFR.

Frequent change has been a constant feature of school examinations in England for many years. The most recent innovation, which appears at the time of writing likely to be short-lived, is the 'unitisation' of GCSE examinations. Whereas GCSEs were formerly taken 'linearly' by students aged 16 at the end of a two year course, now the different 'units' can be taken at various stages throughout the course. This naturally presented a great challenge to the exam boards to ensure that the outcomes on the first large-scale award of the new unitised examinations in June 2011 were in some sense comparable to those on the old linear ones. In the seventh article, Mike Forster, Head of Research & Technical Standards at OCR, describes some of the issues that arose, and research undertaken by OCR in order to develop guidelines for grading procedures in 2011 that would be capable of achieving comparability.

I suspect that few researchers in comparability would deny that the audience for their academic articles is relatively small, comprising a fairly closed circle of individuals writing mostly for each other's benefit. Many comparability stories that make the headlines, on the other hand, come from outside academia. The final article, by Nicky Rushton, Matt Haigh and Gill Elliott, takes an interesting step away from the academic literature on comparability to discuss how comparability issues are presented in the media, and to evaluate the contribution that programmes like "That'll Teach 'em" can make to our understanding of comparability and standards.

It is our hope that this Special Issue will be both thought-provoking and informative, and a useful point of reference for anyone interested in the complex and challenging issues of comparability of examinations and qualifications.

Acknowledgements

I would like to thank the following people* for reviewing the articles in this Special Issue: Beth Black, Robert Coe (CEM Centre, Durham University), Vicki Crisp, Mark Dowling, Gill Elliott, Mike Forster, Jackie Greatorex, Sylvia Green, Matt Haigh, Neil Jones, Paul Newton, Tim Oates, Helen Patrick (independent consultant) Nick Raikes, Stuart Shaw, Irenka Suto, and Carmen Vidal Rodeiro.

Tom Bramley
Cambridge Assessment, October 2011

*Cambridge Assessment unless otherwise stated.

100 years of controversy over standards: an enduring problem

Gill Elliott Head of Comparability Programme, Assessment Research & Development

Why are we so bothered about comparability in public examinations? The issue has been a thorn in the sides of educational commentators for at least a century and, despite numerous attempts to solve it, remains a stubborn problem.

This article introduces some of the key issues within the field of comparability, and provides an historical perspective on some of the current concerns. It traces major developments in the theory, methodology and use of comparability research and looks at the way in which theories of comparability have developed and different viewpoints have emerged.

In 1911 a Consultative Committee was convened by the Board of Education to report on *Examinations in Secondary Schools*. What were the comparability-related issues, and how were those issues described and defined?

The 1911 report contained a list of the functions which examinations, in their widest sense, were expected to fulfil at that time. They were expected to:

- test the ability of the candidate for admission to practice a profession;
- ascertain the relative intellectual position of candidates for academic distinction (e.g. scholarships);
- be used for recruitment to the public (civil) service;
- test the efficiency of teachers;
- diffuse a prescribed ideal of liberal[1] culture ("for the efficient discharge of the duties of citizenship in the more responsible positions in life, or as the necessary qualification for admission to University, every young man should be required to have reached a prescribed standard of all round attainment in liberal studies").

This list is still relevant, given the importance of understanding the purposes to which the results of assessments are put when interpreting claims of comparability.

The report outlined the large number of organisations that were providing examinations for the purposes of matriculation and/or progression into the professions. These included not only universities, but also trade organisations and professional societies, such as the London Chamber of Commerce, the Pharmaceutical Society, the Institution of Civil Engineers and the Incorporated Society of Accountants and Auditors. Whilst the many organisations that required examinations still wanted to preserve their own examinations, in 1911 there was beginning to be a move towards recognition of other examinations as equivalent to their own. The document described a system of equivalents being in place, whereby some organisations were prepared to accept alternate examinations of a corresponding standard to their own. However, as the document went on to report, the system was dogged by innumerable confusing restrictions imposed by the various organisations. The main consequence of the restrictions placed upon the system of equivalents was that the students' choices became very complicated, with an increasing chance of making a poor choice of examination. The document describes it thus:

> *While candidates can obtain their Oxford Senior Certificate by passing in five subjects, no one set of five subjects is accepted by all the exempting bodies. A candidate would have to pass in eleven subjects, viz., Arithmetic, English, Mathematics Higher Geometry, Latin, Greek, English History, Geography, French or German, Chemistry or Physics, and a portion of New Testament in Greek , to be sure that his certificate would be accepted by all the bodies who accept the Oxford Senior Certificate as qualifying a candidate for exemption from their Matriculation or Preliminary Examination. If he only passed in the five subjects required by one particular body, and then for any reason changed his plans... he might find it quite useless to him...*
> (Examinations in Secondary Schools, p.34)

Furthermore, a number of awards simply were not accepted as equivalent:

> *There are at the present moment a large number of external examinations in Secondary Schools, the certificates of which, regarded as entrance qualifications to the various Universities and professional careers, cannot be said to be always accepted as yet as valid equivalents.*
> (Examinations in Secondary Schools, p.38)

Additionally, there was the difficulty of students who had not yet decided upon a career path and needed a more general qualification, which did not exist. Generally these students took two or even three of the available certificates in order to prepare for a variety of paths. However, the document suggested that this approach might have been slightly unfair, in that it gave these students the option to use the best of their performances.

In 1911 the problem of providing access to the examination to the less able students whilst adequately testing the more able was firmly on the agenda. However, the committee was optimistic about the ability of the system to accomplish this without compromising comparability.

> *The levels of attainment reached by different pupils at any one age will of course always differ widely, and it is not supposed that any one set of examination papers will be equally appropriate for them all. But there should be no insuperable difficulty in arriving at a standard which the average pupil should reach at a stated age, and taking this as the criterion by which alternative examinations should be gauged.*
> (Examinations in Secondary Schools, p.90)

1 Liberal in this context can be defined as 'general broadening'.

The 1911 committee advocated a closer relationship among awarding bodies, and between awarding bodies and the schools, and that the 'standard' be fixed on purely educational grounds. In expanding on the latter point, the report blamed the isolation of awarding bodies from each other for many of the problems and for the fact that even when schools of similar type were compared, standards from different awarding bodies were found to be different (according to a very broad definition of 'standards').

The 1911 report highlighted a number of comparability issues, in particular the problem of aligning standards among Awarding Bodies and the problem of adequately providing a system which would allow some students to qualify for entrance to Universities and professions and others to attain a more general qualification.

The committee proposed a system to accommodate these needs which incorporated two examinations – the School Certificate ("breadth without specialism") and the Higher School Certificate (less general and geared more to the needs of Universities and certain professions). However, they also considered a situation where the former examination could serve a dual purpose – a certificate of general ability, plus a distinction level if certain conditions were met. The rationale behind this was explained in a Board of Education circular (1914), quoted in Norwood (1943), *Curriculum and Examinations in Secondary Schools*:

> *(iv) The standard for a pass will be such as may be expected of pupils of reasonable industry and ordinary intelligence in an efficient Secondary School.*
>
> *(v) If the examination is conducted on the principle of easy papers and a high standard of marking, the difference between the standard for a simple pass and that required for matriculation purposes will not be so great as to prevent the same examination being made to serve, as the present school examinations do, both purposes; and with this object a mark of credit will be assigned to those candidates who, in any specific subject or subjects, attain a standard which would be appreciably higher than that required for a simple pass.*
>
> (Curriculum and Examinations in Secondary Schools, 1943, p.27)

It is interesting to note how succinctly these criteria are described, compared with those of today. It is clear that the 'standard' in 1943 was embedded in the notion of the norm.

The following selection of quotations from the Norwood Report (1943) explain how this system began to fall apart.

First, it proved difficult to meet the two distinct purposes of the examination at the same time. The needs of scholars seeking matriculation took precedence, in practice, over those looking for more general certification of educational attainment.

> *Whether there was any chance of these two purposes being achieved simultaneously without one obscuring the other is open to doubt; it is easy to be wise after the event; but the history of the examination has shown that the second purpose rapidly overshadowed the first.*
> (Curriculum and Examinations in Secondary Schools, 1943, p.27)

The Higher Certificate began to present problems because it had been based upon assumptions that the numbers of candidates would be small and the link with universities close. These assumptions proved mistaken. According to the Norwood Report the certificate became increasingly popular, and attracted increasing numbers of students. This led to new courses being added to accommodate the needs of an increasingly diverse body of students. These courses fitted less closely to the original conception of the system where the curriculum was closely linked to needs of students seeking a qualification for matriculation.

> *…Yet its very success has tended to bring about its progressive disintegration. Rapidly winning recognition on all hands, the certificate awarded on the examination has gathered more authority and more significance than was ever intended at the outset, till it has become a highly coveted possession to every pupil leaving a Secondary School. As the curricula of schools have widened to meet the needs of a Secondary School population rapidly growing more diverse in ability and range of interests, the original structure of the examination has changed. Subjects have necessarily been multiplied, whether susceptible to external examination or not; rules which were framed to give a unity to the curriculum tested by examination have been relaxed. Secondary education has become too varied to be confined within a rigid scheme; teachers are becoming too enterprising to be hedged in by set syllabuses, and subjects themselves are gaining in independence and resourcefulness.*
> (Curriculum and Examinations in Secondary Schools, 1943, p.32)

Nevertheless, the Norwood Report was unequivocal about the continuing importance of comparability:

> *…If a test is to carry any weight outside the school, there must be some approximation to uniformity of standard in assessing attainment. The test and the verdict must be objective, and conditions must be equal; there can be no prejudice and no favouritism as between school and school or pupil and pupil. Employers, parents and Professional Bodies need the Certificate; employers ask for a disinterested assessment, and would not be satisfied with a Head Master's certificate; parents look for something which will be a hall-mark of their children, valid wherever in the country they may go.*
> (Curriculum and Examinations in Secondary Schools, 1943, p.31)

Changing use of terminology

Before moving on to discuss how our understanding of comparability has progressed since the 1911 report and the Norwood report, it is important to look at definitions of terms. The 1911 and 1943 reports used three of the key terms used currently, shown below, together with their Concise Oxford Dictionary (COD) (Allen, 1992) definition:

- standards: degree of excellence required for the particular purpose
- equivalence: equal in value, amount or importance
- equate: regard as equal or equivalent

To these we should add several further terms:

- alignment: bring into line, place in a straight line
- comparable: that can be compared; fit to be compared
- examinations
- assessments
- qualifications

Confusingly, comparability research over the years has used the latter three terms almost interchangeably. Partly this is due to the historical background. Originally the term 'examinations' was applied both to the

written papers and the overall award. However, that was when 'examinations' (in the sense of the overall award) comprised entirely written papers. Assessment became a term of use to describe components of awards which were not written – coursework, speaking tests etc – and has now tended to become the preferred term to refer to the overall award. Very strictly defined, 'qualification' means the piece of paper which conveys the award, in the same way that 'certificate' does. However, it is also used as the term for the overall award.

The historical papers discussed so far in this article have tended to refer to 'examinations' as the overarching term for assessments which are part of a system of education and training, leading to further educational or employment opportunities. In the remainder of the article (except where reference is being made to the historical documents), 'qualifications' will be used as the preferred term, as it encompasses a wider variety of assessment practice.

It is important to note that the COD definition of 'standards' includes a qualifier – *for a particular purpose*. This is often lost in debates, media headlines and so on. It is also important to note that 'equivalence' and 'alignment' have different meanings. It is possible for qualifications to be aligned according to their equivalence on one particular aspect but to remain non-aligned on other aspects. For example, the subject of General Studies at A level could be compared with other A level subjects on the basis of the amount of teaching time. A comparison made on the basis of prior attainment of students would give a very different result.

The evolutionary problem in establishing equivalent standards between qualifications

In 1911 the report recognised clearly that the different purposes to which the results of examinations might be put had a bearing upon comparability. The Norwood report identified a key difficulty, which is that, as qualifications evolve, so the underlying assumptions change – which can affect conceptions of comparability. The situation as it developed from 1911 to 1943 is a perfect illustration of this. In 1911 the problem was that multiple qualifications were being used for very similar purposes and they required a degree of inter-changeability. The solution – a single system, with qualifications being used for multiple purposes – was criticised (in 1943) because the qualification in its more multiply-acceptable form attracted more students, who in turn required a greater variety of courses within the system to accommodate their needs. The comparability solutions provided by the original conception of the system were eroded in the face of these challenges.

Both the 1911 report and the 1943 report provide insights into why comparability is so important in the history of English examining. Three main reasons emerge.

First is the relationship between comparability and validity and reliability. The Norwood Report (p.31) is absolutely clear that "*some approximation to uniformity of standard in assessing attainment*" is desirable (if not essential) for examinations to hold any value or currency beyond the school gates. However, it is worth noting the use of 'approximation', and the suggestion of 'uniformity' rather than equivalence. A key aspect of validity is that the inferences made on the basis of the outcomes of any assessment should be meaningful, useful and appropriate, in the particular set of circumstances that they are used. Reliability, which relates to the stability, consistency and precision of an assessment, is strongly linked to validity, because poor reliability compromises validity. Comparability is a part of validity, as alternative pathways or routes within assessments which lead to the same outcome, or the use of the outcomes of different assessments to access the same FE or employment opportunities, imply a degree of equivalence between them which must be borne out.

Second is the need to provide students with a meaningful choice of qualifications which are recognised by employers and higher education institutions. In 1911 it was proposed that these qualifications should be "valid wherever in the country they may go". Nowadays we might expand this to "wherever in the world they may go". In essence, learners, education institutions and businesses need to be assured of the value of the qualifications.

Third is the social responsibility of awarding bodies to provide students with appropriate qualifications, delivered fairly. In 1943, the Norwood Report referred to an objective test and outcome, taken under equal conditions with no prejudice or favouritism. Today it is expressed in the fact that awarding bodies are committed to ensuring that all assessments are fair, have sound ethical underpinning, and operate according to the highest technical standards.

Having explored in some detail the extent to which educational thinkers early in the twentieth century defined and understood issues of comparability, it is worth tracing briefly some of the more recent developments in theory and practice. For a more detailed description of the evolution of comparability from the mid-nineteenth century to the present, see Tattersall (2007).

Crucial amongst these developments was the move nationally towards measuring, monitoring and maintaining standards between qualifications. This was led primarily by the awarding bodies and regulatory authorities. An unpublished summary (Anonymous, 1970) of early comparability studies recently found in the Archives at Cambridge Assessment reveals that, following discussions at the annual meeting of the Secretaries of GCE examining boards in 1951, it was decided to institute inter-board investigations in a whole series of subjects, at both Ordinary and Advanced level. Nineteen separate studies were described in this paper, investigating inter-Board A level standards from eleven boards including those in England, Wales and N. Ireland. The work encompassed 16 different subjects (and included what may have been the only comparability work ever to have addressed the subjects of Zoology or Botany). These studies were carried out between 1953 and 1968. The report also made reference to similar investigations having been held on subjects at Ordinary Level, but so far no documented evidence of these has come to light. The methods used by the majority of studies carried out in the 1950s and 1960s are familiar to researchers today, as they asked panels of examiner judges, to scrutinise script evidence from key grading points, alongside evidence of demand derived from syllabuses, regulations, 'hurdles' (possibly meaning grade boundaries), and mark schemes. A variety of different judgemental methods of using the script evidence were tried. These included simply reading and discussing the scripts in the light of the demand of papers; re-marking exercises; cross-moderation[2] approaches; and a 'conference' approach. The conference approach involved a review of the practices of the various boards in the subject concerned, and did not incorporate any scrutiny of scripts. Three of the four conferences described related to subject areas already addressed by other forms of comparability study (hence 19 studies in

2 Cross-moderation methods have been defined as 'systematic ways of looking at candidates' work, that ought to be of the same standard.' (Adams, 2007, p.212)

only 16 subjects). Although the conference approach omitted any investigation of script evidence, it was considered helpful: in the description of the Geography conference (the only subject where a conference approach was taken without any other type of comparability study being conducted), it was stated that:

> *This conference brought out yet again the very great value which the investigations and conferences have had over the years in bringing together persons concerned with carrying out similar work for different boards. The interchange of ideas has been valuable and there has undoubtedly been much cross-fertilisation, all of which has contributed towards establishing the comparability of the boards in their demands on candidates and the comparability of the awards made.*
> (A review of investigations into subjects at Advanced Level conducted by the GCE boards 1953–1968, p.14.)

Two startling facts about the dedication of the boards to comparability at this time emerge from the summary of comparability studies between 1953 and 1968. In the description of a study carried out in Physics in 1968, the cost of the study is mentioned as being £16,000, which according to two different UK inflation/price conversion tools[3] would equate to about £200,000 today. This was for just one study, albeit one which was designed to test a new method, which included a subject-based reference test taken by a sample of students (the size of the sample was, alas, unrecorded in this summary document) and a questionnaire survey of schools. The second surprising piece of commentary describes the scale of the Mathematics study in 1954:

> *There were 20 syllabuses, 50 question papers and nearly 500 scripts; photocopying was not used for the scripts, and the enquiry therefore took three years to complete.*
> (A review of investigations into subjects at Advanced Level conducted by the GCE boards 1953–1968, p.5.)

Advances in comparability theory and practice between the 1970s and the present day have been widely and extensively documented. Several reviews were completed of the studies carried out in the 1970s and 1980s (Bardell, Forrest, and Shoesmith, 1978; Forrest and Shoesmith, 1985; NEAB, 1996), which largely comprised judgemental cross-moderation approaches. These studies focussed mainly on comparing qualifications on the basis of the perceived demands of the specification and assessment material and/or the perceived quality of examinees' work. As Bramley (2011) has pointed out, both 'perceived demand' and 'perceived quality' might be thought of as higher-order attributes that are built up from lower-order ones and the definition of these attributes suggests that it is appropriate that they be investigated by methods that use the judgement of experts. The development of these methods continued into the 1990s and the use of paired comparisons and Rasch statistical analysis, based upon the work of Louis Thurstone (1959), was added to the research armoury during this period (see Bramley, 2007, for a full history and description of the method). A further refinement to this type of study was the development of a rank-ordering method (Bramley, 2005: Black and Bramley, 2008).

Alongside the development of methods for use with the judgement of experts, alternative statistical methods for assessing the equivalence of qualifications were explored. These statistical comparisons are based upon different attributes to those used for judgemental comparisons. Attributes for statistical comparisons do not include perceptions of quality or of demand; rather they are based upon some statistical measure applied to a particular population, such as 'percentage gaining grade A', or 'average grade conditional on a given level of prior attainment' (Bramley, 2011). A statistical strand was developed alongside the judgemental method applied to large scale inter-board studies in the 1990s (see Adams *et al*., 1990 for an early example; also Fowles, 1995; and Pinot de Moira, 2003). Syllabus/subject pairs work has been a feature of research since the early 1970s (Nuttall *et al*. 1974, chapter III) and methods for deriving a 'putative' grade distribution based on prior attainment have been developed more recently.

The final, important, research strand which should be included in this potted history of the development of comparability theory has been the discussions about what is meant by the terms used to define and discuss comparability. Although this has been alluded to throughout the history of comparability (Massey, 1994) it has increased greatly in more recent years, fuelled by debates between individual researchers (Newton, 2005, 2010; Coe, 2007, 2010) and by public events such as the debate: *School exams: what's really happened to 'standards'?*, hosted by Cambridge Assessment on 29th April 2010. The essence of these arguments relates to whether researchers use common terms when discussing comparability, exactly what each term means and how a more common understanding might be brought about. One of the most important recent developments in thinking about comparability is Newton's insight that:

> *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, 2010, p.288)

Discussion

It is important to be open and honest about the challenges that are inherent in the study of comparability and assessment processes. Comparability has been an issue for the past century and there are still few completely satisfactory solutions. In this respect an important lesson can be learnt from the 1943 review of the 1911 system changes: if the qualifications are changed, there will be an impact on uptake and use of those qualifications, thus raising further comparability issues. In other words, comparability has always and will always evolve as qualifications do.

In order to go forward, a number of issues need to be addressed:

First, it is important to find clear ways of dealing with the different definitions of comparability, especially when applied to the different purposes to which the results of qualifications are put.

Secondly, Newton (2010) has made it clear that it cannot be assumed that different commentators are talking about the same thing, even when similar terminology is used. There are a number of challenges inherent in the process of explaining comparability evidence to the users of qualifications (students, parents and schools and prospective employers). These include: (i) the confusing nature of the terminology; (ii) the claims which are made both by those organisations delivering qualifications and by wider authoritative bodies (e.g. national and

3 The currency conversion websites were: http://safalra.com/other/historical-uk-inflation-price-conversion/ and http://www.nationalarchives.gov.uk/currency/results.asp#mid

international government departments and organisations); and (iii) the fact that much of the comparability information that reaches the wider public is conveyed by a third party, such as the media.

Thirdly, it must always be remembered that most of the methods of determining equivalence between qualifications can only ever be accurate to a certain point. A statistical or judgemental method can provide a very specific measure of equivalence, but care must be taken to ensure that it is not spurious, given the statistical limitations of the grades awarded. As Murphy (2010) has stated:

> *In measurement terms they [GCSE and A level examinations and the grades which they produce] are 'approximate estimates of educational achievement', which need a great deal of interpretation, rather than precise measurements on a highly sophisticated scale*. (Murphy, 2010, p.2).

Finally, as qualifications become more high stakes, it needs to be decided whether comparability is the master, or the slave, or neither. The Quality Assurance Agency for Higher Education (2006), stated that:

> *…it cannot be assumed that students graduating with the same classified degree from different institutions having studied different subjects, will have achieved similar academic standards; (b) it cannot be assumed that students graduating with the same classified degree from a particular institution having studied different subjects, will have achieved similar academic standards; and (c) it cannot be assumed that students graduating with the same classified degree from different institutions having studied the same subject, will have achieved similar academic standards… These implications are implicitly acknowledged and accepted in the higher education (HE) sector. They are of long standing, and many of those who make use of degree classifications couple this information with their judgement and experience when employing graduates, or recommending awards for further study, or determining salaries*. (QAA, 2006, pp.1-2)

It is important to ensure that the drive for comparability, and the arguments about comparability do not obscure other key aspects of the assessment process, such as fitness for purpose. It is clear from the historical perspective provided in this paper that comparability is an enduring issue, not easily resolved, and that systemic changes inevitably produce further comparability problems. Reviewing the history of these can help to anticipate what may happen in future if changes are made.

References

Adams, R. (2007). Cross-moderation methods. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Adams, R.M., Phillips, E.J. & Walker, N.A. (1990). GCSE intergroup comparability study 1989: Music. Organised by the Welsh Joint Education Committee and the Northern Ireland Schools Examination Council for the GCSE and the Joint Council of the GCSE. Available in: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Allen, R.E. (1992). *The concise Oxford dictionary of current English*. 8th Edition. Oxford: Clarendon Press.

Anonymous (1970). *A review of investigations into subjects at Advanced Level conducted by the GCE boards 1953-1968*. Available from Cambridge Assessment archives.

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

Black, B. & Bramley, T. (2008). Investigating a judgemental rank-ordering method for maintaining standards in UK examinations. *Research Papers in Education*, **23**, 3, 357-373.

Board of Education Consultative Committee (1911). Report of the Consultative Committee on examinations in secondary schools. Parliamentary Papers. Session 1911, Vol. XVI, 159. Chairman: Acland, A. H. Dyke.

Bramley, T. (2005). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement*, **6**, 2, 202-223.

Bramley, T. (2007). Paired comparison methods. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Bramley, T. (2011). Comparability of examinations standards: Perspectives from Cambridge Assessment Seminar. April 6th 2011, Cambridge.

Coe, R. (2007). Common examinee methods. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

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

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

Fowles, D. (1995). A comparability study in Advanced level Physics. A study based on the Summer 1994 and 1990 examinations. Organised by the NEAB on behalf of the Standing Research Advisory Committee of the GCE Boards. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Massey, A. J. (1994). Standards are slippery! *British Journal of Curriculum and Assessment*, **5**, 1, 37-38.

Murphy, R. (2004). *Grades of Uncertainty: Reviewing the uses and misuses of examination results*. London: Association of Teachers and Lecturers.

Murphy, R. (2010). *Grades of Uncertainty? Why examination grades will never give us anything more than very approximate answer to questions about changes in standards over time*. Paper presented at 'School exams: what's really happening to 'standards'?': An event held in London on 29th April 2010. Available at: http://www.cambridgeassessment.org.uk/ca/Viewpoints/Viewpoint?id=132622, accessed on July 26th 2010.

NEAB (1996). *Standards in public examinations, 1977 to 1995: A review of the literature*. Conducted for the School Curriculum and Assessment Authority and the Office for Standards in Education by the Northern Examinations and Assessment Authority on behalf of the GCSE and GCE boards. London: School Curriculum and Assessment Agency.

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

Newton, P.E. (2010). Contrasting conceptions of comparability. *Research Papers in Education*, **25**, 30, 285-292.

Norwood Report (1943). *Curriculum and Examinations in Secondary Schools*. Report of the Committee of the Secondary School Examinations Council appointed by the President of the Board of Education in 1941. London: HMSO.

Nuttall, D. L, Backhouse, J. K. & Willmott, A. S. (1974). *Comparability of standards between subjects*. Schools Council Examinations Bulletin 29. London: Evans/Methuen.

Pinot de Moira, A. (2003). An inter-awarding body comparability study. The statistical analysis of results by awarding body for: GCE A level and

AVCE Business; GCE A level Chemistry; GCSE A level Geography; AVCE Health and Social Care. A study based on the Summer 2002 examination and organised by AQA on behalf of the Joint Council for General Qualifications. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Quality Assurance Agency for Higher Education (2006). Background Briefing Note: The classification of degree awards. Gloucester: QAA. Available at: http://www.qaa.ac.uk/education/briefings/classification_20sept06.asp accessed on 25.3.2011.

Tattersall, K. (2007). A brief history of policies, practices and issues relating to comparability. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.), (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Thurstone, L.L. (1959). *The measurement of values*. Chicago, Illinois: University of Chicago Press.

A guide to comparability terminology and methods

Gill Elliott Head of Comparability Programme, Assessment Research & Development

Preface

Comparability has a broader meaning than is often attributed to it. Comparability of examination standards concerns anything related to the comparison of one qualification (or family of qualifications) with another and encompasses many different definitions, methodologies, methods and contexts. Comparability of educational standards is broader still, including comparisons of educational systems and outcomes, again in a number of contexts.

One of the issues which has beset researchers in recent years has been the proliferation of terminology to describe different aspects of comparability research. This makes it especially difficult to explain the issues to non-specialist audiences, including students taking examinations. As the results of an increasing variety of qualifications are put to diverse purposes in a high-stakes environment, the issue of communicating meaningfully about comparability and standards in qualifications becomes ever more important.

This article has been written to provide non-technical readers with an introduction to the terminology and issues which are discussed elsewhere in this edition of *Research Matters*.

The article is divided into three sections. In Section 1, the common terms used in comparability research will be identified and their usage discussed. Section 2 presents a framework for addressing the literature. Finally, Section 3 describes possible methods for investigating comparability, and illustrates how these must be related to the definition of comparability at hand.

Introduction

One of the problems of writing an article such as this is where to start. There is no beginning and no end to the issues which can be identified; rather there is a web of interlinking concepts, few of which can be adequately described without invoking others, and which themselves then need explanation. The issues interweave with one another to such an extent that separating them out for the purposes of explanation runs, to some extent, the risk of losing some of the sense of the whole. With this in mind this introductory section explores some of the key points relating to the holism of the topic which need to be borne in mind when reading the article as a whole.

Comparability is part of validity. In particular, comparability in assessment relates to the validity of inferences about the comparability of students, teachers, schools or the education system as a whole that are made on the basis of assessment outcomes.

Comparisons are manifold. They can apply to the demand of the system or assessment; the curriculum content and domain coverage; the performance of students and the predictive ability of the outcomes. Comparisons can be applied in different ways – between syllabuses including within and between awarding bodies, between subjects and over time. Comparability studies (i.e. actual comparisons) tend to address these issues individually, so a study investigating the demand of two or more qualifications over time will usually have little to contribute about the performance of students between subjects. However, these distinctions are much less apparent in the literature about the philosophies, processes and theories of comparability, which can cause confusion if the reader has a different conceptualisation of comparability from the author. This is why the next point is so important.

Providing adequate definitions of comparability and standards is crucial. The word 'standards' and the phrase 'definition of comparability' do not appear in the title of this article, but they are at the heart of the issues discussed. Comparability terminology, whether used in a general or a specific context, can mean many different things. Unless a commentator clearly specifies exactly what they mean by these concepts, a reader is in danger of drawing misleading conclusions. This has been recognised in point 1 of the summary of recommendations of the report into the Standards Debate hosted by Cambridge Assessment in 2010:

> *Before any discussion about 'standards', terms need to be defined and clarity reached about what kind of standards are being referred to*. (Cambridge Assessment, 2010).

Some terms are deeply inter-related... It is simply not possible to understand how definitions of comparability apply without understanding the related terminology: such as type of comparability, purpose of comparability, context of comparability, and attribute.

...but definitions and methods should always be kept separate. The distinction between definitions and methods is key to understanding some of the issues. A method is a technique for making a comparison, whilst a definition is the rationale and purpose behind the comparison, and it is not the case that they exist in a one-to-one relationship with one another (Newton, 2010). Any definition may be combined with any method – although a proportion of the resulting combinations will be invalid because the method in question will not address the definition. In the past, research concentrated mainly upon methods. Definitions, when provided, were seen as integral to the method. This is now considered undesirable.

