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)