Purposes. Purposes feature frequently in this article, and it is vital to understand that there are different sorts of purposes in comparability. There is the purpose for conducting comparability research in the first place. There is the purpose for selecting the particular entities which are to be compared (i.e. why do these examination systems or these particular qualifications need to be compared with one another?). Finally, there is the purpose of selecting a particular method (i.e. why is this method more suitable than that one?). These should also be distinguished from the purposes to which the outcomes of examinations are put, which are all about what the users of qualifications (students,

FE institutions, employers) are, rightly or wrongly, inferring or expecting from the qualifications.

The distinction between comparability and face comparability. Inasmuch as face validity is about the extent to which something appears valid, the term 'face comparability' can be used to describe the extent to which parallel assessments are expected or are seen to be of the same standard. Thus, if the qualification titles of assessments (e.g. 'A level' (AL), or 'General Certificate of Education') are the same, then users of those assessments will expect them to be comparable, regardless of the subject title or the date of the assessment. Additionally, even when the qualification title is not the same, there may be an expectation of comparability. Sometimes this is because there is an overlap in title, which establishes a link between the qualifications, for example, GCSE and IGCSE. At other times it is merely circumstantial juxtaposition which dictates a measure of face comparability – for example, a candidate presenting three A level grades might be expected to be of a similar general educational standard as a candidate who has taken the International Baccalaureate on the basis that they are taken at the same age, and provide access to similar pathways. In some cases examinations may not necessarily be designed to be equivalent. Nonetheless, if they are structurally the same, and use the same reported grades, they will almost certainly be perceived as equivalent in the public eye.

Having face comparability does not mean that qualifications have had their equivalence put to the test, nor, necessarily, that any claims about their equivalence have been made by the providers of the qualifications.

Section 1: A glossary of common comparability terms and their usage

Figure 1 provides a list of terms used to describe comparability issues. Accompanying each term is a discussion of the way in which it is used within a comparability context. It is not always possible to provide definitive meanings for terms, because different authors use them in different ways.

The list begins with the most commonly used terms – those which are often found in media reports and public documents, and progress to terms used more frequently in a research, rather than public, arena. Terms which are related to one another are grouped together.

Figure 1: A glossary of common comparability terms and their usage

Term	*Usage, examples of use, popular misconceptions and/or problems of interpretation*
Comparability/ Defining comparability/ Definition of comparability	In its most general usage this is an umbrella term covering a large number of different definitions, methodologies, methods and contexts, e.g. "The seminar will be about comparability". However, in comparability research there also exist general definitions of comparability (which are less general than that described above) and specific definitions of comparability. These are discussed in more detail later in this article, but essentially are a more technical usage of the term comparability. General definitions of comparability are those where the author provides an overarching definition of what they understand by comparability. Such use of the term comparability DOES NOT specify the particular context or purpose of the comparison. An example of this is the following: *The extent to which the same awards reached through different routes, or at different times, represent the same or equivalent levels of attainment*. (Ofqual, 2011a). Specific definitions of comparability are those where the author DOES specify the particular context or purpose of the comparison. An example of this is the following: *Comparable grading standards exist if students who score at equivalent grade boundary marks demonstrate an equal amount of the discernible character of their attainments*. (Newton, 2008) One of the problems which has beset both technical and non-technical users of comparability research over the years has been a misunderstanding about what is meant by comparability by particular authors. If a general definition of comparability is provided, it can mislead readers into the assumption that the arguments made or the methods described can be applied to any context or purpose. This is not necessarily the case.
Comparable	This is a classic example of a term with several usages. Strictly speaking if it is stated that two qualifications are comparable, it means that there are grounds upon which a comparison can be drawn. Apples and pears are comparable, in the sense that they share common features and use. Concrete and block paving are comparable, because one might wish to make a choice between them. Apples and concrete are not comparable, because one would never expect to use them for the same purpose. However, the more common usage of the term is to describe two or more qualifications which have been compared and found to be equivalent, e.g. qualification X and qualification Y are comparable. Even more common is the use of the term to describe two or more qualifications which are assumed (but not proved) to be equivalent. This situation tends to reflect face comparability issues, e.g. it is possible to state that, "The UK A level system and the German Abitur system are comparable," and mean that there are some broad similarities between the systems – similar age group of users, similar purposes to which the results are put. This statement does not necessarily mean that there is any evidence that the systems are equivalent.
Non-comparable or not comparable	Strictly speaking, if it is stated that two qualifications are not comparable, it means that there are no grounds upon which a comparison can be drawn, not that they have been compared and found not to be equivalent. However, it is often used to mean the latter.
Types of comparability (also sometimes called modes of comparability	This refers to the nature of the comparison: • between awarding bodies • between alternative syllabuses in the same subject • between alternative components within the same syllabus • between subjects • over time – year-on-year • over time – long term

Standards

"A definite level of excellence, attainment, wealth, or the like, or a definite degree of any quality, viewed as a prescribed object of endeavour or as the measure of what is adequate for some purpose" (OED, 2011).
It is important to note that the definition of 'standards' includes a qualifier – *for some purpose*. This is often lost in debates, media headlines and so on.

Test
Award
Assessment
Examination
Qualification

Comparability research refers to these terms almost interchangeably. In the same research paper (including the present one) 'examination', 'qualification' and 'assessment' may each be used to refer to the award as a whole. Partly this is due to the historic background to the topic. Originally the term 'examinations' was applied both to the written papers and the overall award. However, that was when 'examinations' (in the sense of the overall award) comprised entirely written papers. Assessment later became a term of use to describe components of awards which were assessed in other ways – coursework, speaking tests etc.

A dictionary definition of 'qualification' suggests that it is: "a quality or accomplishment which qualifies or fits a person for a certain position or function; (now esp.) the completion of a course or training programme which confers the status of a recognized practitioner of a profession or activity." (OED, 2011). An alternative meaning attributed to the term is the piece of paper which conveys the award, e.g. "a document attesting that a person is qualified." However, 'certificate' is more commonly used in this context. In common educational usage the term 'qualification' is more frequently defined thus:

An award made by an awarding organisation to demonstrate a learner's achievement or competence. (Ofqual, 2011a).

Alternatively, some users prefer to use 'qualification' to mean a particular class, or family, of award – e.g. A levels or GNVQs or IGCSEs.
In this article 'qualification' is used as the preferred term for referring to the award as a whole.

'Test' has always had a slightly different connotation, relating more to psychometric contexts, such as reading tests or IQ tests.

Syllabus/specification

The document describing what will be assessed and how it will be assessed. Some awarding bodies use the more recent term 'specification' whilst others retain the traditional term 'syllabus'. In this article the term 'syllabus' is used.

Methodology

Science of the method (or group of methods) available for use.

Method

Specific procedure which is followed in order to achieve a comparison.

There is an important distinction to be drawn between methodologies and methods. A methodology provides the reasoning which underlies a method or group of methods. The method itself is the specific procedure carried out on a particular occasion.

Demand

The level of knowledge, skills and competence required of the typical learner.
Defined alternatively by Pollitt *et al*. (1998) as the "requests that examiners make of candidates to perform certain tasks within a question".

Difficulty

How successful a group of students are on a particular exam question or task.
Defined and analysed post-test (Pollitt *et al*., 2007). Difficulty can be represented numerically e.g. as 'facility values' – the mean mark on an item expressed as a proportion of the maximum mark available.

Equate

'Equate' and 'equating', used in the context of assessment, tend to have a very specific meaning.

Equating is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. Equating adjusts for differences in difficulty among forms that are built to be similar in difficulty and content. (Kolen and Brennan, 2004, p.2)

The above definition comes from the US context, but the concept does apply to year-on-year comparability of examinations in the same subject where there have been no changes to the syllabus or assessment structure.

Attainment

The underlying skills, knowledge and understanding (SKU) which can be inferred (approximately) from observed performance.

Purpose or context of comparability

The condition under which the comparison is taking place – which helps to fix its meaning, for example:
- a comparison between the standards of demand (a comparison of the requirements made of the candidates);
- a comparison of standards of attainment/grade standard (the level of performance required at key boundaries).

Attribute

The grounds for the comparison which is being made; for example:
- demand of examinations;
- results of examinations;
- content of syllabuses/domain coverage;
- fitness for a particular purpose of examination outcomes.

Bramley (2011) states, "comparisons among any entities are always on the basis of a particular attribute. For example, an apple and an orange could be compared on the basis of weight, or sweetness, or price". Elliott (2011) demonstrates how, by conducting a comparison on the basis of different attributes amongst fruit, the result of the comparison changes. When strawberries are compared with apples on the basis of weight two thirds of an average apple corresponds to nine average strawberries; when the comparison is made on the basis of vitamin C content nine average strawberries correspond to six average apples. So, nine average strawberries are equivalent both to two-thirds of an apple and to six apples, and this is not contradictory. Applying the same argument to comparability of assessments means that if a study provided evidence that two qualifications were equivalent in terms of content domain coverage, it does not follow that they would also be equivalent in terms of the proportion of students being awarded a particular grade. That attribute must be compared separately and may give an entirely different answer.

Equivalence

The dictionary definition is "equal in value, power, efficacy or import" (OED, 2011). However, in usage the term tends to mean 'a degree of...', or 'extent of...', implying that in practice, equivalence is not absolute.

The meaning of equivalence as 'equal in amount' can be measured in a different way to its meaning as 'equal in value or importance'. Using the definition of equivalence as equal in importance or value, it can be argued that, if two qualifications are regarded as equivalent, the fact that they are used as such is evidence that they are. Whilst this argument may seem circular, it is based upon the fact that 'equivalence' as defined, is about currency and value, which is to an extent a subjective measure. Something can only be considered valuable if somebody has attributed a value to it. And as long as that value continues to be attributed, the object retains its currency.

Alignment

Arrangement in a straight or other determined line. The action of bringing into line; straightening. (OED, 2011)
The definition of alignment implies some action which has been brought about to create equivalence on a particular attribute. However, it must be stressed that alignment on one attribute will not result in alignment on another. Alignment can take place pre-or post- awarding. Alignment of curriculum content of a qualification with another qualification is likely to take place at a very early stage of qualification development. Alignment of grade boundaries (with, say, the previous year) takes place during awarding.

Section 2: Understanding the arguments in the literature

The literature which has built up around the issues of comparability is both complicated and, at times, confusing. This is partly because authors have used different ways to conceptualise the topic, partly because they sometimes use different terms to describe the same thing and sometimes use the same term to describe different things, and partly because there seems to be little underlying agreement about which (or whose) concepts should be used as the basis of comparability practice. This literature is particularly difficult for a non-technical audience, because it is hard to know where to start. A frequent mistake made by non-technical readers is to pick up on just one author's views, and assume that those views are definitive. In fact there is very little literature in comparability research which can be described as definitive, and this presents a problem when attempting to decide upon appropriate practice for monitoring and maintaining standards.

Figure 2 provides a framework for understanding the arguments in the literature. In this framework each box shows a broad area which has been covered by the literature. It is not the case that every piece of literature fits only into one box – a single journal article may touch upon many of the areas. However, the intention of the framework is to try to make clearer what the overarching topics of interest may be. Each box is described in more detail below.

History of comparability methodologies, methods and definitions

These analyses of the methodologies, methods and definitions used throughout the long history of comparability, provide an insight into the question of 'what happened next?' By analysing the reasons why certain approaches to comparability were taken and then how well they succeeded predictions can be made about the outcome of future changes. These retrospectives (e.g. Tattersall, 2007; Newton, 2011) are very valuable (Elliott, 2011).

Categorical schemes for ordering definitions of comparability

A number of authors have provided frameworks for ordering the many different definitions of comparability. Definitions can be grouped into categories or 'families', where certain definitions share particular properties. Such a framework tends to be expressed in terms of 'definitions. A, B and C share particular characteristics and can therefore be termed 'category X' whilst definitions D and E share different characteristics and can be placed into 'category Y'. Inevitably each author presents a different angle about how the categories should be organised, some of which differ only slightly; others radically. Newton (2010) provides a discussion of this, and a description of more than thirty-five definitions and eight separate categorisation schemes.

Definitions of comparability

As mentioned in the introductory section of this article, there are a number of different circumstances under which it is necessary to define comparability:

- In a theoretical paper in order to establish what, exactly, is being discussed.
- In an empirical study, where it is essential to establish the precise nature of the comparison being made.
- In more general public documentation: media reports, awarding body websites, etc.

This has led to both general definitions of comparability and specific definitions of comparability.

General definitions of comparability take the form of a broad description of what comparability constitutes, for example:

> *… the application of the same standard across different examinations*. (Newton, 2007)
>
> *The notion of equivalence between qualifications of the same type offered in different institutions or countries. Comparability does not require complete conformity.* (AEC, 2004)

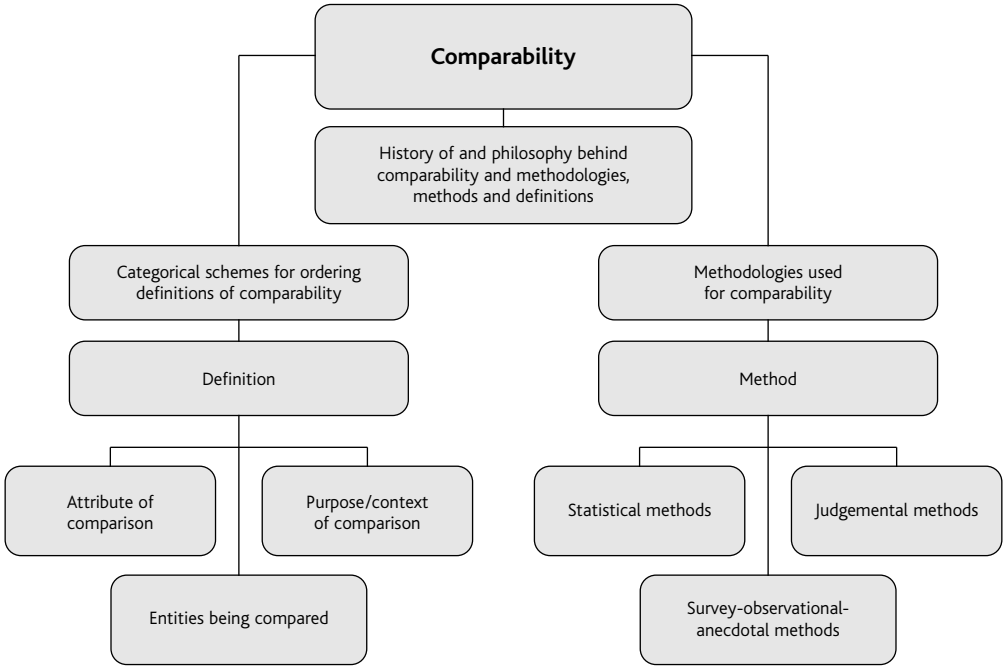


Figure 2: A framework for understanding the arguments in the literature

> *Comparability is the formal acceptance between two or more parties that two or more qualifications are equivalent. Comparability is similar to credit transfer.* (Harvey, 2004–11)

However, such general use of the term comparability does not specify the particular context or purpose of the comparison. Certainly in comparability studies (i.e. comparisons of qualifications) and ideally in detailed articles in the literature there needs to be some considerably more specific definition of the terms being used. Examples of specific definitions of comparability include:

> *Comparable grading standards exist if students who score at equivalent grade boundary marks demonstrate an equal amount of the discernible character of their attainments.* (Newton, 2008)

Specific definitions often comprise a combination of the attribute being compared and the purpose/context of the comparison.

Attribute of comparison

The attribute of the comparison is a key part of the definition. The attribute is the characteristic which forms the basis of the comparison. Using the example given above, the emboldened text describes the attribute.

> *Comparable grading standards exist if students who score at equivalent grade boundary marks demonstrate an equal amount of* ***the discernible character of their attainments***.

Purpose/context of the comparison

The purpose and/or the context of the comparison is also important to the definition. Purpose and context are not entirely the same thing. Purpose is the reason for carrying out the comparison. The context of the comparison refers to 'the standard of ***what***?' Again using Newton's definition as an example, it can be seen that a context is given:

> *Comparable* ***grading*** *standards exist if students who score at equivalent grade boundary marks demonstrate an equal amount of the discernible character of their attainments*.

By including the context of 'grading standards', Newton makes it clear that the comparison in this case is to establish that candidates who are matched in terms of attainment, achieve similar grades in the assessments being compared. There is no implication that they will necessarily perform in similar ways in future, nor that they have covered the same content.

The purpose of the comparison becomes important if one is trying to decide whether a comparability study is worth conducting. An example of this can be found in the adage "things ain't wot they used to be." It is often alleged that examination standards (in some overarching, general sense) have declined over time. Yet were a study to be mounted to 'prove' this one way or another, what would be the purpose of the research? Would it be to discredit the systems which had enabled this to happen? Surely, in this case, the purpose *of the comparison* is not particularly valid. If 'standards' are not currently fit for purpose, then that is an issue of validity which needs to be dealt with, by making them so. The comparison with some point in the past when they were allegedly fit for purpose is arguably largely irrelevant.

Entities being compared

This refers to whether the comparison is being made (for example) between alternative syllabuses within the same subject (either between or within awarding bodies), between alternative components within the same syllabus, between subjects, over time or between different modes of assessment (e.g. pen-and-paper scripts versus online testing).

Methodologies used for comparability

Just as the categorical schemes for ordering definitions group together those definitions which share common features, methodologies provide the reasoning which underlies a method or group of methods.

Methods

Methods are the techniques used to make a comparison. Traditionally, the method section of a scientific paper should be sufficiently detailed to enable the procedure to be replicated. In comparability research there have traditionally been two broad groups of method: statistical and judgemental (Newton *et al*., 2007). Figure 2 also includes a new category of method, which we have termed 'survey-observational-anecdotal'.

Statistical methods

Statistical methods are based upon the principle that the 'standard' can be detected and compared via the data emerging from the assessments; the number and proportion of students achieving given grades, controlled with data pertaining to concurrent, or previous performance, and/or other data such as demographic features.

Judgemental methods

Judgemental methods rely upon human judgement to detect and compare the 'standard' by asking experienced and reliable commentators (often practising examiners) to examine assessment materials and/or candidates' scripts.

Bramley (2011) states that:

> *... when investigating comparability of assessments, or of qualifications, we have focussed 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 judgment of experts.*
>
> *Other bases for comparisons are possible, such as 'percentage gaining grade A', or 'average grade conditional on a given level of prior attainment'. If comparability is defined in terms of this kind of attribute, then statistical methods are necessary for investigating it.*

Survey-observational-anecdotal methods

A third group of methods also exists in comparability research. Here termed 'survey-observational-anecdotal', this is information obtained from 'users' of qualifications, usually by surveys and face-to-face interviews. For example, QCA and Ofqual investigated perceptions of A levels and GCSEs by asking students, teachers and parents about their perceptions of these qualifications in a series of surveys (e.g. QCA, 2003; Ofqual, 2011b). Other examples are a study investigating differences between pathways (Vidal Rodeiro and Nádas, 2011), and changes in particular subjects over time (Elliott, 2008). Whilst these studies were not necessarily targeted at comparability issues directly, they are nonetheless relevant.

Data about patterns of centres (schools) changing which assessments they enter their students for can be illuminating, especially when combined with information about the reasons for such changes, even if this latter information is only anecdotal. For example, if a large group of centres switched from assessment A to assessment B, claiming that assessment B was more challenging, it provides some evidence about the comparability of the two assessments. The fact that the anecdotal evidence (centres' claims about the relative standard of the qualifications) is matched by their behaviour (changing to the alternate syllabus) gives the evidence some credence.

Other anecdotal information can be found amongst the semi-organised vocalisations of the assessment-users' communities, principally on subject or assessment forums on the internet, but also in the less formal publications associated with particular subjects or user groups, and at conferences and INSET events. The benefit of such information is that it can represent the considered reflections of a group of experienced users of qualifications within the subject area, who are reasonably representative of the overall population of users. Sadly, the limitation is that it is not always possible to determine the provenance of the authors. Nevertheless, such information – especially when it can be obtained from a source about whom enough is known to render it reputable – should not be discounted. This third category of methods tends to investigate face comparability. By engaging with users, the issues which emerge may be solely limited to the perceptions held or they may reflect more fundamental, underlying comparability issues.

Section 3: A guide to methods

In this section, a guide to methods is presented. A list of *methods* has been chosen (rather than a list of possible definitions or a chronological study of the literature) for several reasons:

- Methods are arguably less elusive than other elements of comparability.
- A major study of comparability, published as a book by QCA (Newton *et al*., 2007), is arranged by methods. By following the same approach, readers will easily be able to refer back to this seminal work for more detail.

The guide to methods which follows provides the following information:

Method title

Methodology	A description of the methodology (the reasoning which underlies the method). If the method is part of a recognised 'group', such as 'statistical' or 'judgemental' this is also identified here.
Method	The specific procedure which is followed in order to achieve a comparison. In scientific papers the method section is intended to contain sufficient detail to enable other researchers to replicate the study. In this instance, the method is described rather more broadly and is intended to provide readers who are unfamiliar with the method with sufficient outline knowledge to enable them to access the relevant literature.
Example of context	This provides a single example of a context in which the method has or might be used. There may be other contexts than the example given, and some contexts may be more appropriate than others. These are not addressed. The example given is intended to serve the purpose of exemplifying a possible comparison for the benefit of readers who are unfamiliar with it.
Example of a definition could be used with this method	The definition given is **an example only**. There may be other definitions than the example given, and some definitions may be more appropriate than others. The discussion below outlines why this is the case. In some cases more than one example of definition is given in order to make it very clear that there is not a one-to-one relationship between methods and definitions.
References	In this section references for further reading are provided, plus (where available) references to studies which have used the method.

1. Statistical linking, using prior attainment as reference measure

Methodology	Statistical, based upon the reasoning that there will be a relationship between a group of students' mean score on a measure of prior attainment and their score on the qualifications being compared. The measure of prior attainment is the link between the scores of the students on the two (or more) qualifications being compared.
Method	The following results (scores) of students are combined: Cohort 1 students' scores from qualification A Cohort 2 students' scores from qualification B Cohort 1 and 2 students' scores from prior attainment measure. Analysis generally takes the form of scatter plots and regression analyses in order to interpret the relationship between qualifications A and B, but sometimes more advanced statistical techniques are applied.
Example of context	Comparing the GCSE awards from two or more different awarding bodies, based upon prior attainment at Key Stage 2 national tests (taken when the students were 11 years old).
Example of a definition which could be used with this method	Comparable grading standards exist if it can be demonstrated that students with an equal level of prior attainment achieve equivalent results.
References	Elliott *et al*. (2002); Al-Bayatti (2005); Baird and Eason (2004); Bell (undated).

2. Statistical linking, using concurrent attainment as reference measure

Methodology	Statistical, based upon the reasoning that there will be a relationship between a group of students' mean score on a measure of concurrent attainment and their score on the qualifications being compared. The measure of concurrent attainment is the link between the scores of the students on the two (or more) qualifications being compared.
Method	The following results (scores) of students are combined: Cohort 1 students' scores from qualification A Cohort 2 students' scores from qualification B Cohort 1 and 2 students' scores from concurrent attainment measure. Analysis generally takes the form of scatter plots and regression analyses in order to interpret the relationship between qualifications A and B, but sometimes more advanced statistical techniques are applied.
Examples of contexts	Comparing the GCSE awards in a particular subject from two or more different awarding bodies, based upon students' mean GCSE scores across all the subjects they have taken.
Example of a definition which could be used with this method	Comparable grading standards exist if it can be demonstrated that students who score equivalent grade boundary marks demonstrate an equal amount of concurrent attainment.
References	Bell (2000) provides a description of the advantages and limitations of this approach.

3. Statistical linking, using future attainment as reference measure

Methodology	Statistical, based upon the reasoning that there will be a relationship between a group of students' mean score on a measure of future attainment and their score on the qualifications being compared. The measure of future attainment is the link between the scores of the students on the cohorts being compared. (Comparisons between qualifications have not been carried out using this method to date – only comparisons between different subgroups of students.)
Method	A measure of future attainment is identified. Data are collected, by tracing students as they progress through the education system.
Examples of contexts	Investigating whether university students with equivalent grades in A level and Pre-U perform equally well in 1st year undergraduate examinations.
Example of a definition which could be used with this method	Comparable grading standards exist if it can be demonstrated that students with equivalent results demonstrate an equal amount of future attainment. (NB. Essentially this is the same as statistical linking using prior attainment as a reference measure; the difference being in the direction of the prediction.)
References	It is difficult to collect the data for this kind of study – we are not aware of any published examples.

4. Statistical linking, using purpose-designed reference test battery

Methodology	Statistical, based upon the reasoning that there will be a relationship between the scores of a group of students on a purpose-designed reference test[1] and their scores on the qualifications being compared. The reference test provides the link between the scores of the students on the two (or more) qualifications being compared.
Method	The following results (scores) of students are combined: Cohort 1 students' scores from qualification A Cohort 2 students' scores from qualification B Cohort 1 and 2 students' scores from the reference test. Analysis generally takes the form of scatter plots and regression analyses in order to interpret the relationship between qualifications A and B, but sometimes more advanced statistical techniques are applied.
Examples of contexts	Comparing the A level awards across a number of different subjects. Comparing the GCSE awards over time.
Example of a definition which could be used with this method	Comparable grading standards (or standards over time) exist if it can be demonstrated that students with equal scores on the reference test achieve equivalent results.
References	Murphy (2007). The Centre for Evaluation and Monitoring (CEM) (Hendry, 2009) provides an independent, objective monitoring system for schools. The CEM work includes the use of ALIS (Advanced Level Information System) which uses both GCSE data and its own baseline tests as a measure of ability and a performance indicator for post-16 students. The ALIS test incorporates vocabulary, mathematics, and an optional non-verbal section.

5. Subject/syllabus pairs

Methodology	Statistical, based upon the reasoning that any (reasonably large) group of candidates who all take the same two examinations will have a similar distribution of grades in each. The assumption of a broadly equivalent performance by the same cohort of students across different qualifications provides the link between the scores of the students on the two (or more) qualifications being compared. Additionally, if the syllabus under scrutiny is compared in this way with not just one, but a series of others, trends in the relationships will emerge which will be even more informative than the individual pairs' scores alone.

1 Assuming a valid relationship between the SKU tested in the reference test and those tested in the qualifications being compared.

Method	A single group of students is identified who took both (all) qualifications being compared. Then (for example) the mean grades of these students on both the main and comparator syllabus are calculated. The difference between the two mean grades is then reported alongside the mean differences generated by repeating the process with a series of different comparators. The results are presented as tables or as graphs.
Examples of contexts	Comparing the A level awards across a number of different subjects.
Example of a definition which could be used with this method	Comparable grading standards exist if it can be demonstrated that the distribution of students' results was similar in each qualification.
References	Jones (2003); Coe (2007).

6. Statistical equating with a common component

Methodology	Statistical, based upon the reasoning that if there is a component which is common to both/all qualifications being compared, it can be used to link the scores of two or more qualifications.
Method	The common component of the two qualifications is identified. This is often a multiple choice, or coursework component. Candidates' scores on the common component are then used as the measure by which to compare the qualifications.
Examples of contexts	Alternative option choices within the same syllabus. Tiered papers with overlapping grades.
Example of a definition which could be used with this method	Comparable grading standards exist if it can be demonstrated that students who obtain equal scores on the common component achieve equivalent results.
References	Newbould and Massey (1979).

7. Looking at trends in pass rates for common centres (sometimes called 'benchmark centres')

Methodology	Statistical, based on the theory that if a centre has well-established teaching and its cohort remains stable (i.e. no changes in intake policy, or any changes in the nature of the student population for any other reason) the proportion of grades awarded in a syllabus should remain broadly similar over time.
Method	Suitable centres are identified for the syllabus concerned, according to strict criteria which are specified according to the comparison being made. These criteria generally include no known changes to the cohort in relation to previous years, no major changes to teaching practice (including staffing) and this to have been the case for a number of years.
Examples of contexts	Maintaining standards in the same syllabus over time.
Example of a definition which could be used with this method	Comparable grading standards exist if it can be demonstrated that year-on-year, common centres are awarded similar proportions of grades.
References	References to the use of common centres for establishing comparability between qualifications are limited to occasional committee papers, which are not widely available.

8. Statistical records of trends over time (uptake, grades, etc)

Methodology	Observational, based upon trends in publically available statistics.
Method	Data are generally displayed as charts and explanations are sought for the patterns arising.
Examples of contexts	Comparing standards over time in a particular qualification or subject. Used frequently in newspaper reports, but less featured in academic research.
Example of a definition which could be used with this method	Comparable standards exist over time if it can be demonstrated that, after allowing for all differences in cohort, social context and teaching practices, proportions of students awarded different grades are similar.
References	BBC (2010).

9. Other concurrent methods e.g. average marks scaling

Methodology	Statistical, designed specifically for the context of inter-subject comparability. The methodology is based upon the reasoning that 'average performance' can be used as a reference, enabling the relative difficulty of different subjects to be derived.
Method	Methods include Kelly's subject difficulty ratings, average marks scaling and Item Response Theory. The procedures are too complex to describe here – see references below.

Examples of contexts	In the Scottish and Australian education systems, the assumption that all subjects are equal is not always made. Difficulty ratings can be considered alongside graded results or marks in order to facilitate comparison between students with similar grades in different subjects.
Example of a definition which could be used with this method	Comparable standards between subjects at the same level exist when correction factors based upon the overall difficulty of each subject have been applied to all subjects.
References	See Coe (2007); Kelly (1976); Coe (2008).

10. Item banking/pre-testing systems

Methodology	Statistical, based upon pre-calibrated data. If the difficulty of particular items is known in advance, then these items can be used to link the standards of two or more qualifications.
Method	Items are pre-tested, either in an experimental context or as part of a live examination. The relative difficulty of the items is then established for the pre-test group of students. Assuming that this relative difficulty would remain the same for the populations of students taking the qualifications under comparison, then the scores of students on the pre-tested items can be used to equate the qualifications as a whole.
Examples of contexts	Keeping standards stable over time.
Example of a definition which could be used with this method	Comparable grading standards exist if the grade boundaries on two examinations correspond to the same points on the (latent) scale of the item bank. *Or* Two examinations with the same grade boundaries are comparable if the distributions of difficulty of the items from which they are each comprised are known to be equal.
References	Green and Jay (2005); QCDA (2010); Willmott (2005).

11. Simple holistic expert judgement studies

Methodology	Judgemental, based on the theory that a single suitably qualified expert is able to weigh up evidence from assessment materials and scripts to provide a considered opinion about whether the assessments are comparable.
Method	A suitable expert is identified, and required to study the syllabuses of the assessments in detail. They are then required to familiarise themselves with the assessment materials (question papers and mark schemes). Finally they are presented with script evidence and required to compare performances of students at equivalent grade points, allowing for differences in the demand of the question papers. They then prepare a report outlining their findings.
Examples of contexts	Comparing different awarding bodies' syllabuses in the same subject at the same level.
Example of a definition which could be used with this method	Comparable standards of attainment exist if it can be demonstrated that the script evidence of students who scored equivalent grade boundary marks was judged to be of similar standard.
References	Ofqual (2009a); Ofqual (2009b).

12. Holistic expert judgement studies: 'Cross-moderation'

Methodology	Judgemental, based on the theory that a balanced panel of suitably qualified expert judges will be able to detect differences in standards of performance at equivalent grade boundary points by systematic scrutiny of script evidence.
Method	The exact procedure varies slightly between different studies, but in essence comprises the identification of a panel of expert judges (usually balanced according to the assessments under comparison). Judges scrutinise scripts (usually from grade boundaries) according to a predetermined schedule and record their judgement about each script in a systematic way. The results have often been analysed using statistical techniques.
Examples of contexts	Comparing different awarding bodies' syllabuses in the same subject at the same level.
Definition	Comparable standards of attainment exist if it can be demonstrated that the script evidence of students who scored equivalent grade boundary marks was judged to be of similar standard.
References	Adams (2007).

13. Holistic expert judgement studies: Paired comparisons and rank ordering

Methodology	Judgemental, based on the theory that expert judges are able to provide the common element link for latent-trait equating.
Method	Expert judges are identified, and required to rank-order script evidence of candidates/pseudo candidates[2], from both/all syllabuses being compared whilst taking into account the demands of each question paper and the overall demand of the content material within the curriculum.

2 Often the 'whole' work of a single candidate on a given mark is unobtainable, so composite or pseudo candidates are generated, where the script evidence comprises the work of several candidates, chosen to aggregate to the desired total score.

Examples of contexts	Comparing standards of different awarding bodies' syllabuses in the same subject at the same level.
Example of a definition which could be used with this method	Comparable grading standards exist if the grade boundaries on two examinations correspond to the same points on the latent scale of 'perceived quality' constructed from the experts' judgements.
References	Bramley (2007); Bramley and Gill (2010); Bell *et al*. (1997); Greatorex *et al*. (2002).

14. Returns to Qualifications

Methodology	Observational/survey, based upon surveyed evidence of earnings in later life.
Method	A survey is conducted to establish information about respondents' earnings, qualifications, sex, age and years of schooling. The data are analysed in order to establish whether respondents with a particular qualification have higher earnings than those without it, once other factors have been accounted for (e.g. age, years of schooling etc.)
Examples of contexts	Investigating the potential for qualifications to have different impacts on future earnings.
Example of a definition which could be used with this method	Comparable economic values of two or more qualifications exist if the returns to qualifications[3] are similar.
References	Conlon and Patrignani (2010); Greatorex (2011).

3 Returns to qualifications can be defined as a statistical proxy for the productivity of people with a qualification, where productivity refers to the skills, competencies and personality attributes a person uses in a job to provide goods and services of economic value.

Summary

This article has aimed to make the terminology used in comparability research clearer, especially for a non-technical audience. It has also sought to provide a framework for following the arguments presented in the literature and to provide a guide to methods.

The arguments surrounding comparability of assessments in the UK are as heated now as they have ever been, but there is also need to sum up the debate (Cambridge Assessment, 2010), and to move on in a productive way.

Our hope is that researchers will gain a better shared understanding of definitions and methods, and begin to approach some of the many outstanding issues yet to be resolved – for example, whether particular definitions of comparability should be prioritised above others, what to conclude when different methods of addressing the same definition of comparability produce different results, and whether operational procedures for maintaining standards should be tied more explicitly to particular definitions of comparability.

References

Adams, R. (2007). Cross-moderation methods. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.) (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Advanced Level Information System (ALIS). (2004). *A level subject difficulties*. The Advanced Level Information System, Curriculum, Evaluation and Management Centre, University of Durham.

Al-Bayatti, M. (2005). A comparability study in GCSE French. A statistical analysis of results by awarding body. A study based on the summer 2004 examinations. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.) (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Association Européenne des Conservatoires (AEC) (2004). *Glossary of terms used in relation to the Bologna Declaration*. http://www.aecinfo.org/glossary%20and%20faq%20english.pdf, accessed October 2009. Not available at this address April 2011.

Baird, J. & Eason, T. (2004). Statistical screening procedures to investigate inter-awarding body comparability in GCE, VCE, GCSE, Applied GCSE and GCSE short courses. AQA. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.) (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

BBC (2010). *A –levels: Rising grades and changing subjects*. BBC news online. 20 August. Available at http://www.bbc.co.uk/news/education-11011564 Accessed on 24th June 2011.

Bell, J.F. (undated). Methods of aggregating assessment result to predict future examination performance. Available at http://www.cambridgeassessment.org.uk/ca/digitalAssets/188917_JBMethods_of_aggregating_assessment_results_to_predict_future_examination_performace.pdf Accessed on June 27th 2011.

Bell, J. F. (2000). Review of research undertaking comparing qualifications. In: J.F. Bell & J. Greatorex (Eds.) *A Review of Research into Levels, Profiles and Comparability*. A report to QCA. London: Qualifications and Curriculum Authority.

Bell, J. F., Bramley, T. & Raikes, N. (1997). Investigating A level mathematics standards over time. *British Journal of Curriculum and Assessment*, **8**, 2, 7–11.

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

Bramley (2011). Comparability of examinations standards: Perspectives from Cambridge Assessment. Seminar. April 6th 2011, Cambridge.

Bramley, T., & Gill, T. (2010). Evaluating the rank-ordering method for standard maintaining. *Research Papers in Education*, **25**, 3, 293–317.

Cambridge Assessment (2010). Exam Standards: the big debate. Report and Recommendations. Available at http://www.cambridgeassessment.org.uk/ca/digitalAssets/189035_Standards_Report.pdf Accessed on June 23rd 2011.

Coe, R. (2007). Common Examinee Methods. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.) (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

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

Conlon, G. & Patrignani, P. (2010). *Returns to BTEC vocational qualifications*. Final Report for Pearson. http://www.edexcel.com/Policies/Documents/Final%20Report%20Returns%20to%20BTEC%20Vocational%20Qualifications%20Fin%E2%80%A6.pdf

Elliott, G. (2008). *Practical cookery in schools at KS3 and KS4: Opinions of teachers about the issues*. Paper presented at the British Educational Research Association Conference, Edinburgh, September, 2008.

Elliott, G. (2011). Comparability of examinations standards: Perspectives from Cambridge Assessment Seminar. April 6th 2011, Cambridge.

Elliott, G., Forster, M. Greatorex, J. & Bell, J.F. (2002). Back to the future: a methodology for comparing old A-level and new AS standards. *Educational Studies*, **28**, 2, 163–180.

Emery, J. L., Bell, J. F. & Vidal Rodeiro, C.L. (2011). The BMAT for medical student selection – issues of fairness and bias. *Medical Teacher*, **33**, 1, 62–71.

Greatorex, J. (2011). Comparing different types of qualifications: An alternative comparator. *Research Matters: A Cambridge Assessment Publication*, Special Issue 2, 34–41.

Greatorex, J. Elliott, G. & Bell, J.F. (2002). A comparability study in GCE AS Chemistry. A review of the examination requirements and a report on the cross moderation exercise. A study based on the Summer 2001 examination and organised by the Research and Evaluation Division, University of Cambridge Local Examinations Syndicate for OCR on behalf of the Joint Council for General Qualifications. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.) (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Green, T. & Jay, D. (2005). Quality assurance and quality control: Reviewing and pretesting examination material at Cambridge ESOL. *Research Notes*, **21**, 5–7. Available at http://www.cambridgeesol.org/rs_notes/rs_nts21.pdf accessed on June 24th 2011.

Harvey, L. (2004–11). *Analytic Quality Glossary*. Quality Research International. http://www.qualityresearchinternational.com/glossary/ accessed on April 14th 2011.

Hendry, P. (2009). Understanding and using CEM data. Curriculum, Evaluation and Management Centre, University of Durham. Available at: http://www.cemcentre.org.uk/publications accessed on April 19th 2011.

Jones, B.E. (2003). Subject pairs over time: A review of the evidence and the issues. Unpublished research paper RC/220, Assessment and Qualifications Alliance. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.) (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

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

Kolen, M.J., & Brennan, R.L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices*. 2nd ed. New York: Springer.

Murphy, R. (2007). Common test methods. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Newbould, C.A. & Massey, A.J. (1979). *Comparability using a common element*. Cambridge: Test Development and Research Unit.

Newton, P. (2007). Contextualising the comparability of examination standards. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.) (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Newton, P. (2008). *Exploring tacit assumptions about comparability*. Paper presented at the 34th Annual Conference of the International Association for Educational Assessment. 7–12 September 2008. Cambridge, United Kingdom.

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

Newton, P. (2011). Comparability of examinations standards: Perspectives from Cambridge Assessment Seminar. April 6th 2011, Cambridge.

Newton, P., Baird, J.-A., Goldstein, H., Patrick, H., & Tymms, P. (2007). *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

OED (2011). Oxford English Dictionary online. June 2011. Oxford University Press. Available at http://www.oed.com/ accessed on 28th June 2011.

Ofqual (2009a). The new GCSE science examinations. Findings from the monitoring of the new GCSE science specifications: 2007 to 2008. Available at http://www.ofqual.gov.uk/files/ofqual-09-4148_GCSE_science_2007_2008_report.pdf accessed on June 27th 2011.

Ofqual (2009b). Review of standards in GCSE English literature. Available at http://www.ofqual.gov.uk/files/ofqual-09-4154_Review_of_standards_English_lit_2000_2007-1.pdf accessed on June 27th 2011.

Ofqual (2011a). Glossary. Available at http://www.ofqual.gov.uk/help-and-support/94-articles/34-161-glossary#_C accessed on June 21st 2011.

Ofqual (2011b). Perceptions of A levels and GCSEs – Wave 9. Available at http://www.ofqual.gov.uk/research-and-statistics/183/537 accessed on 21/6/11

Pollitt, A., Ahmed, A. & Crisp V. (2007). The demands of examination syllabuses and question papers. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.) (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Pollitt, A., Hughes, S., Ahmed, A., Fisher-Hoch, H. and Bramley, T. (1998). *The effects of structure on the demands in GCSE and A level questions*. Report to the Qualifications and Curriculum Authority, December 2003.

QCA (2003). Public confidence in the A level examination system. Research study conducted for Qualifications and Curriculum Authority. Perceptions of A levels and GCSEs – Wave 1. Available at http://www.ofqual.gov.uk/research-and-statistics/183/537 accessed on 21/6/11.

QCDA (2010). Test development, level setting and maintaining standards. Available at http://orderline.qcda.gov.uk/gempdf/1445908166/QCDA_Assessments_test_setting.pdf accessed on June 24th 2011.

Tattersall, K. (2007). A brief history of policies, practices and issues relating to comparability. In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms (Eds.) (2007), *Techniques for monitoring the comparability of examination standards*. London: Qualifications and Curriculum Authority.

Vidal Rodeiro, C. L. & Nádas, R. (2011). The effects of GCSE modularisation: a comparison between modular and linear examinations in secondary education. *Research Matters: A Cambridge Assessment Publication*, **11**, 7–13. Available at http://www.cambridgeassessment.org.uk/ca/digitalAssets/189984_Rearch_Matters_11_2011.pdf accessed on June 23rd 2011.

Willmott, A. (2005). Thinking Skills and Admissions. A report on the validity and reliability of the TSA and MVAT/BMAT assessments. Available at http://www.cambridgeassessment.org.uk/ca/digitalAssets/113977_Thinking_Skills___Admissions_a_report_on_validity.pdf accessed on June 24th 2011.

A level pass rates and the enduring myth of norm-referencing

Paul Newton Director, Cambridge Assessment Network, Assessment Research & Development

Defining standards and comparability

Comparability can be defined as the application of the same standard across different examinations (Newton, 2007, p.9). If so, then, to understand comparability, you need to understand what it means to apply a standard. The meaning of comparability, in the context of A level examining, has been confused for decades because of a lack of clarity over the definition of the A level standard.

The mythology of A level examining goes as follows: standards were norm-referenced, from the 1960s until the middle of the 1980s, after which they became criterion-referenced. This article will argue that A levels have never been norm-referenced, have never been criterion-referenced and have always been attainment-referenced. However, to make this case, these terms need to be defined with some precision. Crucially, quite distinct versions of these definitions can be identified within different contexts, so we need to focus specifically upon the way in which the terms have traditionally been understood in the context of UK examinations.

The idea of an examination standard being [x]-referenced means that it is linked to, or defined in terms of, the [x]. In criterion-referencing, the standard is defined in terms of written criteria, the satisfaction of which results in the award of a particular grade. As understood in the context of UK examinations, this has typically been related to the notion of 'mastery' testing, such that, for the award of a particular grade, students must have demonstrated a certain level of proficiency across each of the sub-domains that comprise a curriculum area, that is, they must have mastered all of the critical elements of the domain. You can imagine this in terms of a student needing to have achieved at least the passing grade on each of the principal sub-domains of mathematics (e.g. number & algebra; geometry & measures; statistics & probability) for the award of a pass overall. The pass would thereby certify that the student had 'mastered' all elements of the mathematics curriculum.[i] Criterion-referencing involves identifying exactly what students can and cannot do, in each sub-domain of the subject being examined, and then awarding: grade A to those who have satisfied all of the grade A criteria across all of the sub-domains; grade B to those who have satisfied all of the grade B criteria across all of the sub-domains; and so on.

Criterion-referencing contrasts with attainment-referencing, in which the standard is defined more holistically in terms of a certain level of attainment. Instead of judging students on the basis of their profile of attainment across sub-domains, in terms of clearly specified performance criteria, they are judged on the basis of their overall level of attainment in the curriculum area being examined. In effect, instead of there being a set of criteria for the award of the overall grade, there is just a single criterion. In practice, the idea of unambiguously articulating this single criterion, at such a high level of abstraction, turns out to be so implausible as to force the examiner to drop any pretence of referencing standards to written criteria. All that can be done is to provide a general indication of the kinds of knowledge, skill and understanding that might well be associated with the award of a particular grade. In UK examinations, attainment-referenced standards are currently exemplified (not defined) through 'performance descriptions' (not 'performance criteria') relating to hypothetical 'typical' students. Attainment-referencing involves ranking students, in terms of their overall level of attainment, and then awarding: grade A to students with a certain level of attainment (i.e. the level at which students were awarded the same grade in previous years); grade B to students with a lower level of attainment; and so on.[ii]

Finally, in norm-referencing, the standard is defined in terms of a particular norm-group. When used in the context of UK examinations, the norm-group is simply the cohort that took a particular examination at a particular time. So the norm-referenced standard simply represents the level of attainment of a particular student in relation to the level of attainment of all other students who sat the examination in question. Importantly, both attainment-referencing and norm-referencing rank students in exactly the same way, on the basis of their overall level of attainment in the curriculum area. All that differs is how standards are set for the award of each grade. Norm-referencing involves ranking students, in terms of their overall level of attainment, and then awarding: grade A to the top X%; grade B to the next Y%; and so on.

The distinction between norm-referencing and criterion-referencing came from the North American literature on educational measurement. Glaser (1963/1994) explained that: "When such norm-referenced measures are used, a particular student's achievement is evaluated in terms of a comparison between his performance and the performance of other members of the group" (p.7). This broad definition resonates somewhat with the UK usage, although it is not identical, since the latter is specific in referring to the award of grades to fixed percentages of each examination cohort, a practice known in the USA as 'grading on the curve'. Nowadays, in the USA and elsewhere, norm-referencing tends to have a more specific definition, which departs even further from the UK usage: "A norm-referenced test (NRT) is a type of test, assessment, or evaluation which yields an estimate of the position of the tested individual in a predefined population, with respect to the trait being measured" (Wikipedia, 2011). For example, results from a particular administration of an IQ test would not indicate how well you performed in relation to others tested at the same time, but in relation to the spread of scores that might be expected within the entire population. Wiliam (1996) proposed that the term 'cohort-referencing' characterises UK usage more precisely; although we will remain with the more conventional term for the remainder of the present article.

By way of summary, each of these definitions has different implications for comparability: norm-referencing specifies that students with the same rank (from their respective examinations) should be

awarded the same grade; criterion-referencing specifies that students with the same profile of proficiency (from their respective examinations) should be awarded the same grade; attainment-referencing specifies that students with the same overall level of attainment (from their respective examinations) should be awarded the same grade.[iii]

The myth

The primary aim of the present article is to dispel a widely-believed myth, which goes something like this:

> *For the first 25 years or so, the maintenance of standards of A-levels relied substantially on the constraints of so-called norm referencing, i.e. a constant proportion of candidates in each subject was each year awarded the same grade. […] it differentiates only between those who took the test at the same time and the results have no validity from one year to another.* (Stubbs, 2002, p.4)

This is a quotation from an article by Sir William Stubbs, who was Chairman of the QCA until September 2002; certainly not a casual observer of the system. Back in the olden-days, so the story goes, we used to define the A level standard in terms of norm-referencing. This meant that we awarded the same profile of grades across subjects, across boards and each year; regardless of how well the students had actually performed in each subject, each board and each year. Norm-referencing was, therefore, blind to the quality of work produced by students. Indeed, Sir Bill went so far as to describe this detachment from attainment as a source of invalidity, at least in relation to trends over time. The implication is that pass rate trends could not be interpreted as evidence of trends in national attainment over time; i.e. national attainment could be rising or falling but pass rates would still remain the same. The maintenance of A level standards, from a norm-referencing perspective, is straightforward: to apply the same standard, for any examination cohort, all you have to do is to apply the same percentage pass rate.

Although the myth of norm-referencing predates the 1960s, 1960 represents a very important chapter in this story. It saw the Third Report of the Secondary School Examinations Council, on A level examinations and the case for their reform (SSEC, 1960).

The A level was originally a pass/fail examination, certifying that students were qualified for university entry. However, by 1960, it had increasingly become an instrument for competitive selection. This meant that university selectors had started asking for numerical marks; which, in turn, had led to "an unhealthy competition in cramming and mark grubbing" by students (see SSEC, 1960, p.3). Moving from a pass/fail system to a graded system was supposed to remedy this.

The SSEC report proposed that there should be five passing grades, from A to E, and a compensatory O level pass. Although it did not actually specify how standards should be set or maintained, it did recommend that grades should be distributed roughly as described in Figure 1.

Grade	*Cum. %*	*Grade*	*Cum %.*
A	10	A	10
B	25	B	15
C	35	C	10
D	50	D	15
E	70	E	20
O pass	90	O pass	20
Fail	100	Fail	10

Figure 1: Recommendations from SSEC (1960)

The straightforward interpretation of these recommendations was as follows: irrespective of any possible difference in calibre of students between subjects, between boards or from year-to-year, the same percentage of students should be awarded each grade. That is, 70% should pass in German and 70% in Economics; 70% should pass with the Cambridge examining board and 70% with the London examining board; 70% should pass in 1960 and 70% in 1986.

Indeed, when looked at from a certain perspective, evidence does seem to suggest that this happened. A graph from the Centre for Education and Employment Research, at the University of Buckingham (BBC, 2010), nicely illustrated a striking norm-reference-like stability in the overall A level pass rate, from the early 1960s until the early 1980s (presenting data aggregated across subjects and across boards). From the early 1980s onwards, the pass rate rose steadily. An earlier report from the School Curriculum and Assessment Authority (SCAA) with the Office for Standards in Education (Ofsted) observed the same trend. A press release from SCAA (SCAA, 1996), which accompanied the report, read as follows:

> *From 1962 to 1986, the proportion of candidates to be awarded each grade in major A level subjects was effectively fixed, so no increase could take place even if candidates' performance improved. This was changed in 1987, when key grades were matched to the quality of candidates' work. This change from 'norm-referencing' to 'criterion-referencing' has permitted an increase in the proportion of candidates being awarded grades.* (SCAA, 1996)

This is the myth of norm-referencing: A level standards were norm-referenced, from the early 1960s until 1987, when they switched to being criterion-referenced.

The reality

It is straightforward to dispel the myth of norm-referencing, with reference to examinations results data that have entered the public domain every year since the A level came into existence.

Figure 2 represents data from a report published by the University of Cambridge Local Examinations Syndicate (UCLES, 1980). It illustrates differences in pass rates between syllabuses in different subject areas. The lowest pass rate was only 21%, which is clearly a vast distance from the supposedly recommended 70%. Admittedly, only 42 candidates were examined in accounting. Perhaps, then, only large-entry subjects were norm-referenced? The evidence suggests otherwise. The pass rate in

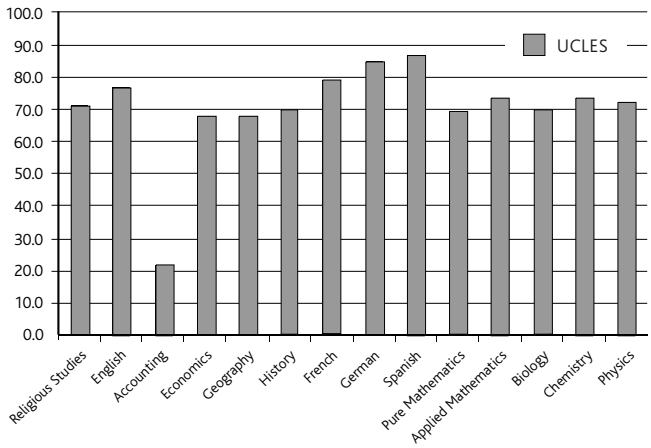


Figure 2: A level pass rate with UCLES, Summer 1980 (Home candidates only)

German was 85% (n=1336); in English literature, 77% (n=7519); and in economics, 68% (n=2699). This does not look much like norm-referencing.

Figure 3 adds data from the Associated Examining Board (AEB, 1980). It illustrates differences in pass rates between boards, within the same subject area. For geography, the pass rate was 68% for UCLES (n=4884) versus 48% for AEB (n=2247); for chemistry, the pass rate was 73% for UCLES (n=3288) versus 44% for AEB (n=2389); for French, the pass rate was 79% for UCLES (n=3335) versus 66% for AEB (n=2318). Moreover, the UCLES pass rates were almost universally higher than the AEB pass rates. Again, this does not look much like norm-referencing. Instead, it seems that the boards were quite clearly aiming to reflect differences in student calibre; both across subjects and across boards.

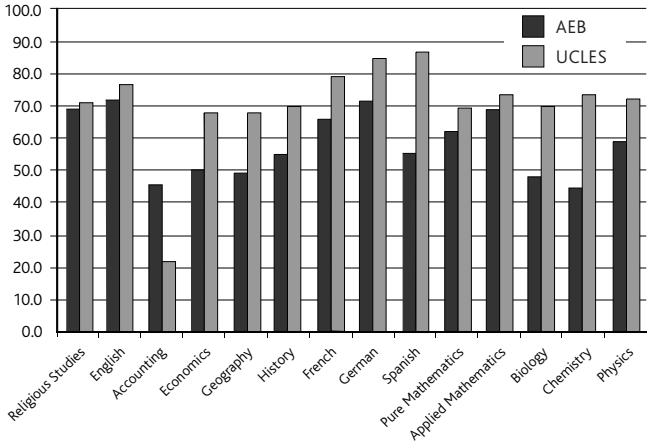


Figure 3: A level pass rate with AEB & UCLES, Summer 1980 (Home candidates only)

Maybe, though, the principle of norm-referencing was only applied within boards, within subjects, from one year to the next? Again, the evidence suggests otherwise. Figure 4 represents data from UCLES Annual Reports, from 1960 until 1986. Data from only four subjects were collated, for illustrative purposes. These are syllabus-level data – the level at which awarding decisions are made – so, if norm-referencing is to be found anywhere, it ought to be found in results like these.

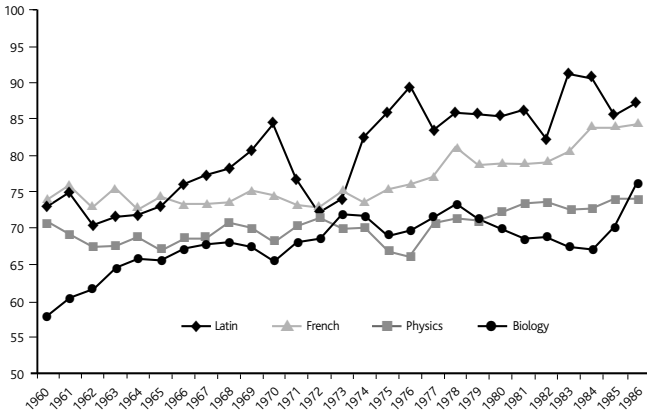


Figure 4: A level pass rates for UCLES (Summer, Home candidates only)

Even at this basic (non-aggregated) level the interpretation of the data is not entirely straightforward, particularly since syllabuses have a tendency to wax and wane in popularity and to be replaced. Where, for instance, there are two syllabuses in a subject area (with different pass rates), which are replaced the following year with a single syllabus (with a new pass rate), which of the two year 1 pass rates should the new year 2 pass rate be linked to? Of the four subjects in Figure 4, physics was the most tricky to interpret in this respect. For instance, in 1964 there was only one syllabus (physics), while in 1965 there were two (physics N and physics T); N gradually lost candidates while T gained them, until 1972, when there was only one syllabus again. To complicate matters, from 1974, physics split into physics and Nuffield physics. Similarly, 1974 saw biology split into Nuffield biology and social biology. Results are only presented for the highest entry 'conventional' syllabus (i.e. excluding Nuffield syllabuses and social biology). Results seem most straightforward to interpret for Latin and French, as these appeared to be essentially the same syllabuses from 1960 to 1986 (although, no doubt, they changed in content and emphasis over time).

Even for Latin, the entry sizes were reasonably large, from a low of 245 (in 1986) to a high of 939 (in 1964). Entries in the other subjects were higher still; in French, for example, ranging from a low of 1779 (in 1960) to a high of 3664 (in 1968). Given the proviso of large entry sizes, notice how:

- French jumped from 77% to 81% in one year (1977 to 1978)
- biology jumped from 70% to 76% in one year (1985 to 1986)
- Latin fell from 84% to 72% in two years (1970 to 1972).

Indeed, during the supposed glory-days of norm-referencing, from 1960 to 1986:

- the physics pass rate rose from 71% to 74% (+3%)
- the French pass rate rose from 74% to 89% (+15%)
- the biology pass rate rose from 58% to 74% (+16%)
- the Latin pass rate rose from 73% to 91% (+18%).

This was clearly not norm-referencing. Even though the pass rates do tend to rise somewhat less in certain subjects than in others, and even though there seems to be somewhat more stability in the 1960s than in the 1980s, it is still clearly impossible to claim that UCLES was norm-referencing at any point in time in any of these subjects.

Indeed, despite considerable research, I have uncovered no evidence that any board ever operated a principle of norm-referencing, although I have uncovered much evidence to the contrary. Choppin (1981) quoted Richard Christopher, Secretary of the Joint Matriculation Board, from 1977:

> *It is often thought that in pursuance of [constant standards] the percentages of candidates passing in a subject are decided in advance […] whereas the deciding factor is in fact the quality of the work presented.* (from Choppin, 1981, p.10)

Of course, when results are aggregated across syllabuses within subject areas, across subject areas and then across boards, year-to-year pass rate changes (for individual examinations) might well average out, giving the appearance of norm-referencing (at the national level). But this is not the same as examination standards actually being norm-referenced.

The confusion

If norm-referencing has never constituted a definition of the A level standard, then why does the myth persist? The answer is that something resembling it used to be used – and still is used – as a central component of grade awarding. I shall call it the Similar Cohort Adage.

The A level standard has always been defined in terms of a certain overall level of attainment, that is, the A level examination has always, essentially, been attainment-referenced. Indeed, this conception predates the A level and was a feature of the Higher School Certificate (from 1918) and of examinations which preceded that. However, equally engrained in the psyche of school examining in England is respect for the following adage: if the cohort hasn't changed much, then don't expect the pass rate to change much either.

As Christopher explained, the deciding factor in grade awarding has always been the quality of the work presented. But there has always been a healthy respect for commonsense and statistics too. So boards triangulate evidence from examiner judgement of scripts, with statistical expectations of pass rate stability, to decide where grade boundaries ought to lie. They did this in 1951, in 1960, in 1987 and we do it today.

If the principle of norm-referencing dictates 'any cohort – same pass rate' the Similar Cohort Adage recommends 'similar cohort – similar pass rate'. It is a rule-of-thumb that the examining boards in England have taken to heart and have integrated within their methodologies for maintaining standards. The Joint Matriculation Board set out its stall very clearly, in 1951, right at the outset of A level examining:

> *Many years ago in the light of its experience the Joint Matriculation Board reached the conclusion that the procedure fairest to the candidates was to award approximately the same percentages of passes, credits, very goods, etc. in every subject each year, the percentages being based upon the number of entries for the subject concerned. The Board insisted however that three strictly limiting conditions must all be satisfied.*
>
> *A. The subject entry must be large, many hundreds of candidates at least.*
>
> *B. The entries must come from a wide enough area to obviate the influence of special local conditions.*
>
> *C. There must be no reason to think that the general nature of the entries is changing.*
>
> (JMB, 1951)

The very fact that limiting conditions were identified illustrates that the JMB was not defining a principle of grade awarding, it was simply describing a rule-of-thumb to support grade awarding practice. The standard was defined in terms of student attainment; attainment-referencing.

Over the years, approaches to operationalising the Similar Cohort Adage have evolved. Early on, the boards had to rely on general impressions of whether, or how, cohorts were changing. To the extent that there were more boards in the 1960s, with smaller entries that were more locally based, this was more manageable. Fortunately, as entries have increased in size and boards have become less locally based, their statisticians have also become more sophisticated. With procedures like the delta analysis, they became better able to adjust statistical expectations, according to gender and school-type differentials (e.g. Eason, 1995). More recently, the boards have routinely established statistical expectations on the basis of prior attainment in the General Certificate of Secondary Examination (e.g. Pinot de Moira, 2008). Yet, conceptually speaking, the boards are not using statistical expectations any differently now from how they were used 50, or even 100, years ago.

The other myth

If the boards have always attainment-referenced and never norm-referenced then they can never have criterion-referenced either. Nor have they. This is the other myth. Criterion-referencing was certainly being considered, during the 1970s and 1980s, as a possible alternative approach to defining standards. Aspirations were particularly high, in some quarters, that the new 16+ examination (which was to become the General Certificate of Secondary Education) would be completely criterion-referenced. Keith Joseph proclaimed as much, in his 1984 North of England speech:

> *Second, we should move towards a greater degree of criterion-referencing in these examinations and away from norm-referencing.* (Joseph, 1984)

Yet, the GCSE was introduced with the traditional approach to grade awarding and with good justification for not criterion-referencing (see Cresswell, 1987; Cresswell & Houston, 1991). Despite high aspirations lingering for some time, criterion-referencing ultimately:

> *[…] died a death, other than taking the much weaker form of grade descriptions.* (Tattersall, 2007, p.70)

What actually happened in 1987

There is a grain of truth in the claim that norm-referencing came to an end in 1987. Only an extremely small grain, though. What happened was the result of a long campaign, spearheaded by the Joint Matriculation Board, to correct the narrow grade C problem; a problem that could be traced to the percentages recommended in 1960 by the SSEC.

Proposals for the reform of A level, within SSEC (1960), included the development of special (S) papers, based upon the same syllabus as A level papers, but examined at a higher level, giving abler candidates the opportunity to demonstrate their excellence. Discussions which preceded the publication of the 1960 report had concluded that S paper grades would only be awarded to students who had achieved at least grade B on their A level papers. Subsequent discussions led to the conclusion that it would be useful to lower this hurdle to grade C, just as long as the number of students who might additionally qualify was not too large. This is why the SSEC recommendations (in Figure 1) proposed a relatively narrow grade C band of 10% of candidates. As it happens, a ruling following the 1965 session lowered the hurdle to grade E. But the broad structure of the A level grade distribution had been established by then.

An unfortunate consequence of the narrow grade C was an increased likelihood of significant error in the award of grades. Students receive marks that fail to reflect their true level of attainment for all sorts of reason, from their own state of concentration on the day of the examination to errors made by clerical staff whilst inputting mark data. When grade boundaries lie only a small number of marks apart, the impact of this kind of error can be significant, for example, a genuine grade B student might end up being awarded grade D (or vice versa). The narrower the width between grade boundaries the more frequently these significant impacts will occur; an effect that is exacerbated when narrow grades fall at the middle of the mark distribution where the largest number of candidates often cluster.

In 1969, the JMB proposed an alternative grading system, based upon the following procedure:

1. first, set the top grade boundary;
2. then, set the passing grade boundary;
3. "All 'passing' candidates between the two fixed points would be placed on the agreed scale, by the application of a simple formula, strictly in accordance with the proportion of marks gained." (JMB, 1969, p.5)

In short, instead of grade distributions which had evolved on the back of a proportional division of *candidates*, the proposed approach would locate grade boundaries on the basis of a proportional division of *marks*. This would mitigate the problem of a narrow grade C. Unfortunately, after a concerted effort on behalf of the Schools Council, and major consultations with stakeholders, the Secretary of State dismissed the proposals in 1972. Of course, this did not resolve the problem.

A decade later, the JMB wrote another paper which, in effect, reminded readers of the problem of a narrow grade C and of the solution recommended in 1969 (JMB, 1983). This was debated for another three years, before a solution was finally agreed. The solution was not easy to reach. Although the similarities between boards in their approaches to maintaining standards far outweighed their differences, they were still somewhat precious over their differences and disagreed over how to resolve the problems of the A level grading system. A compendium of consultation responses received by the Secondary Examinations Council revealed comments such as the following:

> *The Committee accepted the need for a review and is not opposed, in principle, to a change in procedures for Advanced level grading. However, the present proposals are unacceptable to the Committee, primarily because of their effect on the level of difficulty of grades B and C but also because Chief Examiner judgements would come into play at only two grade boundaries*.
> (Letter from M.J. Jones, Examinations Secretary to the Welsh Joint Education Committee, to Sir Wilfred Cockcroft, 25 February 1985)

> *The Delegates point out that the SSEC guidelines which are at present used are no more than that: they are not rules to be rigidly followed, and while the effect of their application is one of the criteria studied by the Delegacy's Awarders in coming to their decisions, the final judgement is always made in the light of the actual work offered by the candidates. Flexibility is vital, in order to be free to do justice in subjects for which rigid rules are inappropriate, and also to avoid any unfairness to candidates which might arise from the use of an intractable framework. This is a matter on which the Delegacy's Awarders feel very strongly. The Delegates understand that there is to be no flexibility allowed in the use of the proposed system; this they deplore.*
> (Letter from C.G. Hunter, Secretary to the Delegates, University of Oxford Delegacy of Local Examinations, to Sir Wilfred Cockcroft, 4 March 1985)

> *The use of two fixed points in fact corresponds to this Board's practice over many years, where determination of A level grade boundaries commences with the establishment of what, in the judgement of the Awarders in consultation with the Secretaries, are the appropriate placings of the B/C and E/O boundaries*.
> (Letter from H.F King (Cambridge Secretary) and K. Schoenberger (Oxford Secretary), Oxford and Cambridge Schools Examinations Board, to Sir Wilfred Cockcroft, 29 March 1985)

> *Our views on which boundaries should be fixed by the quality of work shown in the scripts of candidates are divided. It could be that in the interests of public confidence in standards, the scrutiny of scripts at the A/B boundary is desirable. However the fixing of the B/C boundary in this way, involving as it does the inspection of the work of a larger sample of candidates, could produce a more 'reliable' performance indicator. On balance the weight of the argument seems to lie with determining the A/B boundary in this way rather than the B/C boundary.*
> (Letter from E.J. Bolton, Senior Chief Inspector, Department of Education and Science, to Sir Wilfred Cockcroft, 3 April 1985)

The University of London Schools Examining Board went a step further, arguing that changing the grading system in advance of the final outcome of research into the development of grade criteria was short-sighted. It recommended awaiting the outcome of a more fundamental review.

Ultimately, the solution to the problem of the squeezed middle was as follows:

> *There will be no change in the way grade A is awarded. The cut-off scores for grades B and E will be determined on the basis of the professional judgement of the examiners (and as this is, by and large, what is done at the moment there will be little change).*
>
> *Once the cut-scores for B and E are set, the minimum scores for D and C will be obtained by dividing the mark-scale into three equal sections*.
> (SEC, 1986)

Ironically, for some boards, the agreed procedure meant less reliance upon the judgement of awarders; not more reliance, as might have been assumed would be the outcome of the 'rejection of norm-referencing' (which, of course, it never was). For other boards, the 1986 ruling, to be operationalised from summer 1987, represented little more than business as usual. At the critical pass/fail boundary, at least, there was no change in procedure for maintaining standards, for any board.

A possible alternative explanation

Before concluding, an important alternative explanation for trends in pass rates over time needs to be considered; one that potentially resurrects the idea that examining boards were attempting to norm-reference after all.[iv]

Earlier, the principle of norm-referencing, in the context of UK examinations, was defined in terms of a requirement for awarding bodies to award the same percentage of students each grade in each examination. It was demonstrated that this principle was never observed since, at the syllabus level, different pass rates were evident across subjects, across boards and over time. What, though, if the principle were to be interpreted in terms of the *national* cohort of students attempting an examination in each subject area, rather than in terms of local cohorts of students attempting the same examination across individual boards? Perhaps the SSEC (1960) recommendations should be read as follows: 70% of students should pass each subject, at a national level, with candidate success spread across boards according to the calibre of students entered. For each subject, then, boards attracting more able students would pass more than 70% of students, and boards attracting less able students would pass fewer than 70% of students, such that the

national pass rate would average out to 70%. At first glance, this interpretation seems attractive. After all, there are clear indications that examining boards did indeed adjust pass rates to reflect the calibre of the entry for each examination. But, in doing so, were they ultimately aiming to norm-reference at a national level?

Against the national cohort interpretation of norm-referencing are a number of telling observations. For one, the boards themselves claimed not to be norm-referencing (see Choppin, 1981). Moreover, right from the outset of A level examining, it is possible to find very clear statements from the boards that they: (i) operated what I have called the Similar Cohort Adage at a local cohort level; but (ii) would not actually operate it if they had good reason to suppose that the general nature of an entry had changed (see JMB, 1951). In short, the boards were explicitly open to the possibility of pass rate change; even if, in practice, the national pass rate tended to remain fairly stable from the 1960s to the 1980s. Finally, if norm-referencing (*à la* national cohort interpretation) did indeed provide the explanation for pass rate stability until the 1980s, then who made the decision to stop applying the principle and for what reason? As explained earlier, the change in grade awarding procedure during the late 1980s – which was incorrectly described as the end of norm-referencing – did not significantly affect the way that passing grades were decided. It there had been an explicit policy decision to stop norm-referencing during the early- to mid-1980s, then it would surely have been well documented.

To evaluate the plausibility of the national cohort interpretation in more depth would require an analysis of data from all boards, going back to the 1960s and 1970s, to examine whether the cross-board, subject-level pass rates did indeed tend to average out to around 70%. These data, whilst potentially available, have not been collated for analysis to date. In the mean time, it seems more parsimonious to conclude that the local cohort interpretation of norm-referencing is both the conventional interpretation and a myth.

Conclusion

The idea that A level examination standards operated on a principle of norm-referencing until 1987, when they switched to a principle of criterion-referencing, is mythological but clearly false. In terms of the theory of grade awarding, 1987 saw:

- no rejection of norm-referencing as a principle (since it never has been assumed);
- no adoption of criterion-referencing as a principle (since it never has been assumed);
- no rejection of attainment-referencing as a principle (since it has always been assumed).

In terms of the practice of grade awarding, 1987 saw:

- no adoption of script comparison as a method (since examiner judgement has always been used);
- no rejection of the Similar Cohort Adage as a method (since statistical expectations have always been used).

Although the evidence which demonstrates this state of affairs is not always easy to locate, it is surprising that even official sources buy into the myth. One reason may be that the myth seems to provide a neat explanation for apparent changes in pass rates over time. At a national level, it is the case that pass rates have risen substantially since the 1980s; although, admittedly, they began their ascent during the earlier, rather than later, part of that decade. If A level awarding procedures did not change radically during the 1980s, especially not at the passing grade, then the pass rate trends are doubly remarkable. If we are to interpret the overall, national pass rate trend line at face value, then not only did student attainment rise substantially over the past three decades (during the 1980s, 1990s and 2000s), it rose from a baseline of no substantial change over the preceding two decades (during the 1960s and 1970s).

The only alternative explanation is that, despite the A level awarding process not having changed radically during the 1980s, more subtle changes were taking place, and these somehow affected the way in which grades were being awarded. This is certainly an intriguing possibility; but one beyond the scope of the present article.

References

AEB. (1980). *Statistics June 1980*. Hampshire: Associated Examining Board.

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

BBC. (2010). *A-levels: Rising grades and changing subjects*. 20 August. British Broadcasting Corporation Website. Accessed 24 March 2011. http://www.bbc.co.uk/news/education-11011564.

Choppin, B. (1981). Is education getting better? *British Educational Research Journal*, **7**, 1, 3–16.

Cresswell, M.J. (1987). Describing examination performance: grade criteria in public examinations. *Educational Studies*, **13**, 3, 247–265.

Cresswell, M.J. & Houston, J.G. (1991). Assessment of the national curriculum – some fundamental considerations. *Educational Review*, **43**, 1, 63–78.

Eason, S. (1995). *A review of the Delta Analysis method for comparing subject grade distributions across examining boards*. Guildford: Associated Examining Board.

Glaser, R. (1963/1994). Instructional technology and the measurement of learning outcomes: some questions. *Educational Measurement: Issues and Practice*, **13**, 4, 6–8.

Joseph, K. (1984). *Speech by the Rt Hon Sir Keith Joseph, Secretary of State for Education and Science*. The North of England Education Conference, Sheffield, 6 January.

JMB. (1951). *JMB Standardization and the GCE*. Manchester: Joint Matriculation Board.

JMB. (1969). *Advanced level grades: suggestions for a new system of reporting results*. Manchester: Joint Matriculation Board.

JMB. (1983). *Problems of the GCE Advanced level grading scheme*. Manchester: Joint Matriculation Board.

Newton, P.E. (2007). Contextualising the comparability of examination standards. In: P.E. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.). *Techniques for monitoring the comparability of examination standards*. 9–42. London: Qualifications and Curriculum Authority.

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

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

Newton, P.E. (2010c). Conceptualizing comparability. *Measurement: Interdisciplinary Research and Perspectives*, **8**, 4, 172–179.

Pinot de Moira (2008). *Statistical predictions in award meetings: how confident should we be?* RPA_08_APM_RP_013. Guildford: AQA.

SCAA (1996). *Standards at GCSE and A level*. Press Release, 96/52. 5 December.

SEC. (1986). *SEC News. Number 3*. London: Secondary Examinations Council.

SSEC. (1960). *Examinations in Secondary Schools: Third Report of the Secondary Schools Examinations Council*. *The General Certificate of Education and Sixth Form Studies*. London: HMSO.

Stubbs, W. (2002). Gold standards and A-levels. *Education Today*, **52**, 4, 3–8.

Tattersall, K. (2007). A brief history of policies, practices and issues relating to comparability. In: P.E. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.). Techniques for monitoring the comparability of examination standards. 43–91. London: Qualifications and Curriculum Authority.

UCLES. (1980). *Statistics 1980*. Cambridge: University of Cambridge Local Examinations Syndicate.

Wikipedia. (2011). "Norm-referenced test", http://en.wikipedia.org/wiki/Norm-referenced_test, accessed 03 04 2011.

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

ENDNOTES

i. Of course, this raises the question of exactly how the pass standard for each sub-domain is defined and, likewise, how sub-domain standards beyond the passing grade are defined. The idea of criterion-referencing is most plausible for straightforward judgements (e.g. has vs. has not mastered) of low-level competences (e.g. the ability to add two single-digit numbers). In situations like this, the standard can be specified through fairly unambiguous written criteria (e.g. a student who demonstrates that they can add the full range of single-digit numbers, consistently over time and contexts, satisfies the criteria and can therefore be said to have mastered the ability to add two single-digit numbers). The more complex the judgement required (e.g. grade C standard vs. grade B standard) and the higher-level the competence in question (e.g. proficiency in statistics and probability) the less plausible criterion-referencing becomes. In situations like this, the standard is far less amenable to specification through unambiguous written criteria. Distinguishing the defining characteristics of competence at one level from competence at another becomes extremely complicated, as the competence becomes increasingly multifaceted and as students exhibit competence in a multiplicity of ways. Thus, the quest for unambiguous written criteria for the award of grades soon turns into a struggle to articulate even rough impressions. The rougher the impression conveyed by the written criteria the less meaningful the very idea of criterion-referencing becomes.

ii. The pros and cons of criterion-referencing and attainment-referencing are described in more depth in Baird, Cresswell & Newton (2000). The present article describes as attainment-referencing that which Baird, *et al* (2000) called weak-criterion-referencing. In retrospect, the term weak-criterion-referencing was not ideal. Attainment-referencing is conceptually quite distinct from criterion-referencing, not simply a weaker version.

iii. There are many ways of cutting the comparability cake; that is, many ways of expressing alternative conceptions of comparability (see Newton, 2010a). In terms of the tripartite framework presented in Newton (2010b,c), attainment-referencing and criterion-referencing would most likely be classified as examples of phenomenal definitions, defining comparability in terms of the observable phenomena of attainment (knowledge, skill and understanding). Norm-referencing, however, could not be classified within that framework (as either a phenomenal, causal or predictive definition). A fourth category of definitions would need to be introduced to accommodate it: competitive. Within norm-referencing, the standard is defined in much the same way as it is in any sporting competition: gold for first, silver for second, and so on. Results do not testify to an absolute standard ('the student has attained a, b or c'), merely to a relative one ('student 1 has attained more than student 2').

iv. Thanks to an anonymous reviewer for this suggestion.

Subject difficulty – the analogy with question difficulty

Tom Bramley Assistant Director, Research Division, Assessment Research & Development

Introduction

Concerns about differences in the difficulty of examination subjects are not new. Moreover, there have been considerable differences in opinion over i) how subject difficulty should be defined; ii) whether and how it should be measured (represented numerically); and iii) whether individual results in examination subjects should be adjusted to 'allow' for differences in difficulty as defined and measured in some particular way. See Newton (*in press*) for a review.

The purpose of this article is to explore in some depth one particular way of defining and measuring subject difficulty – a way that will be called the 'IRT approach'. This approach has been investigated in Australia in the context of university admissions by Tognolini and Andrich (1996) and in the Netherlands by Korobko, Glas, Bosker, and Luyten (2008), and has recently been advocated in the UK context by Robert Coe at the CEM centre in Durham (Coe 2008, Coe *et al*., 2008).

This article is structured as follows. First the IRT approach is briefly described. Then the analogy of using the IRT approach when the 'items' are examination subjects is explored. Next the task of defining difficulty from first principles is considered, starting from the simplest case of comparing two dichotomous items within a test. The thinking of Louis Guttman on scales and dimensionality is shown to provide a useful framework for understanding difficulty, and the link between Guttman and IRT is described. Finally, an alternative to the IRT approach, based on producing visual representations of differences in difficulty among just a few (three or four) examinations, is offered as an idea for future exploration.

Item Response Theory

Item Response Theory (IRT) is concerned with modelling the scores obtained on the *items*[1] *on a test*, rather than scores or grades obtained on a whole test (or on a composite of several tests). It (IRT) is not limited to educational tests – for example, it is quite widely applied in psychological testing more generally and in healthcare, but the educational context is the only one considered here. An overview of IRT can be found in Yen and Fitzpatrick (2006).

The organising concept of IRT is that of the 'latent trait' or continuum – an abstract line representing whatever the test concerned is supposed to be measuring. The most commonly used unidimensional IRT models contain a single parameter that represents person location on the trait (usually referred to as their 'ability') and one or more parameters characterising the item. In the simplest IRT model, the 1-parameter IRT model for dichotomous items, each item is characterised by a single parameter representing its location on the trait (usually referred to as its 'difficulty'). The 1-parameter model expresses the probability of a person with a given ability succeeding on (i.e. answering correctly) an item with a given difficulty as a function of the difference between ability and difficulty.

The 2- and 3- parameter IRT models for dichotomous items include extra parameters that represent 'discrimination' and 'guessing' respectively. The latter is often used for multiple-choice tests. IRT models for polytomous (i.e. multiple-mark) items also exist. These contain parameters representing the thresholds between adjacent score categories on the trait. In a multidimensional IRT model a person's ability is represented as an *n*-element vector rather than by a single number.

There are many reasons why IRT models are used, but the one of most relevance to this article is that (when the data fit the model) estimates of person ability and item difficulty on a common scale can be made when people have answered different subsets of items. This is the principle behind item banking and computerised adaptive testing (CAT), two of the main practical applications of IRT.

It is this feature of IRT that suggests it might have something to offer to the problem of comparing examination subject difficulty, because in most examination systems (and in particular for GCSEs and A levels) the examinees do not all take examinations in the same set of subjects. In applying the 'IRT approach' the different examination subjects have the role of different items in a test. A pass-fail examination could therefore be modelled analogously to a dichotomous item, and a graded test modelled analogously to a polytomous item.

The analogy with item-based IRT

The first issue that potentially weakens this analogy is the lack of clarity about the meaning of the trait when examination subjects are modelled with the IRT approach. When individual items are being modelled, as in 'normal' IRT, it might be something like 'maths ability' (for a maths test). The items in such a test will have been designed according to a specification setting out the criteria (e.g. topics and skills tested) that items must meet in order to be included in the test. In an IRT item banking/CAT scenario the items will also have been screened to check that they do in fact fit the model to a satisfactory degree. An important part of test validation (e.g. Kane, 2006) is to provide evidence of 'construct validity' – in other words that the items do conform to the definition of the trait and that their scores enter into the empirical relationships predicted by the theory of the trait.

However, there is no such deliberate design and validation in the case of examination subjects. The set of possible examinations on offer depends heavily on the cultural and political context of the examination system. In the case of A levels there are currently around 80 possibilities including subjects as diverse as Physical Education, English Literature, Accounting, Chemistry, Latin, Travel and Tourism, Music, and Critical Thinking. If these subjects can be located along a single unidimensional

1 In this report, the terms 'item' and 'question' are used interchangeably.

trait it might be called 'General Academic Ability' (Coe, 2008). While it is a bit optimistic to expect every single subject to be adequately represented on a single line, explorations of the data might reveal a subset of subjects that can more reasonably be thus represented. For example Coe (2008) found that by starting from a group of 37 large-entry GCSE subjects, removing ten that did not fit the model well, and then selectively adding in smaller-entry subjects he was able to include 34 subjects in his final model. Coe (ibid) presented a graph showing the relative difficulty of his set of 34 GCSE subjects: Latin, German, Spanish and French were the 'hardest'; Sport/PE, Textiles, Drama and Media Studies were the 'easiest'. Somewhat surprisingly (given the greater choice and fewer examinations taken, see below), Coe *et al*. (2008) found that only one of 33 large-entry A level subjects did not fit a unidimensional model.

A different approach is to use a multidimensional model splitting the subjects into more natural groupings either on an a priori basis (e.g. sciences, languages) or on the basis of investigating the statistical dimensionality of the data. This was tried by Korobko *et al*. (2008) using pre-university examinations taken in the Netherlands by 18 year olds (i.e. at a similar stage to A level students). They found that a unidimensional model did not fit the data nearly as well as a multidimensional model (which is not surprising), but more interestingly they found that some implausible results were obtained from the unidimensional model in terms of the 'expected scores' imputed to examinees for subjects they had not chosen to take. For example, average scores in French and German imputed to examinees who had mostly chosen science subjects were nearly as high as those actually achieved by examinees who had mostly chosen language subjects, despite the fact that these science students clearly appeared to have less 'language ability' than the language students on the basis of their scores on the (compulsory) examinations in Dutch and English. This apparent anomaly disappeared when a multidimensional model was used. Korobko *et al*. (ibid) produced tables showing the estimated grade point averages (GPAs) obtained from their models – that is, the average grades in each subject that would have been obtained if all students had taken each subject (interestingly, Latin came out as the 'easiest' subject, whichever model was used!). Nonetheless, the issue of the meaning of the trait and the interpretation of the 'difficulty' parameter still remains, regardless of how well the data fit any particular IRT model. This is discussed again later in this article.

A second issue that weakens the analogy with item-based IRT is that in most applications of IRT where different examinees have taken different subsets of items they have not had any choice in which items they take. For example, in a CAT the next item will be chosen by the computer according to its item selection algorithm, usually taking account of its current estimate of the examinee's ability plus any content coverage requirements. In on-demand testing where tests are constructed from a calibrated item bank there may be a variety of different test forms (versions) but no choice for the examinee in which form they answer. In contrast, for A levels especially, the examinees have enormous choice open to them in which subjects they take. If these choices are not independent of ability (and it would seem unrealistic to expect them to be) then it is not reasonable to assume that the modelled outcome on not-chosen subjects will be adequately predicted by the model. In statistics the 'missing data' literature (e.g. Rubin, 1976) deals with the circumstances under which the mechanism producing the missing data can be ignored. Korobko *et al*. (2008) tried to incorporate a model for the subject choice process into their IRT model:

> *Since the students can only choose a limited number of subjects, it is reasonable to assume that the probability of choosing a subject as a function of the proficiency dimension … is single peaked: Students will probably choose subjects within a certain region of the proficiency dimension … and avoid subjects that are too difficult or too easy.*
> (Korobko *et al*. 2008, p.144).

This assumption was not supported in a large-scale survey of A level students (Vidal Rodeiro, 2007) where liking for the subject and university/career plans were found to be more important than perceived difficulty as factors influencing subject choice. Nevertheless, it does represent an attempt to tackle the missing data issue. In fact, Korobko *et al*. (ibid) found that including a model for the missing data mechanism did not yield substantively different results, once multidimensionality had been modelled (see above).

A third, perhaps less important, difference between item-based IRT and subject-based IRT is that in the former the ability estimate of examinees will be based on the responses to a relatively large number of items – perhaps 60 dichotomous items, or 60 marks-worth of polytomous items. When a small number of subjects is chosen, in contrast, the ability estimate will be based on only a few 'items' (perhaps three to five in the case of A levels). The number of score categories per subject depends on the grading scale used – it is currently seven for A levels since the introduction of the A* category in 2010. Thus the 'maximum score' for an examinee taking three A levels is 21. Whilst this would not normally be considered a sufficient 'test length' for reliably estimating an individual's 'ability' this is perhaps not such a problem when the focus of the analysis is on estimating the difficulty parameters for the items (i.e. the subjects).

Definition of difficulty in the IRT approach

One of the reasons why debates about comparability, standards and subject difficulty have been so protracted and inconclusive is that those involved have often disagreed about the most appropriate definition of these and related terms. That there is this disagreement is of course recognised:

> *… 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, 2010, p.289)

Two recent articles by Newton (ibid) and Coe (2010) give thoughtful analyses of these definitional problems. Their arguments will not be repeated here, but one important insight of Newton's is the importance of distinguishing between definitions and methods:

> *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, 2010, p.288)

The 'IRT approach' as described in this article has been used as a method for monitoring whether comparability has been achieved, by retrospectively analysing examinations data. How was difficulty defined by the authors of the articles that have been described previously?

Korobko *et al*. noted that using GPAs results in

> *... systematic bias against students enrolled in more rigorous curricula ... A lower GPA may not necessarily mean that the student performs less well than students who have higher GPAs; the students with the lower GPAs may simply be taking courses and studying in fields with more stringent grading standards*.
> (Korobko *et al*., 2008, p.144)

While superficially this sounds very reasonable, without a precisely stated definition of what is meant by 'more rigorous' curricula or 'performs less well' or 'more stringent grading standards' there is the suspicion that a lurking circularity could cloud the interpretation of the findings. Nonetheless, it is clear from reading their text as a whole that for Korobko *et al*.: i) subject difficulty is whatever it is that is represented by the difficulty parameter in the IRT model, and ii) once scores (grades) on subjects not taken have been 'imputed' to examinees based on the parameters of the (best fitting) IRT model, the estimated average scores (grades) in each subject can legitimately be compared. To paraphrase, this is equivalent to saying that the rank ordering of examination subjects in terms of difficulty is the rank order by average grade in a hypothetical scenario where all examinees take all subjects. The IRT model is used to simulate this hypothetical scenario.

Coe (2007; 2008; 2010) has given far more consideration to the conceptual issues behind the use of IRT models for comparing examination subjects. Using the concept 'construct comparability' he argues that examinations can be compared in terms of the amount of some common construct implied by given grades. For example, when fitting a 1-parameter IRT (Rasch) model to GCSE subjects, the common construct is 'general academic ability'. If subjects are to be compared (for example on the basis of their difficulty parameters from an IRT model) then this comparison must be stated in terms of the common construct:

> *So rather than saying that maths is 'harder' than English we must say that a particular grade in maths indicates a higher level of general academic ability than would the same grade in English*.
> (Coe, 2008, p.613)

This approach allows Coe to make interpretations of observed statistical differences in subject grading outcomes without having to commit either to a particular definition of difficulty or of general academic ability, since both emerge from the IRT analysis. It also implicitly assumes that 'common construct' is synonymous with 'latent trait'.

Defining difficulty for items in a test

The previous section considered how difficulty has been defined (or its definition has been circumvented) by those employing an IRT approach to investigate difficulty of examination subjects. In this section the issue is approached from the other end – that is, by considering how difficulty has been defined at the item level.

Before IRT became widely used, the framework now known as 'Classical Test Theory' (CTT) was used to analyse data from educational tests. In many contexts CTT is still the preferred choice because in some respects it is conceptually more straightforward, and it is often simpler mathematically, both of which make it easier to explain to non-specialists.

The familiar index of item difficulty in CTT is the 'facility value', defined as the mean mark (score) on a question divided by the maximum possible mark. If the question is dichotomous, the facility value is also the proportion of examinees who answered correctly. Therefore, on a test consisting entirely of compulsory dichotomous items, if question 4 (say) has a higher facility value than question 7, this means that question 4 was answered correctly by more people than question 7. It seems completely uncontroversial to say in these circumstances that question 7 was more difficult than question 4. Because we are dealing with CTT, there is, or seems to be, no need to invoke a latent trait or construct. The qualifier 'for these examinees' might be added, but only in a context where it makes sense to consider the performance of other examinees who did not happen to take the test.

But there are complications possible even for this apparently simple case. First, what can be said if the difference does not hold for identifiable sub-groups? For example, suppose that more males answered question 7 correctly than question 4, but that the opposite was the case for females. In this instance it seems natural just to add the qualifier 'for females, but not for males' to the statement 'question 7 was more difficult than question 4'. A more interesting example is if the group of examinees is split into two groups, 'high scoring' and 'low scoring', on the basis of their overall test score. Now it is again possible for the order of difficulty of the two questions to be different in the two groups, but now adding the qualifier 'for high scorers on the test overall' *does* raise the question of what the test overall was measuring. This is because if question 4 and question 7 were included in a test with different items (but the same examinees) it is conceivable that their relative difficulty with respect to high and low-scoring groups could change.

A second complication with even this simple case is that it does not consider the individual patterns of performance on the two questions, as illustrated by Table 1 below.

Table 1: Question scores for three questions on an imaginary test taken by ten examinees

Person	*Q1*	*Q2*	*Q3*
1	1	1	1
2	1	1	0
3	1	0	0
4	1	1	1
5	1	0	0
6	1	1	0
7	0	0	1
8	0	0	0
9	0	0	0
10	0	0	1
Facility	0.6	0.4	0.4

According to facility values, Table 1 shows that Q1 is easier than both Q2 and Q3, and that Q2 and Q3 are equally difficult. But there is an interesting contrast between Q2 and Q3 in terms of their relationship with Q1. Every person either scored the same or better on Q1 than they did on Q2, whereas this does not hold for the comparison between Q1 and Q3.

Looking at it another way, if a two-item test were made up of the items Q1 and Q2 then knowledge of the total score on this test would

also be knowledge of which items were answered correctly – a person with a score of 2 out of 2 would have got both right, a person with 1 out of 2 would have got Q1 right and Q2 wrong, and a person with 0 out of 2 would have got both wrong. In contrast, on a 2-item test made up of Q1 and Q3, knowledge of the total score would not permit knowledge of which items were answered correctly.

The kind of relationship between Q1 and Q2 was formalised by Louis Guttman in his work on scalogram analysis (e.g. Guttman, 1944; 1950). In brief, a set of items forms a scale if the item scores[2] are a simple function of the scale scores. Guttman was well aware that achieving a 'perfect scale' was not likely in many practical contexts but found that 90% perfect scales (in terms of the reproducibility of the item scores from the scale score) were usable as efficient approximations of perfect scales. (It should be noted that scalogram analysis does not just apply to dichotomous items).

There are two reasons why Guttman's work on scalogram analysis is of interest from the point of view of the present article. The first is that he considered it to be a method for analysing *qualitative* data. It has become so natural for us to think of the data arising from testing as quantitative that we can sometimes lose sight of the fact that the 'raw data', as it were, usually consists of written answers to written questions. Where do the numbers come in? The mark scheme can be thought of as a coding scheme that assigns numerical values (usually integers) to examinee responses according to a certain rationale. One purpose of scalogram analysis is to discover whether the item level data (set of responses across the items) for each examinee can be represented by a single number. (In most examinations this would be the raw score obtained by adding up the scores on each item). If the questions form a scale in the scalogram sense then the scale (total) scores have a definite interpretation in terms of the item scores.

The second reason is that Guttman's starting point was definitions of the universe of attributes (e.g. items) and the population of objects (e.g. examinees) to be scaled. The universe of attributes is the concept of interest whose scalability is being investigated, conceived as the indefinitely large set of questions that could be asked on that concept. Items belong to the universe based on their content, not on statistical criteria. For example, the set of questions testing topics on a particular maths syllabus might define a universe whose scalability could be investigated. The population of objects could be examinees who have studied the appropriate course and prepared for an examination in it. The question of scalability then becomes a matter of empirical investigation that can be carried out on a particular sample of items and examinees. A scalable set of items is by definition unidimensional.

Guttman's approach, in my view, represents the closest thing to 'starting from first principles' in developing definitions of difficulty and comparability. For dichotomous items, if two items P and Q are from a scalable universe then item P is more difficult than item Q if some people (from a defined population) get item Q right and P wrong, but no-one gets Q wrong and P right. Unfortunately, extending even this simple definition to polytomous items runs into problems, as shown in Tables 2a and 2b.

The data for Q4 and Q5 in Table 2a meet the scale definition in that if a scale score is made (e.g. by summing the two responses) then the item scores are perfectly reproducible from the scale scores. Everyone scores at least as well on Q4 as they do on Q5, so Q4 could be said to be 'easier' than Q5.

Table 2a: Question scores for two questions on an imaginary test taken by ten examinees

Person	*Q4*	*Q5*	*Score*
1	2	2	4
2	2	2	4
3	2	1	3
4	2	1	3
5	1	1	2
6	1	1	2
7	1	0	1
8	1	0	1
9	0	0	0
10	0	0	0
Facility	0.6	0.4	

Table 2b: Question scores for two questions on another imaginary test taken by ten examinees

Person	*Q6*	*Q7*	*Total*
1	2	2	4
2	2	2	4
3	1	2	3
4	1	2	3
5	1	1	2
6	1	1	2
7	1	0	1
8	1	0	1
9	1	0	1
10	0	0	0
Facility	0.55	0.5	

However, in Table 2b, although the item scores are perfectly reproducible from the total score it is not the case that everyone scores at least as well on one item as the other. Perhaps the most that can be said is that it is easier to score 1 or more on Q6 than Q7, but easier to score 2 on Q7 than Q6.

This last example makes clear that even the ordering of two items by facility value is ambiguous for polytomous (multiple-mark) items. With a different assignment of scores to response categories, the order could change. For example, in Table 2b if the responses scored '2' were scored '2.8' then Q7 would have a higher facility value than Q6.

To summarise, Guttman's work on scalogram analysis provides a definition of unidimensionality and a definition of what it means for one item to be more difficult than another (for dichotomous items at least).

The link between Guttman and IRT

Unfortunately, item level data from real educational tests never conforms exactly to Guttman's pattern. But there is a strong connection between one particular IRT model, the Rasch model (Rasch, 1960), and Guttman's scale pattern (Andrich, 1985). The expected (i.e. modelled) scores from the Rasch model meet the ordering requirements of the Guttman pattern in that people with higher ability have higher expected scores on every item than people with lower ability, and people of all abilities are expected to score higher on a dichotomous item with a lower difficulty than on one with a higher difficulty. This is not necessarily true for other IRT models. It is also noteworthy that Rasch introduced the concept of

2 The item scores need not be numerical – they could represent responses of 'yes' or 'no' to attitude questions, for example.

'specific objectivity', the 'specific' part of which emphasised that the model only held within a specified frame of reference describing the persons and items, a parallel to Guttman's stressing the need for definitions of the universe of attributes and the population of objects whose scalability was to be investigated.

In fact, Guttman did recognise the concept of a quasi-scale – one where the item responses are not highly reproducible from the scale score but where the 'errors' occur in a gradient (Guttman, 1950), in a manner that seems to conform very closely to the pattern of misfit expected from a Rasch model. The significance of a quasi-scale is that the scale score can still predict an outside variable as well as any weighted combination of the individual item scores (as is the case with a perfect scale). The counterpart of this in Rasch analysis is that the total score is a sufficient statistic for estimating ability (Andersen, 1977) – this means that when the data fit the model there is no additional information about ability in the pattern of item responses. People who have attempted the same items and received the same total score will get the same ability estimate regardless of any differences in scores on the individual items.

This suggests that when data fit the Rasch model, it is possible to define difficulty (for dichotomous items) in a reasonably straightforward way: one item is more difficult than another if any arbitrarily selected person has a lower probability[3] of success on it than on the other item.

As with facility values, and as with the Guttman scale, there is no way round the inherent ambiguity of the concept of difficulty for polytomous items when analysed with a Rasch model. For example, the Rasch partial credit model (Masters, 1982) estimates difficulty threshold parameters representing the points on the latent trait where adjacent score categories are equally probable. There are different possible ways of using these threshold estimates to come up with a number representing 'overall difficulty'. For example, the average of the threshold estimates represents the point on the trait where the lowest and highest score categories are equally probable. Alternatively, it is possible to find the point on the latent trait where the expected score is equal to 'half marks' on the item. Because these are different definitions of difficulty, it would be possible for the ordering of two items to differ depending on which definition was used.

Of course, there is not necessarily any need to produce a number representing 'overall difficulty' – it may be more informative to make comparisons at each category. This was the approach taken by Coe (2008) in comparing relative difficulty of GCSE subjects by grade category. (See Andrich, de Jong and Sheridan, 1997; and Linacre, 2010, for a discussion of some of the issues involved in interpreting Rasch threshold parameters).

While the followers of Rasch seem keen to cite Guttman with approval, essentially regarding the Rasch model as a probabilistic form of the Guttman scale, it is not clear whether this approval was reciprocated. Guttman seemed to avoid using the concept of a latent trait. He also made the following comment about a conventional (CTT) item analysis:

> *This idea of scale construction is a most comfortable one: it is virtually guaranteed to succeed for the kinds of data concerned. I know of no instance in which all items were rejected. In other words, item analysis does not test any hypothesis of scalability. It assumes that scalability exists, and that its task is merely to cull out inappropriate items.*
> (Guttman, 1971, p.343)

3 The interpretation of probability in this context is beyond the scope of this article. See Holland (1990) for some discussion.

Rasch practitioners might feel that this criticism does not apply to them, because they are very keen to stress the primacy of the model over the data (e.g. Andrich, 1989; Wright, 1999), but without an a priori definition of the trait it is probably true in some cases in practice that misfitting items are culled and the resulting set of items provides the 'best' measure of an ill-defined concept. It could be argued that this is what happens when attempts are made to model subject difficulty with the Rasch model (e.g. Coe, 2008; Coe *et al*., 2008). Without starting from a definition of 'general academic ability' it is not clear what the estimated values of subject difficulty with respect to this variable actually mean.

Spatial representations of subject difficulty

For Guttman, it was clear that the dimensionality of the data was something to be discovered rather than imposed. If the empirical evidence showed that two items did not form part of the same unidimensional scale then 'not comparable' was a valid experimental finding. In the later part of his career he developed some of the methods that have become part of the field known as 'multidimensional scaling' or MDS (see, for example, van Deun and Delbeke, 2000). Very broadly speaking, the aim of this kind of analysis is to represent objects in the lowest dimensional space that preserves certain aspects of empirically discovered relationships between them. These relationships could be (for example) indices of similarity or of monotonicity. The final spatial representation might attempt to preserve actual differences in terms of these indices ('metric MDS'), or just their order ('non-metric MDS'). For Guttman, the purpose of these spatial representations was to test hypotheses (made in advance on non-statistical grounds) about how the objects would group into regions of the multidimensional space (see, for example, Schlesinger and Guttman, 1969).

A new direction for investigations of subject difficulty might be to explore such an approach. Given that two objects can always be represented in a single dimension, and generally n objects can be represented in n-1 dimensions, a very simple 2-dimensional example can be contrived by considering 3 subjects. There are several reasonable choices for an index of similarity. If there was no need or desire to maintain any connection with an IRT approach then the difference in mean grade achieved by examinees common to each pair of subjects could be used. This is the index of difficulty familiar from subject pairs analyses (see Coe, 2007, for a description of this and related methods).

However, to stay close to the spirit of Rasch it seems interesting to explore an index of difference that has a close connection with the Rasch model for dichotomous items. In this model, one way of estimating item difficulties is the paired method (Choppin, 1968) where an estimate of the difference in difficulty between any two items A and B is the logarithm of the ratio of the number of examinees succeeding on A and failing on B to the number failing on A and succeeding on B. In the context of examinations rather than items we could choose to make them dichotomous by defining success as 'grade x or above' and failure as 'below grade x'. In the example below the A grade has been chosen as the grade for x. The data have been invented for the purpose of the illustration.

Table 3a shows that 300 people got an A in Psychology but not in Biology, whereas only 50 people got an A in Biology but not in Psychology. On the index of difficulty we are using, Biology is thus log (300/50) ≈ 1.8 logits 'harder' than Psychology.

Table 3a: Biology and Psychology grade A

		Psychology	
Biology	Below A	Grade A	Total
Below A	900	300	1200
Grade A	50	200	250
Total	950	500	1450

Table 3b: English Literature and Biology grade A

		Biology	
English	Below A	Grade A	Total
Below A	400	20	420
Grade A	100	120	220
Total	500	140	640

From Table 3b we see that Biology is log (100/20) ≈ 1.6 logits 'harder' than English Literature, and from Table 3c we see that English Literature is log (160/100) ≈ 0.5 logits 'harder' than Psychology.

Table 3c: English Literature and Psychology grade A

		Psychology	
English	Below A	Grade A	Total
Below A	1100	160	1260
Grade A	100	200	300
Total	1200	360	1560

Because these three differences satisfy the 'triangle inequality'[4] in that the sum of any two differences is larger than the remaining one, it is possible to represent these results diagrammatically as in Figure 1 below.

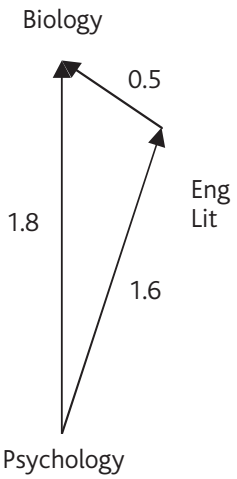


Figure 1: Visual representation of differences in difficulty when triangle inequality is satisfied

4 http://en.wikipedia.org/wiki/Triangle_inequality Accessed 12/04/11.

The length of the arrow represents the logit difference between any two subjects, and the head of the arrow points to the 'more difficult' subject. The closer the three points are to lying on a straight line with arrowheads pointing in the same direction, the more comparable they are as a triplet in terms of difficulty, in the sense that the direct comparison between two subjects is the same as the indirect comparison via a third subject.

Suppose, however, that instead of (1.8, 1.6, 0.5) the three logit differences had been (2.0, 1.5, 0.3). Then the triangle inequality would not have been satisfied and it would not be possible to represent the results as in Figure 1. An alternative depiction of such a scenario is shown in Figure 2.

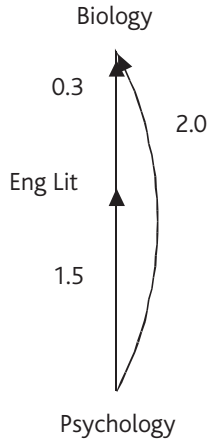


Figure 2: Visual representation of differences in difficulty when triangle inequality is not satisfied

As in Figure 1, the lengths of the arrows represent the logit differences and the heads of the arrows point to the more difficult subjects. (The curved line is part of a circle arc with the straight part as a chord).

With good graphical software it might be possible to represent differences between four subjects (i.e. as a 2D projection of the 'correct' 3D configuration). For higher numbers of dimensions the correct configuration would not be visualisable without either applying some data reduction technique to achieve the best lower dimensional solution according to some criterion, or producing several projections. This is an area for further research.

Conclusion

Using an IRT approach to investigate differences in difficulty among examinations relies on an analogy with using the same approach in its original context – differences in difficulty among items in a test. The software used for the IRT analysis is of course blind to where its inputs have come from and in this sense the outputs of the analysis can be subjected to the usual tests of reliability and model fit.

However, doing this places a greater burden on the analyst to interpret both the latent dimension of the IRT model and the difficulty parameter in that model. This article has shown that it is not entirely straightforward to define difficulty even in the simplest possible case of two dichotomous items in a test. The complications increase as we move to scenarios with polytomous items, scenarios with missing (not presented) items, scenarios with missing (not chosen) items, and finally to scenarios where whole examination subjects are treated as items and there is no a priori defined single trait (dimension) or traits.

This is not to say that an IRT approach is necessarily inadvisable or misleading – the results just need to be interpreted very carefully. It may even be one of the better approaches in cases where there is a pragmatic operational need to produce global rankings of examinees on the basis of overall attainment (as in Tognolini and Andrich, 1996). However, for investigations of differences among subjects, I suggest that it might also be worth going back to the principles first articulated by Guttman, and building up slowly from ground level, considering differences among just a few subjects and representing these visually – searching for stable patterns and always being prepared to accept that 'not comparable' is a reasonable outcome.

References

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika*, **42**, 1, 69–81.

Andrich, D. (1985). An elaboration of Guttman scaling with Rasch models for measurement. In: N. Brandon-Tuma (Ed.), *Sociological Methodology*. 33–80. San Francisco: Jossey-Bass.

Andrich, D. (1989). Distinctions between assumptions and requirements in measurement in the social sciences. In: J. A. Keats, R. Taft, R. A. Heath, & S. H. Lovibond (Eds.), *Mathematical and theoretical systems*. 7–16. New York: North-Holland.

Andrich, D., de Jong, J.H.A.L., & Sheridan, B.E. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. In: J. Rost & R. Langeheine (Eds.), *Application of Latent Trait and Latent Class Models in the Social Sciences*. 59–70. New York: Waxmann Münster.

Choppin, B. (1968). Item bank using sample-free calibration. *Nature*, **219**, 870–872.

Coe, R. (2007). Common examinee methods. In: P.E. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards*. 331–367. London: Qualifications and Curriculum Authority.

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

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

Coe, R., Searle, J., Barmby, P., Jones, K., & Higgins, S. (2008). *Relative difficulty of examinations in different subjects*. Report for SCORE (Science Community Supporting Education). Durham: CEM Centre, Durham University.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, **9**, 139–150.

Guttman, L. (1950). The basis for scalogram analysis. In: S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and Prediction*. 60–90. Princeton, NJ: Princeton University Press.

Guttman, L. (1971). Measurement as structural theory. *Psychometrika*, **36**, 4, 329–247.

Holland, P.W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, **55**, 4, 577–601.

Kane, M.T. (2006). Validation. In: R. L. Brennan (Ed.), *Educational Measurement*. 17–64. Westport, CT: ACE/Praeger series on higher education.

Korobko, O.B., Glas, C.A.W., Bosker, R.J., & Luyten, J.W. (2008). Comparing the difficulty of examination subjects with item response theory. *Journal of Educational Measurement*, **45**, 2, 139–157.

Linacre, J.M. (2010). Transitional categories and usefully disordered thresholds. Online Educational Research Journal, **1**, 3. Retrieved from www.oerj.org 11/01/11.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47**, 2, 149–174.

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

Newton, P.E. (in press). Making sense of decades of debate on inter-subject comparability in England. *Assessment in Education: Principles, Policy & Practice*.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 3, 581–592.

Schlesinger, I.M., & Guttman, L. (1969). Smallest space analysis of intelligence and achievement tests. *Psychological Bulletin*, **71**, 2, 95–100.

Tognolini, J., & Andrich, D. (1996). Analysis of profiles of students applying for entrance to Universities. *Applied Measurement in Education*, **9**, 4, 323–353.

van Deun, K. & Delbeke, L. (2000). Multidimensional Scaling. http://www.mathpsyc.uni-bonn.de/doc/delbeke/delbeke.htm Accessed 12/04/11.

Vidal Rodeiro, C. L. (2007). *A level subject choice in England: patterns of uptake and factors affecting subject preferences*. Cambridge Assessment report. http://www.cambridgeassessment.org.uk/ca/digitalAssets/114189_Survey_Report_-_Final.pdf Accessed 12/04/11.

Wright, B.D. (1999). Fundamental measurement for psychology. In: S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: what every psychologist and educator should know*. 65–104. Mahwah, NJ: Lawrence Erlbaum Associates.

Yen, W.M., & Fitzpatrick, A.R. (2006). Item Response Theory. In: R. L. Brennan (Ed.), *Educational Measurement*. 111–153. Westport, CT: ACE/Praeger series on higher education.

Comparing different types of qualifications: An alternative comparator

Jackie Greatorex Head of the Core Research Team, Assessment Research & Development

Introduction

Returns to qualifications measure how much more is earned on average by people with a particular qualification compared to people with similar demographic characteristics who do not have the qualification. Awarding bodies and the national regulator do not generally use this research method in comparability studies. This article considers what returns to qualifications comparability research can offer awarding bodies and shows that it enables researchers to make comparisons which cannot be satisfactorily achieved by other methods, for instance, comparisons between different types of qualifications, occupations, sectors and progression routes. However, as with all research approaches, returns to qualifications has its limitations.

Background

The English qualification system is complex and for some time government reports have noted this complexity (Foster, 2005; Leitch, 2006; Wolf, 2011). There are thousands of qualifications of several types and at different levels; for details see Isaacs (2010) and Ofqual (2010a to c, 2011a). A glossary of qualifications and qualification types is given in Appendix 1, and a glossary of technical terms, common abbreviations and acronyms relevant to this article is given in Appendix 2.

Different types of cognate qualifications can lead to the same job or programme of study. The results of comparability studies contrasting such qualifications can highlight easy or difficult routes to jobs or further study, and the results may be provided to appropriate authorities who determine what action is necessary to reduce any disparity.

Research methods for comparing the quality of examinees' performance are frequently considered in the literature (e.g. Newton *et al*., 2007) and used in comparability studies. Many of these methods are unsuitable when comparing qualifications that are not predominantly assessed by national examinations. For these comparisons an alternative comparator is required. An example of an alternative comparator, and the focus of this article, is *returns to qualifications*.

Customary comparators

The comparators listed below have often been used in comparability research:

- The demand of the examination items (e.g. QCA, 2006; Crisp and Novaković, 2009a and b)
- The quality of learners' performance as illustrated by their responses to examination items (D'Arcy, 1997; Bramley, 2005; Yim and Shaw, 2009)
- Prior measures of attainment (Bell and Dexter, 2000; Schagen and Hutchinson, 2007)
- Concurrent measures of attainment (Bell and Dexter, 2000; Murphy, 2007).

Each of the customary comparators has different requirements. A robust sample of examination items is needed if their demand is being contrasted. Similarly, a robust sample of learners' work is needed to compare their examination performance. Prior and concurrent measures of attainment both require large datasets with multiple measures of educational attainment.

The studies listed above compare the same type of qualification. Few studies comparing different types of qualifications utilise the customary comparators. Exceptions are that Guthrie (2003) and Arlett (2002; 2003) compared the quality of learners' performance in GCE versus VCE qualifications and Bell and Vidal Rodeiro (2006) used prior and concurrent measures in attainment to compare GCSE versus VGCSE performance in similar subjects.

There are some circumstances in which these customary comparators cannot be used. For example, studies based on comparing the quality of work produced require physical examples of that work, which might not be available. This can happen when performance is assessed by observing work-based practice, or after examination scripts have been destroyed. Appendix 3 displays the requirements for some specific comparability studies and lists some of the circumstances when these cannot be met.

When one or more qualification(s) in a comparison do not fit the requirements for the customary comparators an alternative comparator is needed. The focus of this article is an overview of an alternative comparator: the *returns to qualifications*. The article describes this comparator and analyses its strengths and weaknesses from a comparability research perspective.

Returns to qualifications

There is an established literature about returns to qualifications in economics (e.g. Psacharopoulos, 1947, 1981; Morgan and David, 1963; Ziderman, 1973; Dearden *et al*., 2000; Conlon and Patrignani, 2010). This field of research is influential and has featured in government reviews such as Leitch (2006) and Wolf (2011).

Returns to qualifications are a relative statistical measure which show how much more on average is earned by people with a given qualification in contrast to people with similar demographic characteristics who do not have the qualification (Wolf, 2011).

A recent example of returns to qualifications research in the field of education can be found in Dolton *et al*. (2001) who applied several statistical models and contrasted returns to qualifications for men and women. They found statistically significant returns for non-government funded apprenticeships and degrees for men, and for degrees and NVQ level 2 or more for women. To date there is little awarding body research in this area. Arguably the exception is the study by Conlon and Patrignani (2010). They found that people with level 2 vocational qualifications showed a relatively strong return when compared to people with no

qualifications. Other examples of research and their findings are given in Table 1.

Table 1: Examples of returns to qualifications research and selected quotes

Returns to…	*Example of comparison*	*Selected quotes*
Types of qualification	Vocational vs. academic	"Considerable variation was however uncovered in the wage returns to different types of qualification, with academic qualifications generally earning higher rewards." (Sianesi, 2003, pp.1–2)
Level of qualification	Level 2 vs. level 3	"In aggregate, the returns to qualifications are quite similar for full-time men and women. The rate of return to level 1 qualifications is negligible or zero; while at level 2 and above, the returns are positive and significant, and quite substantial – around 13%–16% for both level 2 and level 3 qualifications, and rising to 23%–31% for level 4 and level 5 qualifications." (Dickerson and Vignoles, 2007, p.V)
Awarding bodies	EdExcel vs. City and Guilds vs. RSA	"[R]eturns associated with level 2 vocational qualifications are relatively strong compared to those in possession of no formally recognised qualifications, with individuals in possession of RSA Level 2, City & Guilds Level 2 and BTEC Level 2 qualifications achieving 38.4%, 15.6% and 13.1% earnings premiums." (Conlon and Patrignani, 2010, p.6)
Occupations	Sales vs. Machine operators	"[W]e find that in particular occupations (such as skilled manual occupations and personal services) and particular industries (such as public administration, education and health), the estimated returns to NVQ2 qualifications are positive and statistically significant." (McIntosh and Garrett, 2009, p.79)
Sectors	Automotive skills vs. Financial Services	"Only the Energy & Utility Skills and People 1st sectors show a positive significant return to level 2 vocational qualifications for males, for example. For women, the return to level 2 vocational qualifications is positively significant in just one SSC, Automotive Skills (albeit with the relatively low sample size of 47)." (Dickerson and Vignoles, 2007, p.15)
Qualifications in different years	Year vs. year	"The rate of return to all levels of education for men remained fairly stable or slightly increased over time while the returns to all educational qualifications noticeably declined for women." (Silles, 2007, pp.411–412)
Progression route	Vocational vs. academic	"For men on the vocational route the extra pay which results from progressing to a higher qualification is less impressive. Having an HND/HNC rather than an OND/ONC yields only an extra 11 percentage points, compared with the 16 percentage point gain in earnings when a man with 2 A levels attains a first degree." (Robinson, 1997, p.12)

Note that these quotes are only a small selection of the findings reported in the returns to qualifications literature and do not indicate overall patterns of findings.

How is the alternative comparator interpreted?

There are several ways of interpreting returns to qualifications. The research often construes them as a proxy for people's productivity (Stasz, 2001; Sianesi, 2003). Within this broad agreement there are two main contrasting hypotheses: 'signalling' and 'human capital'. The signalling hypothesis proposes that returns to qualifications indicate the learners' skills and motivation levels or productivity from *before* entering the qualification's learning programme, and that the qualification does not necessarily improve productivity. The human capital hypothesis proposes that education leading to qualifications improves learners' productivity, which leads to higher earnings and thereby higher returns to qualifications. The weight of evidence supports the human capital hypothesis (Machin and Vignoles, 2005). For further detail on the debate see Sianesi (2003) or Powdthavee and Vignoles (2006). If the human capital hypothesis is correct it makes returns to qualifications a more useful comparator for awarding body purposes, as they would then indicate the value of the learning associated with the qualification.

When qualifications give similar returns they are comparable in terms of economic value and the productivity of the qualification holders is comparable. This is not to say that the knowledge, skills, competence and personality attributes of people with different qualifications are the same or similar. Therefore the results of returns to qualifications analyses will not necessarily align with outcomes from other comparability research using the customary comparators.

Methods for researching returns to qualifications

Usually the research involves analysing longitudinal survey data. The most suitable UK longitudinal datasets are the Youth Cohort Study, the National Child Development Study, and the 1970 British Cohort Study (Wolf, 2011). These datasets are detailed and comprehensive. For instance, the 1970 British Cohort Study achieved a cross-sectional sample of 16,571 in 1970 and 11,261 in 1999–2000 (Centre for Longitudinal Studies, 2009). The Universities of Essex and Manchester (2008) provide information about the 1970 British Cohort Study including a follow-up 38 years later in 2008–2009. The follow-up dataset contains variables representing type of residence, sex, income from earnings, socio-economic group, managerial duties and health. Additionally, it contains variables about qualifications gained, dates they were achieved, whether study was full or part time and who paid for them. The survey covers A levels, GCSEs, O levels, NVQs, degrees, City and Guilds, RSA, HNC, HND and other qualifications.

Comparing the average wage of people with a qualification with the average wage of similar people without the qualification (Wolf, 2011) is a staple of investigation. This is achieved by creating samples from within the data, one with and one without the qualification, which have similar profiles on other variables.

A family of statistics known as regression or regression modelling is utilised to calculate returns to qualifications. An attribute of regression is that analysts can control for the effects of variables (e.g. García-Mainar and Montuenga-Gómez, 2005). The effect of the controlled variable is removed to avoid influencing the effect of the variables under investigation. For example, Robinson (1997) controlled for *years of experience* to contrast returns to qualifications thereby avoiding a comparison of the effects of both *years of experience* and *qualifications* on wages. This is important because unqualified people are often older. Dearden *et al*. (2002) used the 1991 sweep of the National Child Development Study and the 1998 Labour Force Survey and found that the returns to vocational qualifications were more similar to those of academic qualifications when they controlled for time taken to gain the qualification. Length of time to gain a qualification is important to control as vocational qualifications often take a shorter time to gain than academic qualifications. Dearden *et al*. (2002) also investigated the bias that can occur when regression models do not control for variables like *ability* and *measurement error*. They found that returns to qualifications analyses that did not control for ability tended to be biased upwards, and those that did not control for *measurement error* tended to be biased downwards. These biases might cancel one another out. They analysed the National Child Development Study data controlling for *ability* and

measurement error and the Labour Force survey data without controlling for either. The results were similar suggesting the biases do indeed offset each other. In summary, returns to qualifications analyses which do not control for variables such as *ability* and *measurement error* can sometimes give reasonable estimates of returns to qualifications.

Some studies aggregate qualifications together to calculate returns to groups of qualifications. Different aggregations of qualifications enable comparisons between:

- Types of qualification
- Levels of qualification
- Awarding bodies
- Occupations
- Sectors
- Different years
- Progression routes

Examples of results from research comparing the above were given in Table 1 previously.

Strengths of returns to qualifications as a comparator

Returns to qualifications are a valid comparator even when comparing qualifications which are not cognate, as in McIntosh and Garrett (2009), and Dickerson and Vignoles (2007). In contrast the customary comparators such as quality of learners' performance and the demand of examination items tend to be used to compare cognate qualifications. Examples include D'Arcy (1997) and Yim and Shaw (2009).

An advantage of returns to qualifications is that they are often more independent of the awarding body and qualification system than the customary comparators. The reasons for the customary comparators being embedded in the awarding body and qualifications system are as follows:

- Judgements about the demand of items and the quality of learners' performance are often made by senior assessors, moderators and verifiers from the qualifications. Bramley (2007) considered the design of fourteen inter-board comparability studies and reported that in ten of the fourteen studies no independent judges participated and in the four studies using independent judges less than a third of the judges were independent, although Forster and Gray (2000) found no evidence that the judgements made by independent judges were different from those made by board-affiliated judges.
- Measures of prior and concurrent attainment are often derived from qualifications or examinations offered by the main awarding bodies or the regulator. Examples include Bell and Dexter (2000), Elliott *et al*. (2002) and Bell (2000).
- The customary comparators can be an accumulation of awarding body/qualifications system decisions (i.e. the decisions by senior assessors, moderators and verifiers along with awarding body staff).

Returns to qualifications, however, are mostly the outcome of decisions by employers. In summary, many employers' decisions contributing to the measure of returns to qualifications renders it more independent than the customary comparators.

There are strategies for increasing the independence of the customary comparators. All experts judging item demand and the quality of learners' performance can be recruited using the criteria that they are experts in the field but independent of the awarding bodies and qualification system under investigation. Some studies claim to use only independent judges but the criteria for recruitment are not explicit, so the exact meaning of 'independent' is unclear – see, for example, Ofqual (2010d). The measures of prior or concurrent attainment can be chosen from outside the awarding body or qualifications system, such as a reference test developed by an independent group. Murphy (2007) provides a comprehensive discussion of reference tests.

Weaknesses of returns to qualifications as a comparator

There are multiple opinions regarding what returns to qualifications measure – for example, the signalling versus human capital hypothesis. Some interpretations are better than others for awarding body purposes, as discussed above.

Strengths of research methods associated with returns to qualifications

Exploiting large longitudinal datasets makes returns to qualifications a robust and powerful research technique. The longitudinal data is preferable to self-reported data which relies on people remembering information, such as examination results (Wolf, 2011). This is a strength of the approach as the fallibility of self-reported examination results is well known. For instance, Kuncel *et al*. (2005) considered several studies and found that the validity of self-reported test scores, grade point averages and class ranks were moderated by school performance and cognitive ability; they suggest using such data with caution. A further strength is that analysts can control for confounding variables, which facilitates purer measures of returns to qualifications (see earlier). A final strength is that the qualifications can be aggregated to make several different types of comparison, for example, comparisons between different levels, occupations, sectors and progression routes, as shown in Table 1. These are infrequently researched by awarding bodies and therefore the returns to qualifications research is offering new comparability evidence.

Weaknesses of research methods associated with returns to qualifications

There are several weaknesses in the returns to qualifications research. Qualifications and other variables do not necessarily 'cause' returns to qualifications (Sianesi, 2003). Sianesi is concerned about how people might apply the repeated research finding that NVQ level 1 and level 2 qualifications are associated with negative returns (e.g. Jenkins *et al.*, 2007 in Wolf, 2011). If NVQs are believed to cause the negative returns, then the qualifications are arguably valueless. This is not necessarily the case, as people with NVQ level 1 and 2 have a higher probability of subsequent employment than those in matched groups without the qualifications (e.g. Dearden *et al*., 2000, and Jenkins *et al*., 2002, in Sianesi, 2003).

Another weakness is that some variables can be influenced by unobserved variables. McIntosh and Garrett (2009) describe steps that can be taken to try to reduce the likelihood of this happening.

Inferences about skills, knowledge, motivation and productivity can be somewhat oversimplified by returns to qualifications analyses. Personality characteristics, competence, skills and knowledge are often treated as unidimensional; that is, they are combined into one measure of returns to qualifications (Stasz, 2001). This is not necessarily realistic as there is evidence that some academic performance is

multidimensional (e.g. Jackson, 1985) and research indicates that personality attributes are multidimensional (e.g. Gow *et al*., 2005). On the other hand educational attainment and personality attributes are connected (Chamorro-Premuzic and Furnham, 2003; Richardson and Abraham, 2009) and personality theorists have research evidence for a personality disposition which integrates most general non cognitive dimensions of personality (e.g. Musek, 2007). Therefore, one scale for returns to qualifications might not be a pure measure, but given how knowledge, skills and personality can be linked, returns to qualifications is likely to be a reasonable proxy for productivity. A related point is that returns to qualifications research assumes that the learning from a qualification transfers to the workplace (Stasz, 2001). However, research shows that knowledge, skills and so on from one context do not readily transfer to another (Lave, 1988; Carraher, 1991).

The statistical results are somewhat dependent on how the statistical model is specified (Wolf, 2011). It is possible to have two valid statistical models using the same data which produce different results (Wolf, 2011); therefore reoccurring patterns of results are more trustworthy than findings from one statistical model (McIntosh and Garrett, 2009).

Conclusion

Returns to qualifications are a statistical measure contrasting the average earnings (often interpreted as productivity) of people who have a particular qualification(s) with the average earnings of those without the qualification. Thus far, returns to qualifications are relatively unexplored by awarding bodies, although they are prominent in government reviews of vocational qualifications. This comparator enables researchers to make comparisons which cannot be achieved by other methods and has the advantage that it is more independent than customary comparators used in many comparability studies. The alternative comparator and associated methods have strengths and weaknesses but provide some robust comparability evidence. The strongest comparability evidence is when there is a clear pattern in the results of several studies using different established research methods and independent data sets. Therefore results from returns to qualifications research combined with results from the customary comparators would provide a strong research evidence base.

Acknowledgements

I would like to thank John Bell for introducing me to the returns to qualifications research.

References

Arlett, S. (2002). *A comparability study in VCE Health and Social Care units 1, 2 and 5. A review of the examination requirements and a report on the cross moderation exercise. A study based on the summer 2002 examinations*. Organised by the Assessment and Qualification Alliance on behalf of the Joint Council for General Qualifications.

Arlett, S. (2003). *A comparability study in VCE Health and Social Care units 3, 4 and 6. A review of the examination requirements and a report on the cross moderation exercise. A study based on the summer 2002 examinations*. Organised by the Assessment and Qualification Alliance on behalf of the Joint Council for General Qualifications.

Bates, I. (1990). "No Roughs and No Really Brainy Ones": The Interaction Between Family Background, Gender and Vocational Training on a BTEC Fashion Design Course. *Journal of Education and Work*, **4**, 1, 79–90.

Bell, J. F. (2000). *Methods of aggregating GCSE results to predict A-level performance*. Paper presented at the British Educational Research Association Annual Conference, Cardiff University, September 7–10. Retrieved 27 May 2011 from http://www.leeds.ac.uk/educol/documents/00001506.htm

Bell, J. & Dexter, T. (2000). *Using Multilevel models to assess the comparability of Examinations*. Paper presented at Fifth International Conference on Social Science Methodology, Cologne, October 3–6, Retrieved 27 May 2011 from http://www.leeds.ac.uk/educol/documents/00001528.htm

Bell, J. F. & Vidal Rodeiro, C. L. (2006). *Performance in GCSE examinations in vocational subjects (GCSEvs) 2004–2005*. A report from the Research Division of Cambridge Assessment commissioned by the Qualifications and Curriculum Authority. Retrieved 27 May 2011 from http://www.ofqual.gov.uk/files/GCSEvs_annex2_ca_appendices_mar07.pdf

Bramley, T. (2005). A rank ordering method for equating tests by expert judgement. *Journal of Applied Measurement*, **6**, 2, 202–223.

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

Carraher, D. (1991). Mathematics in and out of school: a selective review of studies from Brazil. In: M. Harris (Ed.), *Schools, Mathematics, and Work*, 169–201. London: The Falmer Press. Cited in Stasz, C. (2001). Assessing skills for work: two perspectives. *Oxford Economics Papers*, **3**, 385–405.

Centre for Longitudinal Studies (2009). British Cohort Study. Retrieved 27 May 2011 from http://www.cls.ioe.ac.uk/studies.asp?section=000100020002

Chamorro-Premuzic, T. & Furnham, A. (2003). Personality predicts academic performance: Evidence from two longitudinal university samples. *Journal of Research in Personality*, **37**, 4, 319–338.

Coles, M. & Matthews. A. (1995). *Fitness for purpose. A means of comparing qualifications*. London: A report to Sir Ron Dearing.

Conlon, G. & Patrignani, P. (2010). *Returns to BTEC vocational qualifications*. Final Report for Pearson. Retrieved 27 May 2011 from http://www.edexcel.com/Policies/Documents/Final%20Report%20Returns%20to%20BTEC%20Vocational%20Qualifications%20Fin%E2%80%A6.pdf

Crisp, V. & Novaković, N. (2009a). Are all assessments equal? The comparability of demands of college-based assessments in a vocationally related qualification. *Research in Post-Compulsory Education*, **14**, 1, 1–18.

Crisp, V. & Novaković, N. (2009b). Is this year's exam as demanding as last year's? Using a pilot method to evaluate the consistency of examination demands over time. *Evaluation and Research in Education*, **22**, 1, 3–15.

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

Dearden, L., McIntosh, S., Myck, M. & Vignoles, A. (2000). *The returns to academic, vocational and basic skills in Britain. DfEE Research Report No. 250 and Research Brief No. 250 and Centre for Economics of Education Discussion Papers Nos. 3 and 4*. Cited in Sianesi, B. (2003). *Returns to Education: A Non-Technical Summary of CEE Work and Policy Discussion*. Institute for Fiscal Studies and Centre for the Economics of Education. Retrieved 27 May 2011 from http://www.ifs.org.uk/docs/cee_summ.pdf

Dearden, L., McIntosh, S., Myck, M. & Vignoles, A. (2002). The returns to academic and vocational qualifications. *Bulletin of Economic Research*, **54**, 3, 0307–3378.

Dickerson, A. & Vignoles, A. (2007). *The distribution and returns to qualifications in the sector skills councils. Research Report 21*. Skills for Business. Retrieved 27 May 2011 from http://www.ukces.org.uk/the-distribution-and-returns-to-qualifications-in-the-four-countries-ofthe-uk-research-report-21a/*/changeNav/002003/outputFormat/print

Directgov (2011a). Education and learning. BTECs, City and Guilds and OCR Nationals. Retrieved 27 May 2011 from http://www.direct.gov.uk/en/EducationAndLearning/QualificationsExplained/DG_10039020

Directgov (2011b). Education and Learning. Higher National Certificates and Higher National Diplomas. Retrieved on 27 May 2011 from http://www.direct.gov.uk/en/EducationAndLearning/QualificationsExplained/DG_10039026

Dolton, P.L., Makepeace, G. H. & Gannon, B.M. (2001). The earnings and employment effects of young people's vocational training in Britain. *The Manchester School*, **69**, 4, 387–417.

Elliott, G., Forster, M., Greatorex, J. & Bell, J. F. (2002). Back to the future: A methodology for comparing A-level and new AS standards. *Educational Studies*, **28**, 2, 163–180.

Forster, M. & Gray, E. (2000, September). *Impact of independent judges in comparability studies conducted by awarding bodies*. Paper presented at the British Educational Association annual conference, University of Cardiff.

Foster, A. (2005). *Realising the Potential. A review of the future role of further education colleges*. Annesley: Department for Education and Skills. Retrieved 27 May 2011 from http://image.guardian.co.uk/sys-files/Education/documents/2005/11/15/fosterreport.pdf

García-Mainar, I. & Montuenga-Gómez. V. M. (2005). Education returns of wage earners and self-employed workers: Portugal vs. Spain. *Economics of Education Review*, **24**, 161–170.

Gow, A. J. Whiteman, M. C., Pattie, A. & Deary, I. J. (2005). Goldberg's 'IPIP' Big-Five factor markers: Internal consistency and concurrent validation in Scotland. *Personality and Individual Differences*, **39**, 2, 317–329.

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

The Information Authority (undated). The Information Authority/ILR/ILR. Retrieved 27 May 2011 from http://www.theia.org.uk/ilr/

Isaacs, T. (2010). Educational Assessment in England. *Assessment in Education: Principles, policy and practice*, **17**, 3, 315–334.

Jackson, I., (1985). On detecting aptitude effects in undergraduate academic achievement scores. *Assessment and Evaluation in Higher Education*, **10**, 1, 71–88. Cited in Harvey, L., Drew, S. and Smith, M. (2006). *The first-year experience: a review of literature for the Higher Education Academy*. Centre for research and evaluation, Sheffield Hallam University. Retrieved 27 May 2011 from http://www.heacademy.ac.uk/assets/York/documents/ourwork/archive/first_year_experience_full_report.pdf

Jenkins, A., Greenwood, C & Vignoles, A. (2007). *The Returns to Qualifications in England: Updating the Evidence Base on Level 2 and Level 3 Vocational Qualifications*. Centre for the Economics of Education. Retrieved 27 May 2011 from http://eprints.lse.ac.uk/19378/1/The_Returns_to_Qualifications_in_England_Updating_the_Evidence_Base_on_Level_2_and_Level_3_Vocational_Qualifications.pdf

Jenkins, A., Vignoles, A., Wolf, A. & Galindo-Rueda, F. (2002). *The Determinants and Labour Market Effects of Lifelong Learning Discussion Paper No 19*. Centre for Economics of Education. Cited in Sianesi, B. (2003). *Returns to Education: A Non-Technical Summary of CEE Work and Policy Discussion*. Institute for Fiscal Studies and Centre for the Economics of Education. Retrieved 27 May 2011 from http://www.ifs.org.uk/docs/cee_summ.pdf

Kuncel, N. R., Credé, M. & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, **75**, 1, 63–82.

Lave, J. (1988). *Cognition in Practice: Mind, Mathematics and Culture in Everyday Life*. Cambridge, MA: Cambridge University Press.

Leitch, S. (2006). *Leitch Review of Skills. Prosperity for all in the global economy – world class skills*. Final report. London: The Stationery Office. Retrieved 27 May 2011 from http://www.delni.gov.uk/leitch_finalreport051206[1]-2.pdf

Machin, S. & Vignoles, A. (2005). *What's the Good of Education?: The Economics of Education in the UK*, Princeton and Oxford: Princeton University Press. Cited in Powdthavee, N. & Vignoles, A. (2006). *Using rate of return analyses to understand sector skill needs, CEE DP70*. Centre for the Economics of Education. Retrieved 27 May 2011 from http://cee.lse.ac.uk/cee%20dps/ceedp70.pdf

McIntosh, S. & Garrett, R. (2009). *The Economic Value of Intermediate Vocational Education and Qualifications. Evidence Report 11*. UK Commission for Employment and Skills. Retrieved 27 May 2011 from http://www.ukces.org.uk/upload/pdf/UKCES%20Full%20Report%2011.pdf

Morgan, J. & David, M. (1963). Education and income. *The Quarterly Journal of Economics*, **77**, 3, 423–437.

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

Musek, J. (2007). A general factor of personality: Evidence for the Big One in the five-factor model. *Journal of Research in Personality*, **41**, 6, 1213–1233.

Newton, P., Baird, J., Goldstein, H., Patrick, H. and Tymms. P. (Eds) (2007). *Techniques for monitoring comparability of examination standards*. 377–441. London: QCA. (CD version).

Ofqual (2010a). The Register. Retrieved 27 May 2011 from http://register.ofqual.gov.uk/

Ofqual (2010b). Explaining the Qualifications and Credit Framework. Retrieved 27 May 2011 from http://www.ofqual.gov.uk/qualification-and-assessment-framework/89-articles/145-explaining-the-qualifications-and-credit-framework

Ofqual (2010c). Explaining the National Qualifications Framework. Retrieved 27 May 2011 from http://www.ofqual.gov.uk/qualification-and-assessment-framework/89-articles/250-explaining-the-national-qualifications-framework

Ofqual (2010d) Comparability of ICT Level 2 Qualifications. An analysis of internal units. Ofqual/10/4796. Retrieved on 10th June from http://www.ofqual.gov.uk/files/10-12-01-ICT-comparability-study.pdf

Ofqual (2010e). Glossary. Retrieved 27 May 2011 from http://www.ofqual.gov.uk/help-and-support/94-articles/34-161-glossary

Ofqual (2011a). Qualification and Assessment Framework. Retrieved 27 May 2011 from http://www.ofqual.gov.uk/qualification-and-assessment-framework

Ofqual (2011b). Our regulatory approach. Retrieved 27 May 2011 from http://www.ofqual.gov.uk/for-awarding-organisations/96-articles/609-our-regulatory-approach

Palmer, G. (undated). English National Pupil Database. Retrieved 27 May 2011 from http://poverty.org.uk/technical/npd.shtml

Powdthavee, N. & Vignoles, A. (2006). *Using rate of return analyses to understand sector skill needs, CEE DP70*. Centre for the Economics of Education. Retrieved 27 May 2011 from http://cee.lse.ac.uk/cee%20dps/ceedp70.pdf

Psacharopoulos, G (1947). The economic returns to higher education in twenty five countries. *Higher Education Quarterly*, **1**, 2, 141–158.

Psacharopoulos, G (1981). Returns to education: an updated international comparison. *Comparative Education*, **17**, 3, 321–341.

QCA (2006). Comparability study of assessment practice: Personal Licence holder qualifications QCA/06/2709. London: QCA. Retrieved 27 May 2011 from http://www.ofqual.gov.uk/files/personal_licence_holder_quals_comparability_study.pdf.

Richardson, M. & Abraham, C. (2009). Conscientiousness and achievement motivation predict performance. *European Journal of Personality*, **23**, 7, 589–605.

Robinson, P. (1997). *The myth of parity of esteem: earnings and qualifications*. *Discussion Paper No 354*. Centre for Economic Performance. Retrieved 27 May 2011 from http://cep.lse.ac.uk/pubs/download/dp0354.pdf

Schagen, I. & Hutchison, D. (2007). Multilevel modelling methods In: P. Newton, J. Baird, H. Goldstein, H. Patrick, and P. Tymms. (Eds.) (2007), *Techniques for monitoring comparability of examination standards*. 377–441. London: QCA. (CD version).

Sianesi, B. (2003). *Returns to Education: A Non-Technical Summary of CEE Work and Policy Discussion*, Institute for Fiscal Studies and Centre for the Economics of Education. Retrieved 27 May 2011 from http://www.ifs.org.uk/docs/cee_summ.pdf

Silles, M. A. (2007). The returns to education for the United Kingdom. *Journal of Applied Economics*, **X**, 2, 391–413.

Stasz, C. (2001). Assessing skills for work: two perspectives. *Oxford Economics Papers*, **3**, 385–405.

UK Commission for Employment and Skills (undated). About Sector Skills Councils. Retrieved 27 May 2011 from http://www.ukces.org.uk/sector-skills-councils/about-sscs/

Universities of Essex and Manchester (2008). Economic and Social Data Service, 1970 British Cohort Study: Thirty-Eight-Year Follow-up, 2008–2009. Variables. Retrieved 27 May 2011 from http://www.esds.ac.uk/findingData/variableList.asp?sn=6557&recid=1&class=0&from=sn#gs

University of Hull (2007). CSE. Retrieved 27 May 2011 from http://slb-ltsu.hull.ac.uk/awe/index.php?title=CSE

Wolf, A. (2011). *Review of Vocational Education – The Wolf Report*. Department for Education. Retrieved 27 May 2011 from http://www.education.gov.uk/publications/standard/publicationDetail/Page1/DFE-00031-2011

Yim, L. & Shaw, S. D. (2009). *A comparability study using a rank-ordering methodology at syllabus level between examination boards*. Paper presented at the International Association for Educational Assessment conference, 13–18 September, Brisbane, Australia. Retrieved 27 May 2011 from http://www.iaea2009.com/abstract/82.asp

Ziderman, A. (1973). Does it pay to take a degree. The profitability of private investment in university education in Britain. *Oxford Economics Papers*, **25**, 2, 262–274.

APPENDIX 1:
GLOSSARY OF QUALIFICATIONS AND QUALIFICATION TYPES

A level	General Certificate of Education Advanced Level. Typically taken by 18 year olds after a two year study programme. Currently the assessments are unitised and some interim examinations contributed to final grades. A type of general qualification.
BTEC	Business and Technology Education Council. Sometimes used to refer to a type of vocational or work related qualification. For details see Bates (1990) and Directgov (2011a).
City and Guilds	Awarding body. Sometimes used to refer to a type of vocational or work related qualification. For details see Directgov (2011a).
CSE	Certificate of Secondary Education. Typically taken by 16 year olds after a two year study programme. O levels catered for the higher ability students. Lower ability students took the Certificate of Secondary Education. The qualification was available between 1965 and 1987. For details see University of Hull (2007). A type of general qualification.
GCE	General Certificate of Education (see also O level and A level).
GCSE	General Certificate of Secondary Education. Typically GCSE is taken by 16 year olds as part of a two year course. Sometimes the examinations are all taken at the end of the two years and at other times they are spread throughout the two years. The qualification replaced O levels and CSEs in England in 1988. A type of general qualification.
GNVQ	General National Vocational Qualifications. They were designed to be an alternative to GCSE and GCE, but also to be different in size, content and assessment approach (Coles and Matthews, 1995). These qualifications are no longer available.
HND/HNC	Higher National Diploma/Higher National Certificate. Type of work related or vocational higher education qualification. For details see Directgov (2011b).
NVQ	National Vocational Qualifications. NVQs are available at several levels and are therefore taken by learners of varied age. NVQs are based on national occupational standards. They are competence based qualifications. Wolf (2011) explains that NVQs are now scheduled to disappear with the exception of some qualifications preserved by some of the Sector Skills Councils. A type of vocational qualification.
O level	General Certificate of Education Ordinary Level. The last year of national testing of O levels was 1987. Typically 16 year olds took the examinations after two years of study. O levels catered for the higher ability students. Lower ability students took the Certificate of Secondary Education. A type of general qualification.
OND/ONC	Ordinary National Diploma/ Ordinary National Certificate. A type of vocational qualification, a BTEC qualification.
RSA	RSA Examinations Board or Royal Society of Arts Examinations Board. This awarding body is now part of OCR (Oxford, Cambridge and RSA examinations). Sometimes used to refer to a type of vocational or work related qualification.
VCE	Vocational Certificate of Education. VCEs replaced GNVQs at level 3. These are no longer available. A type of vocational qualification.
VGCSE	Vocational General Certificate of Secondary Education. GCSEs in vocational subjects were introduced in 2000. However, the term 'vocational' was dropped in 2004. A type of vocational qualification.

APPENDIX 2:
GLOSSARY OF ASSESSMENT TERMS USED IN THIS ARTICLE

Accredited qualification	A qualification and specification are accredited by Ofqual when they meet regulatory criteria. For details see Ofqual (2010e, 2011b).
Accreditation of prior (experiential) learning	The recognition (award) of academic credit for demonstrated learning and achievement from formal education, life or work. The process is used by learners to gain entry to a learning programme or to claim credit for part of a qualification.
City and Guilds	Awarding body.
Cognate	The same subject/discipline/occupation.
Comparability	Extent of the similarity or equivalence of qualification(s) or unit(s).
Comparator	A device for comparing qualification(s)/unit(s) and determining their comparability. It might be a numerical measure like returns to qualifications or a concept like factual recall.
Controlled assessment	Assessments taken under supervised conditions. They are set by the awarding body and assessed by the learner's teacher or set by the learner's teacher and assessed by an assessor contracted by the awarding body. Many UK qualifications now have controlled assessment rather than coursework.
Coursework	Assessments, often project work, which were devised by the learner/teacher/awarding body within awarding body guidelines. Generally assessed by the learner's teacher.
Demand	The level of knowledge, skills and competence required of typical learners.
External moderator	A subject/occupational expert contracted by the awarding body to check the assessment judgements of assessors employed by schools, colleges and employers.
Internal moderator	A subject/occupational expert employed by schools, colleges and employers to check the assessment judgements of assessors from the same organisation.
External verifier	A subject/occupational expert contracted by the awarding body to check the assessment judgements of assessors employed by schools, colleges and employers. The external verifiers also consider audit trails.
Internal verifier	A subject/occupational expert employed by schools, colleges and employers to check the audit trail and assessment judgements of assessors from the same organisation.
NQF	National Qualifications Framework. For details see Ofqual (2011).
Ofqual	National regulator of qualifications in England and vocational qualifications in Northern Ireland.
Productivity	The skills, knowledge, competence and personality attributes a person uses in a job to produce goods and services of economic value.
QCA	Qualifications and Curriculum Authority. QCA was the predecessor of Ofqual.
QCF	Qualifications and Credit Framework. For details see Ofqual (2011a).
Qualification level	Qualification levels are within qualification frameworks (e.g. NQF, QCF). Each level contains qualifications deemed to be of similar demand. The qualifications in a level vary in subject, content and assessment design.
Returns to qualifications	A statistical proxy of the productivity of people who have a particular qualification(s) compared with the productivity of those without the qualification.
RSA	RSA Examinations Board or Royal Society of Arts Examinations Board. This awarding body is now part of OCR (Oxford, Cambridge and RSA examinations).
SSC	Sector Skills Council. SSCs are employer driven, UK wide organisations that aim to ensure the UK has the skills needed for the present and the future, and to improve productivity and performance. Each SSC covers a particular industry. For details see UK Commission for Employment and Skills (undated).
Type of qualification	Qualifications with a particular characteristic, or from a particular grouping e.g. A levels, vocational qualifications, BTEC.

APPENDIX 3:
REQUIREMENTS FOR COMPARABILITY STUDIES AND EXAMPLES OF CIRCUMSTANCES WHEN THESE ARE NOT AVAILABLE

Comparator	*What is needed*	*Circumstances when these are not available*
Demand of the examination items	The examination items (or equivalent) answered by many learners. A representative sample might suffice.	• When performance is assessed by observing work practice in situ and asking supplementary questions as needed. Examples include many NVQ assessments. • Internally assessed units when the assessment task is devised or adapted by the learner/teacher and the learner's performance is assessed by the teacher. These include some coursework/controlled assessment tasks in GCSE, Diplomas and other qualifications. An example of a study when researchers attempted to collect internally assessed tasks for a vocational qualification in administration, with limited success, is Crisp and Novaković (2009a). • Cases of the accreditation of prior (experiential) learning.
Quality of learners' performance	A representative sample of learners' responses to the examination items.	• Once scripts are destroyed. Scripts from examinations are destroyed after a certain length of time once certificates have been issued. The exception is a small number of scripts on some grade boundaries. • For some internally assessed units. The awarding body has limited access to the artefacts produced by learners. The artefacts are often retained by schools, colleges or learners. Crisp and Novaković (2009a) collected artefacts for a research study with limited success. • The assessment does not require the learners to produce an artefact or a recording of the learners' performance such as a video of a drama performance. Examples include some NVQ assessment as mentioned above.
Prior measures of attainment	Marks or grades from both prior measures of attainment and the current mark or grade for each learner or for a representative sample of learners.	When the qualifications/learners under investigation are not well represented in databases with multiple measures of educational attainment. There are several government owned databases which have prior, current and concurrent measures of attainment, examples include the National Pupil Database (NPD) and Individualised Learner Record (ILR). For details of the NPD see Palmer (undated) and for ILR see The Information Authority (undated). However, less well represented learners are: • Not from state maintained schools (independent schools are not required to follow the National Curriculum and to take the statutory national tests) • Not of typical test taking age • Too old to have key stage test results • Taking certain qualifications, usually vocational qualifications • Taking unaccredited qualifications
Concurrent measures of attainment	Marks or grades from both prior measures of attainment and the current mark or grade for each learner or for a representative sample of learners.	

Linking assessments to international frameworks of language proficiency: the Common European Framework of Reference

Neil Jones Assistant Director, Research & Validation, Cambridge ESOL

Introduction

Cambridge ESOL, the exam board within Cambridge Assessment which provides English language proficiency tests to 3.5 million candidates a year worldwide, uses the Common European Framework of Reference for Languages (CEFR) as an essential element of how we define and interpret exam levels. Many in the UK who are familiar with UK language qualifications may still be unfamiliar with the CEFR, because most of these qualifications pay little attention to proficiency – how well a GCSE grade C candidate can actually communicate in French, for example, or whether this is comparable with the same grade in German. The issues of comparability which the CEFR addresses are thus effectively different in kind from those that occupy schools exams in the UK, even if the comparisons made – over time, or across subjects – sound on the face of it similar. This article offers a brief introduction to the CEFR for those unfamiliar with it.

Given its remarkable rise to prominence as an instrument of language policy within Europe, the CEFR has acquired detractors as well as advocates, the former painting it as a methodologically outdated, bureaucratic menace. Of those more positively disposed, some see it as a closed system, while others stress its open and unfinished nature. This article takes the latter view. It discusses the nature of constructing a link to the CEFR, and makes the case that extending the scope of the present framework to deal effectively with many linguistically complex contexts of learning is both necessary and possible.

An introduction to the CEFR

Frameworks for language proficiency can take many forms and operate on many levels. The one which this article focuses on is the Common European Framework of Reference for Languages (CEFR), which has become uniquely influential in a European context, as well as beyond Europe. What exactly is the CEFR? At one level, it is a book (Council of Europe, 2001), though one which probably few people read from cover to cover, and many misunderstand. The book is complemented by some additional material on the Council of Europe website. At another level the CEFR can be seen as a major ongoing project, an area of activity which is focusing the efforts, coordinated or uncoordinated, of many language specialists across Europe and beyond: policy makers, testing bodies, curriculum designers and teachers.

For readers unfamiliar with the CEFR it is worth outlining its distinctive features:

- It is a *proficiency* framework, with quite different aims to the currently-in-development European Qualifications Framework (EQF), whose purpose is to make national qualifications more readable across Europe. Generally, qualifications frameworks need not relate strongly to language proficiency frameworks.
- It is comprehensive in scope: as its title states, it is a framework for learning, teaching and assessment.
- It is a framework for *all* European languages (and has been applied to many non-European languages).
- Its aim is to support language learning, within the Council of Europe's general remit to promote communication, exchange and intercultural awareness within Europe.
- It is *not* an assessment system, something which frustrates those who expect to make easy comparisons with test-linked scales such as the ACTFL (American Council on the Teaching of Foreign Languages) Oral Proficiency Interview.
- It has no direct mandate, because neither the Council of Europe, who produced it, nor the European Commission, which has adopted it as an instrument of policy, has any direct authority over education policy in European member countries. However, many countries do reference it explicitly in teaching and assessment policy.

The CEFR is in fact two kinds of framework – a conceptual one, and a set of reference levels.

Conceptually, the CEFR offers a comprehensive discussion of the many ways in which contexts of learning differ. Every context of learning is unique, having its own aims and objectives, reflecting the purposes for which a language is learned, the skills to be emphasised, the teaching methodology adopted, the place of the language within a wider languages curriculum, and so on. The CEFR lays out the range of choices which must be made. This is its first purpose.

The CEFR's second purpose is to provide a set of reference proficiency levels. It claims that *despite* the differences between contexts of language learning it is possible and useful to compare them in terms of level. The levels are offered as a neutral point to which any specific context of learning can be referred. The levels are illustrated by a large number of scales: the summary table below shows the *Common Reference Levels: global scale* (Council of Europe, 2001:24).

There is no doubt that since its publication in 2001 the CEFR has acquired great prominence in Europe and beyond, particularly as an instrument of language policy, for defining learning objectives and assessing outcomes. For language testing organisations with an international market, linking their exam levels to the CEFR and providing evidence for their claims has become almost essential.

Common Reference Levels: global scale

Proficient user	C2	Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices.
Independent user	B2	Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	B1	Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans.
Basic user	A2	Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need.
	A1	Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

The CEFR proficiency levels

Where do the CEFR proficiency levels come from? Taylor and Jones (2006) provide the following account. The levels formalise conceptual levels with which English Language Teaching (schools, teachers and publishers) had operated for some years – with familiar labels such as 'intermediate' or 'advanced'. North, one of the CEFR's authors, confirms its origins in traditional English Language Teaching levels: "the CEFR levels did not suddenly appear from nowhere." (North, 2006:8). North outlines the gradual emergence of the concept of levels, referring to the Cambridge Proficiency and the First Certificate exams, now associated with C2 and B2, as well as the Council of Europe-sponsored Threshold and Waystage learning objectives, first published in the 1970s as defining useful levels of language competence now associated with B1 and A2. According to North, "The first time all these concepts were described as a possible set of 'Council of Europe levels' was in a presentation by David Wilkins (author of 'The Functional Approach') at the 1977 Ludwighaven Symposium."

What this account suggests is that the CEFR levels reflect an existing reality of some kind inherent in large populations of language learners. These learners progress through a series of stages in their learning career, each stage supported by appropriate courses, coursebooks and tests, which spring up as needed around each language. The levels are as they are because they reflect a progression of steps which are sufficiently accessible as learning targets but sufficiently distinct as learning achievements (Jones, 2005). They have developed in an organic way in response to demand, and in this sense it is not unreasonable to refer to them as 'natural' (North, 2006:8).

At the same time there is clearly a conventional element to the levels. Each educational context, and each widely-learned language, may have developed well-embedded understandings of levels (what is intended by 'intermediate' or 'advanced', for example), and accreditation systems with well-embedded standards.

Thus it seems inevitable that particular contexts or particular studied languages will tend to refer the CEFR level descriptors to somewhat different realities, and in consequence interpret them differently.

A common understanding of levels is clearly a goal worth pursuing, within education, for setting objectives and comparing performance with other contexts, and beyond education, for example in matching language competence to jobs.

However, given the nature of the CEFR there are currently no ways of enforcing a common understanding of levels, and as will be discussed below, it is by no means clear that enforcement is desirable, even if possible. What we might expect to happen is a gradual convergence of use across countries and languages, informed by authoritative points of reference. These will of necessity arise from studies with an explicitly multilingual focus.

A further issue is the adequacy of the CEFR's conception of proficiency for the range of contexts which we might wish to relate to it. The CEFR states explicitly that it is a framework for foreign language learning. However, foreign language learning is but one aspect of language education policy, and many educational contexts are characterised by considerable complexity. Language is an object of study but also the medium (whether as a first, second or foreign language) through which other subjects are studied. Increasingly, language testers are engaging in educational contexts demanding a single conceptual framework that encompasses this complexity. Another project of the Council of Europe Languages Policy Division, initiated after the completion of the CEFR, is the *Platform of resources and references for plurilingual and intercultural education* (also called the Languages of Schooling project). This group has avoided the term 'framework', and any notion of reference levels, indicating a concern with educational and social values rather than empirical scaling of proficiency. None the less, the issues which engage this group clearly complement those addressed by the CEFR, and point directions in which it might be extended. I will return to this below.

Is linking to the CEFR worthwhile?

Let us agree that the creation of common standards relating to the CEFR's reference levels is an aim worth pursuing. As stated above, this is in the intention of its authors the secondary purpose of the CEFR, its primary purpose being to offer a comprehensive, non-prescriptive presentation of the myriad options teachers and course designers face when deciding what to teach and how to teach it. It invites reflection.

As the authors state (Council of Europe, 2001:1) "We have not set out to tell people what to do or how to do it".

This openness, however, does not imply an absence of policy, and we should consider whether by buying into the CEFR we in some way risk adopting a policy which limits or misdirects the development of our approach to language education.

The CEFR refers to Council of Europe statements of policy which emphasise the satisfaction of learners' "communicative needs" including dealing with the business of everyday life, exchanging information and ideas, and achieving a wider and deeper intercultural understanding. This is to be achieved by "basing language teaching and learning on the needs, motivations, characteristics and resources of learners", and "defining worthwhile and realistic objectives as explicitly as possible" (p.3). This conveys the CEFR's basic communicative, action-oriented approach.

Some have interpreted the CEFR's approach as outdated. McNamara and Roever (2006, p.212) are typical when they criticise "the fundamental underlying construct of the assessment [sic], a 1970's notional/functionalism that was given its clearest expression in the work of Van Ek and Trim". The criticism is understandable, given the way readers are continually prompted to "consider and where appropriate state" their choices with respect to content, particularly throughout chapters four and five – *Language use and the language learner; The learner's competences* – which is where the descriptor scales appear. The apparent notional/functional emphasis thus partly results from the unintended prominence of the descriptor scales in most readers' understanding of the CEFR. In fact, the prompts in chapter 6 – *Language learning and teaching* – and the remaining chapters are almost entirely methodological in focus: what assumptions users make about the process of learning; which of a list of general approaches they use; what they take to be the relative roles and responsibilities of teachers and learners, and so on. These little-read invitations to methodological reflection allow us to see the CEFR as more open than it is generally given credit for.

The CEFR's approach is broad and should be coherent with the aims of most school language learning. It leaves scope for a range of implementations.

Furthermore, the simple notion of orienting language learning towards a proficiency framework is itself of great potential value. This, at least, was the view of the Nuffield Languages Inquiry (Nuffield Languages Inquiry, 2000; Nuffield Languages Programme, 2002), which criticised existing UK language qualifications as being bad for learning and "confusing and uninformative about the levels of competence they represented" (idem: 8). They regretted that for the most part, "beyond 14, student attainment in languages is mainly related to examination targets, and not to performance criteria in 'can do' terms" (idem: 9). The Inquiry's conclusion was that a new assessment framework should be made available based on graduated and meaningful proficiency levels. The CEFR was cited as a model.

The Inquiry's findings helped define the National Languages Strategy, launched in 2001 in the context of a deepening crisis in UK foreign language learning. A proficiency framework was defined called the *Languages Ladder* which was broadly comparable to the CEFR. *Asset Languages* was the name given to the corresponding assessment framework, developed by Cambridge Assessment for the Department of Education (then the DfES), building on an approach to construct definition, item writing and scale construction developed by Cambridge ESOL over many years of testing English as a foreign language.

The Asset Languages framework is complex, comprising 25 languages, four skills, six levels, and a degree of differentiation of age groups (as a lifelong learning framework it encompasses both children and adults). The empirical construction of this CEFR-linked framework provides a case study on the theoretical and practical challenges involved in such a multilingual enterprise (Jones, 2005; Jones, Ashton and Walker, 2010).

Beyond the technical challenges, the Asset Languages story also illustrates the practical challenge of introducing a proficiency-focused language exam into an educational system more accustomed to interpreting performance simply in terms of exam grades. Clearly, linking assessments to the CEFR will impact positively on language learning to the extent that the goals of testing and teaching are aligned (Jones, 2009).

There are critics of the CEFR who see it as a clear force for evil, a tool of authority and control – "manipulated unthinkingly by juggernaut-like centralizing institutions" (Davies, 2008:438, cited by Fulcher, 2008:21). Consider, for example, this recent recommendation by the Council of Ministers (Council of Europe, 2008b), which calls on countries to make reference to the CEFR, and specifically in relation to assessment, to:

> *ensure that all tests, examinations and assessment procedures leading to officially recognised language qualifications take full account of the relevant aspects of language use and language competences as set out in the CEFR, that they are conducted in accordance with internationally recognised principles of good practice and quality management, and that the procedures to relate these tests and examinations to the common reference levels (A1–C2) of the CEFR are carried out in a reliable and transparent manner.*

Such statements could certainly be seen as conducive to a bureaucratised adoption of the CEFR, notwithstanding the benign intention of its authors. As Trim, one of those authors, concedes: "there will always be people who are trying to use it as an instrument of power" (Saville, 2005: 282).

Language assessment providers should of course be accountable for the quality of their exams. But how should this be done? Some would see this as a process which can and should be standardised, and even policed by some suitably-instituted authority (Alderson, 2007: 662). A basis for such standardisation might be seen in the *Manual for relating language examinations to the Common European Framework of Reference for Languages (CEFR)* (Council of Europe, 2008a), which together with an extensive reference supplement and various further materials offers practical instructions. Should this be the core of an essentially regulatory and bureaucratic process?

The Council of Europe has rejected the suggestion of fulfilling a policing role, and the authors of the Manual themselves disclaim the idea that it defines a necessary and sufficient process.

The nature of linking to the CEFR

The main problem with understanding the issue as one of regulation or standardisation is that it seems to require, and would militate in the direction of, a closed, static system rather than an open and developing one. The construction of a comprehensive language proficiency framework must be seen as a work in progress, still needing much work to be done. This is a creative process because there are many contexts of learning that might usefully be linked to the CEFR, but which are not currently well described by the CEFR.

So it is the context which is critical. Jones and Saville (2009:54–5) put it thus:

> *… some people speak of applying the CEFR to some context, as a hammer gets applied to a nail. We should speak rather of referring a context to the CEFR. The transitivity is the other way round. The argument for an alignment is to be constructed, the basis of comparison to be established. It is the specific context which determines the final meaning of the claim. By engaging with the process in this way we put the CEFR in its correct place as a point of reference, and also contribute to its future evolution.*

The CEFR levels are illustrated by a large number of descriptor scales describing activities (*addressing audiences; reports and essays*) and competences (*vocabulary control; grammatical accuracy*). We should look critically at these. They aim to be *context-free* but *context-relevant*, that is, relatable to or translatable into each and every relevant context (Council of Europe, 2001:21). A framework of reference should describe no specific context of language learning, but be framed in terms which allow widely differing contexts to find common points of reference, and implicitly, of comparison. This is easier said than done. A great virtue of the descriptor scales in the body of the CEFR is that they were developed through an empirical study (North, 2000); but this also makes them specific to the context of that study, which most closely resembles a standard language school setting. School contexts involving young children, or with instruction through the medium of a foreign language, for example, might require quite different description.

Moreover, despite the use of the term 'illustrative', it is clear that the scales function as *definitions* of the reference levels, in the way they are selected from to compile the global descriptor scales, or in a discussion of the salient features of the levels, where each level is epitomised through a compilation of selected descriptors (Council of Europe, 2001: 3.6). The description seems complete: it is hard to imagine how a particular context of learning could be *differently* characterised.

Milanovic (2009) points out that in an earlier draft of the CEFR the illustrative descriptors were included in an appendix, a layout which "visibly reinforced the different status and function of the general *reference levels* and more specific *illustrative scales*." He criticises the 'overly prescriptive' way in which the illustrative scales have come to be used, citing the earlier draft, in which it is acknowledged that:

> *The establishment of a set of common reference points in no way limits how different sectors in different pedagogic cultures may choose to organise or describe their system of levels and modules. It is also* ***to be expected that the precise formulation of the set of common reference points, the wording of the descriptors, will develop over time*** *as the experience of member states and of institutions with related expertise is incorporated into the description.*
> (Council of Europe, 1998:131; emphasis added)

So each context, if sufficiently distinct, may need its own illustration.

We should also be cautious of characterising levels and progression solely in terms of behavioural outcomes illustrated by can-do descriptors. The CEFR scales tend to emphasise these, because as the authors state, being observable, such language activities provide "a convenient basis for the scaling of language ability" (Council of Europe, 2001:57). Weir (2005) criticises the absence of a theoretical model of cognitive development, without which, he argues, the CEFR does not equip testers to defend the validity or comparability of their tests.

Extending the CEFR framework

What range of contexts can the CEFR encompass? As Coste, one of the CEFR's authors has said, contextual uses can take "various forms, apply on different levels, have different aims, and involve different types of player". In his view: "All of these many contextual applications are legitimate and meaningful but, just as the Framework itself offers a range of (as it were) built-in options, so some of the contextual applications exploit it more fully, while others extend or transcend it." (Coste 2007).

Relating contexts to the CEFR inevitably leads us to extend or transcend it. I have already mentioned contexts which are not well described by the present CEFR even within its stated remit as a framework for foreign languages:

- Young children, that is, situations where what learners can do is defined both by language proficiency and cognitive stage.
- CLIL (Content and Language Integrated Learning) situations, where the content of a school subject is taught through the medium of the language being studied.

We can easily see the current CEFR as an instance of a more general framework, which happens to be parameterised and illustrated for the case of foreign language learning in certain contexts. More parameters could be added where needed, extending the framework to other contexts without changing its relevance or meaning in contexts which it already encompasses. As Cambridge ESOL engages increasingly with linguistically complex educational contexts the need for such an extended framework becomes increasingly evident, and it is fairly clear in what respects the CEFR needs extending. Additional dimensions to be developed include:

- Cognitive development stages, which are closely linked to all linguistic development, as well as to the process of concept formation, which from school age is largely mediated through language.
- Language as the medium of schooling, as distinct from language for social interaction. This is Cummin's distinction between Cognitive Academic Language Proficiency (CALP), a high level of competence necessary for academic success, and Basic Interpersonal Communicative Skills (BICS), which can be more readily acquired through social interaction (Cummins, 1984). In the CEFR 'CALP' is very much the stuff of the C levels, but where a child is acquiring schooling through the medium of a second language, it is involved from the outset.
- Foreign Language (language for its own sake) as distinct from Second Language (language for some extrinsic purpose).
- Mother tongue language (MTL), which is characterised by the linguistic reflexes of a developed socio-cultural competence (culture in the 'broad' sense): a shared grasp of idiom, cultural allusion, folk wisdoms, etc. MTL speakers may master both *restricted* and *elaborated* codes (Bernstein, 1973).

Such an inclusive framework will enable a coherent approach to language education, recognising synergies between different language competences, and the different purposes of language use in an educational setting and in society. Interestingly, in a foreword to a newly-revised ALTE guide to assessment, Joe Shiels, Head of the Council of Europe Languages Policy Division, points to the Council's efforts to promote a "global approach to all languages in and for education" and

calls on language testers to address the "new challenges for curriculum development, teaching and assessment, not least that of assessing learners' proficiency in using their plurilingual and intercultural repertoire" (ALTE, 2011). Is this an invitation to extend the CEFR in the way outlined here? We need such an inclusive framework because learners with different language backgrounds co-exist and intersect within educational settings which direct their learning, and qualifications frameworks which compare and judge them, on their language, or other skills mediated by language. Beyond education, they share all the personal and professional opportunities that specific language skills afford.

An example will illustrate how the extended framework will make it easier to describe and compare different groups. According to the CEFR: "Level C2 … is not intended to imply native-speaker or near native-speaker competence. What is intended is to characterise the degree of precision, appropriateness and ease with the language which typifies the speech of those who have been highly successful learners" (Council of Europe, 2001:36).

But some C2 descriptors of educated competences evidently denote levels of skill well beyond the capacity of many native speakers. So if native speakers are lower than C2 in some respects, in what respects might they be higher, and do we need a D level to describe them? As noted above, MTL speakers possess a socio-cultural competence (culture in the 'broad' sense) which few foreign language learners will acquire. They may master several codes, and naturally move between them.

By distinguishing these skills from the educated, CALP competences which native speakers may well not acquire, while foreign learners can, we can describe two distinct kinds of C-ness and avoid setting one above the other.

The heterogeneous nature of the dimensions in the extended framework do not prevent a coherent approach to defining levels. As the history of the development of the CEFR levels illustrates, the lowest identified level is the first point at which there is any significant competence to describe (where 'significant' represents a social value judgement). It is interesting that as far as ESOL goes, over the years that level has moved progressively lower: in 1913 it stood at C2, with the Cambridge Proficiency (CPE) exam. By 1939 it had moved down to B2 with what became First Certificate. Then in the 1970s it moved down through B1 (Threshold level) to A2 (Waystage). Currently it stands at A1, but there are already many contexts where A1 is being sub-divided to provide a lower first objective.

The highest identified level is the last one worth describing because it is observed sufficiently frequently in the relevant population to be useful; that is, we exclude exceptional cases of literary, intellectual or linguistic brilliance. For ESOL, the CPE exam still exemplifies the C2 level. Some people argue that the CEFR description of C2 is a higher level than CPE, but a counter-argument to that is: if C2 were any higher, it would not exist, because a sufficiently large group of learners seeking accreditation at that level would not exist. In this way the need for a D level is eliminated, unless we wish to reserve a category for the truly exceptional (interpretation, for example, might qualify, as a skill quite beyond ordinary language use).

Conclusion

In this article I have introduced the CEFR and claimed that its reference levels have a kind of reality inherent in populations of learners; but that this means that different educational contexts may tend to have different understandings of them. I made a positive case for linking assessment to the CEFR, but argued against the view that linking to the CEFR could or should be a formally standardised or policed process, and in favour of a conception of linking which treats each context of learning on its own terms, and in this way progressively enriches the CEFR and leads to improvements in its articulation. Finally, I made specific proposals for extending the CEFR so that those who have the requirement to work in linguistically complex contexts should be able to do so within a single coherent framework.

I have not gone into detail here regarding the technical and practical issues involved in aligning language tests or setting standards within a proficiency framework, even though Asset Languages (Jones, Ashton and Walker, 2010), and the currently in-progress European Survey on Language Competences (www.surveylang.org), are two projects which have offered ample first-hand experience and a number of lessons. This is material for a different article.

I believe the aim of linking different languages and contexts to a common framework is a meaningful one which can bring benefits. The explicitly multilingual assessment context is the one which has most to offer the CEFR project, if our goal is to move progressively towards something like a common understanding of levels. Comparison across languages and contexts is vital. We should, as far as possible, base our comparisons on what we can discover about learners, rather than their performance on tests. Finally, I think that in constructing the argument that links a learning context to the CEFR we could focus with benefit on the partial, scaffolded nature of classroom language competence (Jones, 2009). There is formative potential in articulating the chain of activities and observations that link the inputs to learning to their intended outcomes in communication.

References

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal*, **91**, 4, 659–663.

ALTE/Council of Europe (2011). *Manual for Language Test Development and Examining for use with the CEFR*. Strasbourg: Council of Europe.

Bernstein, B. (1973). *Class, codes and control, vol. 1*. London: Routledge & Kegan Paul.

Coste, D. (2007). *Contextualising uses of the Common European Framework of Reference for Languages*. Paper presented at Council of Europe Policy Forum on use of the CEFR, Strasbourg 2007, available online www.coe.int/T/DG4/Linguistic/Source/ SourceForum07/D-Coste_Contextualise_EN.doc

Council of Europe (1998). *Modern Languages: Learning, Teaching, Assessment. A Common European Framework of Reference*. Strasbourg: Language Policy Division.

Council of Europe (2001). *Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press. Also available for download at: http://www.coe.int/T/DG4/Linguistic/Default_en.asp

Council of Europe (2008a). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) A Manual*. Strasbourg, France: Council of Europe.

Council of Europe (2008b). Recommendation CM/Rec(2008)7 of the Committee of Ministers to member states on the use of the Council of Europe's Common European Framework of Reference for Languages (CEFR) and the promotion of plurilingualism. Strasbourg: Council of Europe.

Cummins, J. (1984). *Bilingualism and special education*. Clevedon: Multilingual Matters.

Davies, A. (2008). Ethics and professionalism. In: E. Shohamy (Ed.), *Language Testing and Assessment*. *Vol*. *7 Encyclopedia of Language and Education*, 429–443. New York: Springer.

Fulcher, G. (2008). Testing times ahead? *Liaison Magazine*, Issue 1, July 2008.

Jones, N. (2005). Raising the Languages Ladder: constructing a new framework for accrediting foreign language skills. *Research Notes No 19*. Cambridge ESOL. Retrieved from: http://www.cambridgeesol.org/rs_notes/rs_nts19.pdf

Jones, N. (2009). The classroom and the Common European Framework: towards a model for formative assessment. *Research Notes Issue 36*, May 2009. Available from: http://www.cambridgeesol.org/rs_notes/offprints/pdfs/RN36p2-8.pdf

Jones N., Ashton K. & Walker T. (2010). Asset Languages: a case study of piloting the CEFR Manual. In: M. Milanovic & C.J. Weir (Eds), *Relating language examinations to the Common European Framework of Reference for Languages: Case studies and reflections on the use of the Council of Europe's Draft Manual*. *SiLT volume 33*. Cambridge: CUP.

Jones, N. & Saville, N. (2009). European language policy: assessment, learning and the CEFR. *Annual Review of Applied Linguistics*, **29**, 51–63.

McNamara, T. & Roever, C. (2006). *Language Testing: The Social Dimension*. Oxford: Blackwell.

Milanovic, M. (2009). Cambridge ESOL and the CEFR. *Research Notes 37*. Cambridge ESOL.

North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Lang.

North, B. (2006). *The Common European Framework of Reference: Development, Theoretical and Practical Issues*. Paper presented at the symposium 'A New Direction in Foreign Language Education: The Potential of the Common European Framework of Reference for Languages', Osaka University of Foreign Studies, Japan, March 2006.

Nuffield Languages Inquiry (2000). *Languages: the next generation*.

Nuffield Languages Programme (2002). *A Learning Ladder for Languages: possibilities, risks and benefits*. Retrieved from: http://languages.nuffieldfoundation.org/filelibrary/pdf/learning_ladder.pdf

Saville, N. (2005). An interview with John Trim at 80. *Language Assessment Quarterly*, **2**, 4, 263–288.

Taylor, L. & Jones, N. (2006). Cambridge ESOL exams and the Common European Framework of Reference (CEFR). *Research Notes 24*. Cambridge ESOL.

Weir, C. J. (2005). Limitations of the Council of Europe's Framework of reference (CEFR) in developing comparable examinations and tests, *Language Testing*, **22**, 3, 281–300.

The challenges of ensuring year-on-year comparability when moving from linear to unitised schemes at GCSE

Mike Forster Head of Research & Technical Standards, OCR

Introduction – the new GCSE suite

In September 2009, teaching began on a new suite of GCSE specifications. These were developed by the UK awarding bodies in order to meet new subject criteria specified by QCA. With the exception of English, mathematics, science and ICT (which were being developed to a different timescale), these new specifications covered all the available subjects. The outgoing specifications were mostly 'linear', whereby candidates took all the components of their assessment at the end of their course of study. The new GCSE specifications were all 'unitised' (also known as 'modular'). This meant that candidates could take units of assessment at different stages of their course, with the first units assessed in January 2010. As the nature of unitised and linear specifications is very different, it was imperative to ensure that appropriate unit grade boundaries were set, so that the first full course aggregation outcomes in June 2011 were appropriate.

OCR was in a unique position to be able to offer advice on this, in that it had been running a selection of unitised GCSEs for a number of years. Specifications in ICT, Business Studies, English and English Literature had been available with unitised assessment for up to nine years, and underwent little modification throughout the lifetime of the specifications. These were therefore the most appropriate specifications to use to illustrate the issues and difficulties facing those awarding (i.e. setting grade boundaries on) new units in the new specifications. Further background to the issue of modularisation can be found in D'Arcy (1997) and Vidal Rodeiro and Nádas (2010).

Comparability

There are many different definitions and contexts of comparability. In this article year-on-year comparability is the main concern. In an open letter to secondary schools and colleges dated 14th March 2011, Ofqual (2011) noted that the principle followed for the first awards of the new A levels in summer 2010, namely that there should be consistent standards at subject level between the old and the new specifications, would apply to the new suite of GCSEs in summer 2011. As such, Ofqual noted "we anticipate that the overall national results in summer 2011 for a particular subject will be reasonably close to the results in that subject in previous years". This particular definition of comparability is based on statistical comparability, as opposed to one based on expert judgement of the quality of candidates' work.

Thus the underlying imperative was to ensure that candidates of a given ability would achieve the same grade in the new specification as they would have done in the old (legacy) specification. In a time of relative stability, in comparability terms this would be what Newton *et al*. (2007) described as the "straightforward situation" of a parallel test, where the test is essentially assessing the same content in the same way, but with different questions. However, the restructuring of the new specifications and the change in the assessment model meant this was not that straightforward. Newton *et al*. also described a more complex situation – comparability of non-parallel versions of the test – in which the versions of the test are different in more than just the questions alone, for example with changes to both content and structure. It is this variant of year-on-year comparability that is central to this report.

Unitisation – the main issues

When linear specifications are awarded, the overall effect of component grade boundary decisions can be seen in the aggregation as a whole. If overall outcomes are not deemed appropriate, it is possible to revisit component boundaries until a satisfactory outcome is achieved. With unitised specifications, this is not possible in the same way. Aggregation outcomes for the new specifications were not available until June 2011. For any units taken before this date, decisions had to be taken whose impacts on overall outcomes would not be known until June 2011. Unit grade boundaries thus needed to be set with a view to satisfactory aggregation outcomes at a later date.

One advantage for candidates taking unitised specifications is the opportunity to resit units to improve their overall outcome. In the new unitised GCSEs, candidates were permitted one retake opportunity for each unit before requesting certification, and usually this would either maintain, or improve, overall grade outcomes, as the best unit outcome would be used towards the aggregation (subject to resit and terminal rules[1]).

One of the artefacts of the assessment of linear GCSE specifications in the UK is the use of two indicators for setting overall aggregation boundaries. The two indicators represent two possible boundary marks, and the chosen boundary is *normally* the lower of the two (as noted in Appendix 2 of the Ofqual code of practice (Ofqual, 2010)). Indicator 1 is the sum of the weighted boundary marks on each component, whilst Indicator 2 uses a weighted average percentage calculation to produce a boundary which yields an overall outcome more akin to the component outcomes. At the top judgemental grades (e.g. grade A on the Higher tier, and grade C on the Foundation tier) the Indicator 2 boundary is normally lower than the Indicator 1 boundary, with the opposite true for the lower grades (below the mean mark). This means, for example, that candidates could achieve a grade A overall (just!) by getting a top grade B on each component. In a unitised qualification, the overall specification boundaries are simply the sum of the unit uniform mark boundaries

1 The resit rules allow two attempts at each unit before certification (sometimes called 'cash-in'). Once a candidate has certificated, another two attempts are permitted before further certification, and so on. The terminal rule states that at least 40% of the assessment must be taken in the same session as certification. The marks from these units must contribute to the final mark, even if a better mark is available from an earlier attempt at the unit.

(see later for an explanation of the uniform mark scheme, UMS) – there is no Indicator 2. This meant that, unless allowance was made at unit level, candidates taking the new unitised specifications could have been disadvantaged at the top grades on each tier in comparison with candidates taking the old linear specifications[2].

The new suite of GCSEs included a 'terminal requirement'. This required candidates to take at least 40% of their assessment for a specification in the same session in which they aggregated (the process of aggregating all unit marks into one total, and receiving an overall grade – also called 'certification' or 'cash-in'). These units had to count towards the overall aggregation, and as such may have negated some of the benefit of resitting. It should be noted that no units were particularly designated as 'terminal units' – the terminal requirement could be met using any units (subject to meeting the 40% requirement and the resit rule).

A number of issues that could have affected unit outcomes needed to be considered, particularly in the first award of the unit. These could have made grade distributions appear odd, even if the outcome in grade terms was entirely appropriate. One such issue was candidate maturity. Candidates could take units throughout the course of study, and after as little as one term of teaching. As such, candidates who entered units early may not have had the maturity and knowledge of their linear counterparts. The performance required to achieve a given grade was the same regardless of the session of entry or maturity of the candidates, that is, the full GCSE standard was applied to all units. Assessors did not know the age of the candidates whose work they were marking, and no allowance was made for a lack of maturity. Therefore any lack of maturity or subject knowledge would have been evidenced by lower grade outcomes. This was especially the case for subjects such as modern foreign languages, where the nature of the cognitive learning process is more cumulative. It is also worth noting the difficulty in making judgemental decisions about the quality of work on units that assess (usually) smaller chunks of the specification, and add together in a different way from the legacy components.

An issue working in the opposite direction was that candidates could have gained an advantage (and hence improved their grade) as a result of the course assessment being broken down into 'bite-size' chunks, as they only needed to focus on one area of the specification at a time. Again, however, this benefit was constrained to some extent by the terminal rule, as candidates had to take at least 40% of the assessment at the end of their period of study. Ofqual's expectation was that there would be similar outcomes under the unitised scheme to those under the linear scheme. It was clear, therefore, that the pros and cons of the unitised scheme would to some extent cancel out, thus helping to ensure the structure of the assessment per se did not advantage or disadvantage the first cohort taking the new assessments.

Finally, centre entry strategies might also have produced misleading unit outcomes. Some centres might have entered only their most able candidates in the early sessions, whilst others might have entered all candidates to allow them to get a feel for what was expected. If the latter occurred in large enough numbers, the outcomes could have been very misleading indeed (even if they were entirely appropriate). Chairs of Examiners were therefore provided with age data about their cohort, which was used to support the awarding process.

Moving to a uniform mark scheme

This article has already identified a number of issues that could have had an impact when moving from a linear to a unitised specification. One such issue was the effect of introducing a uniform mark scheme, a necessity for a GCSE assessment that permitted units to be taken on different occasions.

As linear specifications assess candidates in one assessment window, the means by which candidates' scores are combined in order to produce an overall score is straightforward. When specifications are unitised, candidates can take units on different occasions, and it is therefore necessary to ensure parity between these units. This is achieved through a common mark scale – the uniform mark scheme. Raw scores are transposed onto a common mark scale such that equivalent scores (in terms of performance, not in terms of marks) from different sessions achieve the same number of uniform marks. Thus the candidate who scores the A boundary mark on a very difficult paper will get the same number of uniform marks as the candidate who gains the A boundary mark on a much easier version of the paper in another session. (See AQA, 2009; or Gray and Shaw, 2009 for further details). The mark transformations used in aggregating linear specifications are linear, according to the weighting of the components. In unitised specifications, the conversion rate for raw to uniform marks is not necessarily constant across the mark range[3]. This can result in compression or stretching of the raw-UMS conversion scale.

OCR replicated this effect by 'unitising' a number of existing linear specifications, to see the effect on grade outcomes. The outcomes varied from specification to specification, but there were some identifiable trends:

- On untiered specifications, most candidates tended to get the same grade following 'unitisation' as they had originally, but where grade changes did occur they tended to be downwards at the top grades, and upwards at the bottom grades.
- On the Foundation tier, most candidates tended to get the same grade as they had originally, but where grade changes did occur they tended to be downwards. On some specifications, this downward trend was restricted to the top grades, and on other specifications it was across all grades.
- On the Higher tier, as on the Foundation tier, most candidates tended to get the same grade as they had originally. Where there were grade changes, they tended to be downwards at the top grades, and upwards at the bottom grades.

These trends fitted with the expected impact of the removal of Indicator 2, namely that the proportion at the top grades would fall, but that at the lower grades the changes would be much smaller (or not there at all, if the boundary was set at Indicator 1). This supported the need to identify and act on the impact of the removal of Indicator 2 (see section below). However, there were also fluctuations at the bottom of the grade distribution, which suggested that subject-specific variations were also occurring.

2 Where the lower indicator is not chosen as the boundary mark (and assuming allowances are not made elsewhere), candidates at the lower grades on each tier who took a unitised specification would be advantaged over those who took a linear specification.

3 Between judgemental grades the conversion is linear (or quasi-linear if the intermediate gap is not equally divisible).

Unit versus aggregation outcomes

One of the major challenges facing awarders of the new GCSE specifications was setting new unit standards that would lead to acceptable specification outcomes when candidates aggregated for the first time. However, analysis of the existing unitised specifications showed little pattern between unit and aggregation outcomes. In some cases candidates gained the same grade on the units they had taken earliest as they gained when they ultimately aggregated. In other cases they did notably worse on the earliest units. Also, the introduction of new specifications (whether linear or unitised) invariably leads to a change in cohort, and most specifications also take time to find stability. As such, the outcomes on new units in new specifications may bear little resemblance to outcomes on the equivalent parts of old specifications. This was especially the case in some of the units taken in January 2010, whereby changes to the cohort in terms of centre type and age profile led to outcomes very unlike those seen in the legacy specifications.

To help ensure comparability year-on-year, OCR was able to account for the removal of Indicator 2 in the new unitised specifications by making unit-level adjustments to account for aggregate-level differences. Chairs of Examiners, who are responsible for the standards set in their awards, were presented with data which demonstrated the likely impact for each specification of the removal of Indicator 2. This was a basic numerical calculation showing the difference in the percentage in grade at the Indicator 1 boundary and the chosen boundary (usually Indicator 2). These impacts were then factored in to the awards at unit level to ensure overall outcomes were appropriate.

The issue of regression to the mean was also relevant. This is the situation in which component-level outcomes at the top and bottom grades are not maintained at aggregate level. Once aggregated, overall outcomes at the top grade tend to be lower than are found at component level, whilst overall outcomes at the bottom grade tend to be higher than are found on the components. The impact of this regression to the mean is determined by the correlation between each component (unit), and the number of components (units) in the assessment. If the number of components in the legacy linear specification was less than the number of units in the new unitised scheme, then there would have been greater regression to the mean in the unitised scheme, which would have affected overall outcomes.

Resitting pattern

The resit patterns for specifications that had previously been unitised (ICT, Business Studies, English, English Literature) were investigated. The resit rule permitted only one retake of any unit. The patterns for each unit, and each specification, varied somewhat. On the whole, the majority of candidates who resat a unit improved their mark, but this was not always the case. There appeared to be no overall pattern as to which session was the best to resit a unit in, with individual units showing different characteristics. The size of the mark changes varied too, although candidates who resat and improved their UMS mark tended to show larger mark changes than those who resat and got a lower UMS mark. This is not surprising since the resit candidates are a self-selecting sample, and few candidates would embark on their resit expecting to gain a lower UMS mark.

Year 10s and Year 11s

Unitisation offers the opportunity for candidates to sit a unit at any session in their course of study. The age profile of the cohort can have an effect on the outcomes for any unit. The majority of candidates who took the first units in the new suite of GCSEs in 2010 were from Year 10 (14 or 15 year olds), and as such their unit grade outcomes were below those seen in the components of the legacy linear specifications, which were mostly taken by Year 11 candidates (15 or 16 year olds). In comparison with the age profiles on the legacy specifications, this was much more variation in the age of candidates taking unitised specifications, and this could have affected the grade distributions for these units. To help with the setting of grade boundaries, Chairs of Examiners received the age profile of the cohort taking each unit.

Other statistical data

For new units in new specifications, the issues discussed presented a number of challenges to awarders. In summer 2011, these specifications certificated for the first time. Since achieving year-on-year comparability was paramount, it was possible to support the unit awards in this session with data about expected outcomes at specification level, based on measures of prior attainment, which Chairs of Examiners could use as one of the many indicators to help in setting appropriate standards. One of the main indicators for these specifications was a cohort-level prediction based on the average performance of the cohort at Key Stage 2 (KS2) – their average results in the national tests taken at age 11 in English, Maths and Science. This was achieved by establishing a matrix of mean KS2 performance against GCSE outcome for each subject, based on a historical pattern, and applying this pattern to the 2011 cohort for each subject to create a predicted outcome.

Summary

This article has noted the main issues that arose when the new unitised GCSE suite was examined for the first time in January 2010, and subsequently certificated in June 2011. The availability of resits, the loss of Indicator 2, the effect of maturity, the terminal requirement, and the introduction of a uniform mark scheme all had an impact on grade outcomes. The loss of Indicator 2 meant that, without correction at unit level, the proportion of candidates in the top grades (on both tiers) would have fallen. This was a fairly consistent finding. However, the other evidence was not so predictable. The analysis of unit outcomes against overall outcomes again showed a mixed pattern.

Data about resits on the existing unitised schemes also showed a mixed pattern. Most candidates improved when they resat a unit, but this was not always the case. Nor was there consistency in performance by session, with candidates in some specifications benefitting from a late resit (i.e. at the time of aggregation), whilst other candidates showed a similar improvement in each session. The make-up of the cohort by year group showed interesting outcomes, but again there was a lack of consistency. In most instances the Year 11 candidates out-performed the Year 10 candidates, but in some specifications the opposite tended to be the case.

The evidence from this article highlights the difficulties in accurately predicting what would happen when the new unitised GCSE

specifications were awarded for the first time. There were numerous factors influencing the outcomes, and whilst these were for the most part identifiable, there was no consistency in the patterns seen. What was true for one specification did not necessarily hold true for another. It was, therefore, crucial that Chairs of Examiners, and their awarding committees, used their judgement and experience, coupled with the statistical data available, to achieve outcomes that were comparable with previous years, and hence which did not give an advantage or disadvantage to the first cohort taking the new qualifications.

References

AQA (2009). Uniform marks in GCE, GCSE and Functional Skills exams and points in the Diploma. http://store.aqa.org.uk/over/stat_pdf/UNIFORMMARKS-LEAFLET.PDF Accessed 16/02/10.

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

Gray, E., & Shaw, S. (2009). De-mystifying the role of the uniform mark in assessment practice: concepts, confusions and challenges. *Research Matters: A Cambridge Assessment Publication*, **7**, 32–37.

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

Ofqual (2010). *GCSE, GCE, principal learning and project code of practice*. Coventry: Ofqual.

Ofqual (2011). *Open letter to secondary schools and colleges about the summer 2011 awards of new unitised GCSEs*. Coventry: Ofqual.

Vidal Rodeiro, C. & Nádas R (2010). Effects of modularisation. http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers

The pitfalls and positives of pop comparability

Nicky Rushton, Matt Haigh and Gill Elliott Assessment Research & Development Comparability Team

Introduction

During recent years the media debate about standards in public examinations has become something of an August ritual. The debate tends to be polarised with reports of 'slipping standards' at odds with those claiming that educational prowess has increased (Lebus, 2009). Some organisations have taken matters into their own hands, and have carried out their own studies investigating this. Some of these are similar to academic papers; others are closer in nature to a media campaign. In the same way as 'pop psychology' is a term used to describe psychological concepts which attain popularity amongst the wider public, so 'pop comparability' can be used to describe the evolution of a lay-person's view of comparability. Studies, articles or programmes which influence this wider view fall into this category and are often accessed by a much larger audience than academic papers. In this article, five of these studies are considered: Series 1 of the televised social experiment "That'll Teach 'em", conducted by TwentyTwenty Television; The Five-Decade Challenge, mounted by the Royal Society of Chemistry; the *Guardian*'s and the *Times*' journalists (re)sitting examinations in order to experience their difficulty first-hand; a feature by the BBC Radio 4 programme, 'Today' (2009), which asked current GCSE students to examine and discuss exam papers from 1936; and a book of O level past papers and an associated newspaper article which described students' experiences of sitting the O level exams.

Experiments like these are largely unreported amongst the academic community, but they are influential within the popular press. This article explores the strengths and weaknesses of the studies, questions whether they should be taken into greater account by the academic community, and investigates the extent to which they help or hinder public perceptions of standards of qualifications in schools.

"That'll Teach 'em"

"That'll Teach 'em" was a television series which achieved worldwide success. Versions of the format were developed in Holland, Germany, Belgium, France, Norway and Spain. There have been three series of the show in the UK: the first airing in 2003 (recreating a 1950's grammar school and featuring academically high-achieving pupils); the second in 2004 (a 1960's secondary modern, focused upon vocational skills); and the third in 2006 (a grammar school again, this time focusing on single-sex classes). Series 1, which will be the focus of discussion in this article, was watched by 3.25 million viewers.

The purpose of the programme was to provide both entertainment and an investigation of examination standards in the UK. Thirty students who had just finished sitting their GCSE examinations undertook to board at the 'school' set up by the programme makers. They had four weeks of 1950's style lessons as well as experiencing the living conditions, food and discipline of the era. At the end of the experiment they sat a partial GCE 'O' level exam in four subjects (Maths, English, English Literature and History) that was marked to the standards of the 1950s.

The experiment addressed a number of features which are often unrecognised when long-term standards over time (i.e. when comparisons span a large number of intervening years) are addressed in the media. Students were:

- removed from their usual environment and placed into a situation resembling that of the period of history being compared as closely as possible;
- taught a 1950's curriculum for a period of four weeks;
- taught according to the 1950's style for a period of four weeks;
- fed 1950's food.

This being a television show, there were additional concerns above and beyond the social experiment – the programme needed to make interesting viewing and to be accessible for a wide audience. Thus, the actual televised episodes would have been edited with this in mind, which might have detracted from the explanation and investigation of standards over time. Also, whilst the students were experiencing the teaching style and living conditions of the 1950s they were also being followed by a camera crew, which may have caused distraction.

The strengths of the programme included debating the topic of standards over time in a public context in a way in which the context of changes in society in the corresponding time were not only acknowledged, but put into the heart of the debate. It was not just about how well students might fare when given question papers from the era, but about what the whole experience of education was like. Much of the discussion was not about how 'standards' differ, but about how experiences differ, and that is a crucial distinction when considering long-term standards over time.

The major limitations of this study as an exercise in investigating standards over time were that: (i) the student sample was small, so it was not possible to draw a great deal from the ultimate examination results; (ii) the experiment was conducted for purposes of entertainment as well as investigation, so a more 'academic' report on its findings was not commissioned; and (iii) although the programme makers went much further than many other commentators in engaging with the social context of the time in question, the students were still modern students experiencing a previous culture, rather than truly representative of the previous era. So, although the students 'experienced' a 1950's curriculum and, to an extent, the lifestyle of the time, their underlying knowledge of the school equipment, teaching styles, and home and school environments pertaining to the 21st century would inevitably have influenced their learning and behaviour during the course of the experiment. The limited time scale of the experiment was also a drawback – students were not undertaking a two-year course, and were

only taking shortened versions of O level papers. Also their motivation for undertaking the experiment and all that it entailed would have been very different from their motivation towards their high-stakes GCSE examinations.

The students' exam results (from the partial O level set at the end of the study) showed a relationship with their GCSE results (which had been taken just before the experiment). In all four subjects there was a trend for students who attained a grade A* at GCSE to score higher on average in the partial O level examination than students who achieved an A at GCSE, who themselves scored higher on average than students who achieved B at GCSE. This trend held for English Literature, English Language and Maths at GCSE grades A*–D (no students scored lower) and in History at grades A* to B. These statistics were based upon very small samples of students, and hence were highly unreliable, but nevertheless showed a reassuring trend.

The Five-Decade Challenge

This study was carried out by the Royal Society for Chemistry (RSC) in 2008. Over a thousand students sat an examination paper containing numerical and analytical chemistry questions drawn from O level and GCSE exams spanning 1965 to 2005. The average scores from the questions in each decade were used as a measure of standards over time. The results are given in Table 1.

Table 1: Five-Decade Challenge results

Decade	*Average score*
1960s	15.4%
1970s	18.2%
1980s	22.2%
1990s	34.9%
2000s	35.3%
All questions	25.5%

The report concluded that:

> *Performance against each decade showed a remarkably steady step-wise progression, with the average scored for the 1960's questions being 15%, rising to 35% for the current 2000's decade. Changes to the syllabus and to the language used in examinations since the 1960s may partially explain this progression, but are unlikely to provide a complete explanation.* (Royal Society of Chemistry, 2008, p.2).

It is interesting to note that the report did not make any explicit reference to declining standards, but focused its recommendations on changing curriculum content to emphasise quantitative and analytical science skills. However, the headline writers used the report to make significant pronouncements on science standards:

- *The proof that science exams are easier now* (*Daily Mail*, 2008)
- *Dumbed-down science is 'failing a generation'* (*Daily Telegraph*, 2008)
- *School science standards are slipping, says study* (*Guardian*, 2008)
- *GCSE students flunk past papers in experiment that exposes decline in standards* (*Independent*, 2008)

The study had strengths in its sample size and addressed issues of student motivation by offering financial reward for top scores. However, the method for selecting questions for the test may have meant that the items from each decade were not representative of the 'standard' at the time. It can also be argued that the use of average score to represent a standard was meaningless without being referenced to the overall 'pass-mark' at that time. For example, it could be that in the 1960s a score of 15% represented a 'pass' but in the 2000s a score of 40% represented a 'pass'. If that were the case, the results in Table 1 would have to be interpreted very differently with respect to the relative difficulties of the exam questions.

The study did acknowledge the potential influence of curriculum changes on the outcomes. Usefully, the report identified which questions were part of the current Chemistry curriculum. Restricting analysis to this set, which was not done in the original report, provides the alternative results in Table 2.

Table 2: Five-Decade Challenge results restricted to questions in the current curriculum

Decade	*Average score*
1960s	51.5%
1970s	27.5%
1980s	39.6%
1990s	34.2%
2000s	35.2%

Although the restricted analysis is based on fewer questions, the outcomes are very different. If average score on questions testing the current curriculum is used as a proxy for standards over time, then science exams in the 1960s were much easier than they are today.

The overt lobbying for curriculum change contained in the report and the lack of peer review probably make the findings less credible for academic researchers. However, the method and ideas contained within the study could provide a stimulus for further research.

The report provided helpful commentary to aid the public perception of standards; however, the media representation radically oversimplified and made judgements on standards not supported by evidence in the report. The RSC seemed to suggest that comparing standards over time is worth pursuing, but acknowledged the complexity of confounding factors. It is interesting to note the government response contained in the *Daily Telegraph* article reporting on the study:

> *... exam standards are rigorously maintained by independent regulators and we would rather listen to the experts whose specific job it is to monitor standards over time.* (*Daily Telegraph*, 2008)

Commentators re-sitting examinations

Some smaller scale experiments reported in the popular press have involved journalists sitting A level papers in order to draw conclusions about whether the exams are as easy as has popularly been claimed. The two instances in the past few years, reported in the *Times* and the *Guardian*, both used this method, but differed slightly in their approach.

Journalists at the *Times* were given a choice of subjects, and were then allowed to pick which A level paper they sat within that subject. In contrast, the journalist at the *Guardian*, whilst given a free choice of

subject, sat all the papers in order to take the full A level. Unsurprisingly, most of the journalists picked subjects that they were familiar with – either subjects that they had studied at university, or subjects that were related to their current jobs. Sometimes the links between subject and job were obvious, such as the Berlin correspondent choosing German or the political correspondent choosing politics. Other journalists were slightly more adventurous in their choice: the linguist who chose English Literature having always wanted to study it at A level or the writer who chose Critical Thinking as it was something that should relate to journalism. Inevitably, subject choice affects the results of the studies. A levels are not intended for those who already have a degree in the subject; therefore (as pointed out at the end of the *Times* article) you would expect the journalists to do well in the subject areas that they worked in. Even the linguist did not have experience representative of an 18 year old's, as she had clearly encountered literary study as part of her university degree.

Another feature of these studies is often the short amount of time that is given to prepare for and sit the examinations. The journalist at the *Guardian* took A level English (including AS) in one year, whilst the *Times* journalists appeared only to have been given a few days. The argument seems to be that if an A level can be studied successfully in such a short period of time, then it cannot be worthwhile, or it must have been devalued. These arguments ignore the experience which journalists bring to their examinations. They talk of exam techniques such as "quoting authorities", timing, and choosing "bluffable subjects", in addition to their already considerable writing skills. "Making a plausible argument is something that I have been paid to do for the past 25 years" (journalist taking critical thinking paper – Mary Ann Sieghart). This makes their experience rather different to that of the average 18 year old. Nor do they have to cope with the difficulty of learning three, possibly four, new subjects at once.

Perhaps the most useful output of these studies is the insight that they give journalists into the experiences of those taking the exams. None of them reported that they found the experience easy, with several of them experiencing the same nervousness that they had when taking A levels in the past. Whilst some questioned whether they deserved their grades, none of them concluded that A levels are easier now. In fact one started by saying, "The one I sat was as demanding as any I tackled in the mid-1960s". The power of these studies is that they make the general public realise that gaining good marks on an A level paper is not as easy as the press sometimes claims that it is.

Discussions of/commentary on historic examinations by current students

Exam results and standards are frequently discussed in the media. Every year, on results day, there is the predictable round of stories of higher results than ever, taken to imply 'dumbing down' of the A level system:

- *GCSEs hit new high as experts criticise tests* (*Daily Telegraph*, 2010)
- *So easy a five-year-old has passed and a seven-year-old got an A star: Record numbers achieve GCSE top marks* (*Daily Mail*, 2010)

These headlines represent a common misunderstanding of the relationship between results and standards. Whilst the percentage of students achieving particular grades may have increased, this does not automatically imply that standards have fallen, or that the exams have got easier. All the students achieving a grade will have met the standard required for it.

Often, as in the examples above, news stories are based on results alone, but occasionally discussion is informed by the inclusion of additional evidence. One such discussion was aired on BBC Radio 4's 'Today' programme in 2009, where the discussion included extracts of pupils comparing examination papers from the 1930s with today's papers.

This particular method is useful as it uses pupils to make the comparisons. As they are of the relevant age, they are arguably better able to make a judgement about how difficult they would have found the papers. These pupils have not studied beyond the level expected in the papers, so do not have the issues of adults' additional knowledge and skills which could make papers seem easier.

In this instance the discussion of papers was directed towards similarities and differences in the papers, the skills required by the papers, and the purposes of them. The pupils identified differences in skills such as the need to learn the text in the 1930's paper versus needing the skills to analyse it in the papers today; however, they found it difficult to agree on the difficulty of the papers. Some pupils thought the memorisation required would make the 1930's paper easy, whilst others thought that would make it much harder.

A drawback of this sort of study is that only small extracts from the discussion were reported in the programme. That makes it difficult to know whether the extracts were representative of the discussion as a whole, or whether they were chosen to illustrate particular points that the editors wanted to highlight. Whilst this programme concluded that some pupils preferred the old exam, it did not say what proportion of the pupils this represented, nor did it go into detail about their reasons for this preference. In addition, as the discussion only took place in one classroom, it is impossible to generalise that the pupils' experiences of the papers would be the same for pupils in all schools. These small discussions are not able to produce firm conclusions about the difficulty of papers, but they are useful in drawing attention to the differences in style and purpose of the papers for the general public.

O level papers

Comparisons using past exam papers are not limited to radio programmes. In 2008 'The O Level Book' (Anon, 2008) was published, containing a collection of past O level papers from 1955 to 1959. Readers were challenged to attempt the papers in a variety of subjects and compare their answers to those provided by experts in the subject. The book formed the basis for an article in the *Times* (Griffiths, 2008).

The book is useful for such comparisons as it contained complete papers, rather than selections of one or two questions. However, closer inspection reveals that the so-called 'complete papers' were actually a collection of questions taken from different years. Whilst they probably retained the structure of the original papers, they may not have been representative of the real challenge. The foreword and editor's notes made reference to the challenge of the O levels, describing them as a "…stinkingly hard, fact-based exam…", and the questions as "…doable, if tough." This suggests that there may have been deliberate selection of difficult questions.

There was an attempt to account for the differences between O levels and today's exams. In the foreword, an interesting comment was made

that at O level it was the number of subjects that mattered, not the grades achieved, which was contrasted with the situation today. The editor's note also commented on changes to the context and content of the exams, drawing attention to changes in teaching and examining, as well as the more obvious changes in content for subjects such as science and history. These are important observations in the context of comparing standards, as all these things will affect the experience of students sitting the exam.

In the *Times* article by Griffiths mentioned above, two of the examination papers in the book were used to test the claim that exams have been dumbed down. Five GCSE pupils sat English and Mathematics O level papers taken from the book in examination conditions just after they had completed their GCSE exams. The teenagers quoted in the article seemed to suggest that the O level papers in mathematics were more challenging, and this was backed up by their results. All of them were predicted Bs and above in their GCSEs, yet only two pupils achieved a pass mark in the Mathematics exam. Their reactions to the difficulty of the English papers was mixed, but none obtained the highest grade in English, despite several of them being predicted As and A*s at GCSE.

These results might seem to confirm that GCSEs are easier than O levels, but there are other factors influencing the results. Whilst the students had studied the subjects recently, they had not received any teaching or preparation for the O level papers. The style of the questions was not the same as a GCSE's, nor were the tasks required of them identical. In English they were asked to summarise passages and explain the meanings of different parts of speech (pronoun, conjunction). Whilst the students did not comment on the content of the mathematics exam, they did mention that they could not use calculators. Their deputy headmistress, who was quoted in the article, acknowledged these differences and several other contributing factors, but nevertheless concluded that the O level papers were harder.

The book of O level questions provides an interesting resource for comparison, but in our opinion there are too many varying factors of unknown effect for a conclusion to be drawn about the relative difficulty of exams today and in the 1950s.

Discussion

There are two key strengths to the type of study discussed in this article. First, they are often able to reach a much larger and broader audience than academic papers can. Secondly, they encourage debate in these areas, which is important. However, these sorts of studies also have weaknesses – the most crucial of which is that, just like academic studies of long-term standards over time, it is not possible to control for all the changes in social context. With the exception of the RSC study, they rely on 'case study' approaches: these have advantages in extracting a rich description of the issues, but are a shaky foundation on which to make generalised statements about national exam standards.

The academic community needs to find a better way of reaching more people, and a way of describing comparability in a clear way. Some of the studies described in this paper are useful in illuminating issues which might be overlooked in more academic research. For example. the depiction of teaching methods from the 1950s in "That'll Teach 'em", brought alive the differences in context in a way that would be difficult to achieve in an academic paper. On the basis of these studies, attracting a wider audience seems to rely on the use of a broad range of media, and on a simplification of the issues. The former is likely to be more readily accepted by the academic community than the latter.

"That'll Teach 'em", probably helped public perceptions because it illustrated the issue of contextualisation in a dramatic way. Readers of the pieces by journalists who re-sat qualifications also gained a greater insight into the complications of the issue and of the fact that long-term comparisons of standards over time are not straightforward, either to conceptualise or to interpret.

One issue, particularly evident in the RSC study, is the relationship of newspaper headlines to the outcomes of a study. There is clearly a tension between a headline accurately representing the content of a report in a handful of words and the need for a headline to sell a story. It can be the case that a headline is far removed from the data on which it is based, and in these cases there is a danger that the benefits from gaining readership are then lost in the misrepresentation of the research.

Comparing standards over time is beset by limitations – the effects of changes in technology, social expectations, culture, educational priorities and knowledge all have to be taken into account when making these types of comparison. The five examples used in this article show that clearly, as do studies from elsewhere within the educational research community. If there is a common theme to be found running through all the studies it is that there is more to standards of time research than at first meets the eye.

These five examples highlight how the 'standards over time' debate has been taken up by television, professional associations, newspaper and radio. It can be argued that the motivation for these studies has moved from the 'contribution to knowledge' of academic research towards viewing and listening figures, newspaper sales, and government lobbying. All acknowledge the complexities of the standards issue, but perhaps rely too much on over-simplification in order to reach a wider audience.

References

Anon (2008). *The O level book: genuine questions from yesteryear*. St Ives: Michael O'Mara Books Ltd.

Griffiths, S. (2008). GCSEs v O-levels, let the battle begin. *The Times* , 6 July [online] http://www.timesonline.co.uk/tol/news/uk/education/article4275054.ece (Accessed 25 May 2011)

Groskop, V. (2010). Why I took an A-level at 37. *The Guardian*, 17 August [online] http://www.guardian.co.uk/global/2010/aug/17/took-a-level-at-37?INTCMP=SRCH (Accessed 24 May 2011)

Lebus, S. (2009). Are A levels getting easier? Speech made to the Global Student Education Forum (GSEF). Wednesday 20th May 2009. http://www.dailymotion.com/video/xghrr7_gsef-talk-q-a-simon-lebus_news, accessed on 12.2.

Riddell, P., Sieghart, M. A., Boyes, R., & Wighton, K. (2008). A levels: put to the test. *The Times*, 14 August, [online] http://www.thetimes.co.uk/tto/life/families/article1758015.ece (Accessed 24 May 2011)

Royal Society of Chemistry (2008). The Five-Decade Challenge. Research report. Retrieved from http://www.rsc.org/images/ExamReport_tcm18-139067.pdf on 19th April 2011.

The Daily Mail (2008). The proof that science exams are easier now. *Daily Mail*, 27 November 2008, p.12.

The Daily Mail (2010). So easy a five-year-old has passed and a seven-year-old got an A star: Record numbers achieve GCSE top marks. 25 August, [online] http://www.dailymail.co.uk/news/article-1305746/GCSE-results-2010-Students-future-uncertain-university-rejects-snap-college-places.html# (Accessed 25 May 2011)

The Daily Telegraph (2008). Dumbed-down science is 'failing a generation and will end in catastrophe' Experts warn of 'catastrophic' science teaching. *Daily Telegraph*, 27 November 2008, p.16.

The Daily Telegraph (2010). GCSEs hit new high as experts criticise tests. 23 August, [online] http://www.telegraph.co.uk/education/educationnews/7958673/GCSEs-hit-new-high-as-experts-criticise-tests.html (Accessed 25 May 2011)

The Guardian (2008). School science standards are slipping, says study. *Guardian*, 27 November 2008, p.14.

The Independent (2008). 0% What this year's top science pupils would have got in 1965. GCSE students flunk past papers in experiment that exposes decline in standards. *Independent*, 27 November 2008, p.45.

The Today Programme (2009). BBC Radio 4. Thursday 31st April.

Cambridge Assessment
1 Hills Road
Cambridge
CB1 2EU

Tel: 01223 553854
Fax: 01223 552700
Email: ResearchProgrammes@cambridgeassessment.org.uk
http://www.cambridgeassessment.org.uk

ISSN: 1755–6031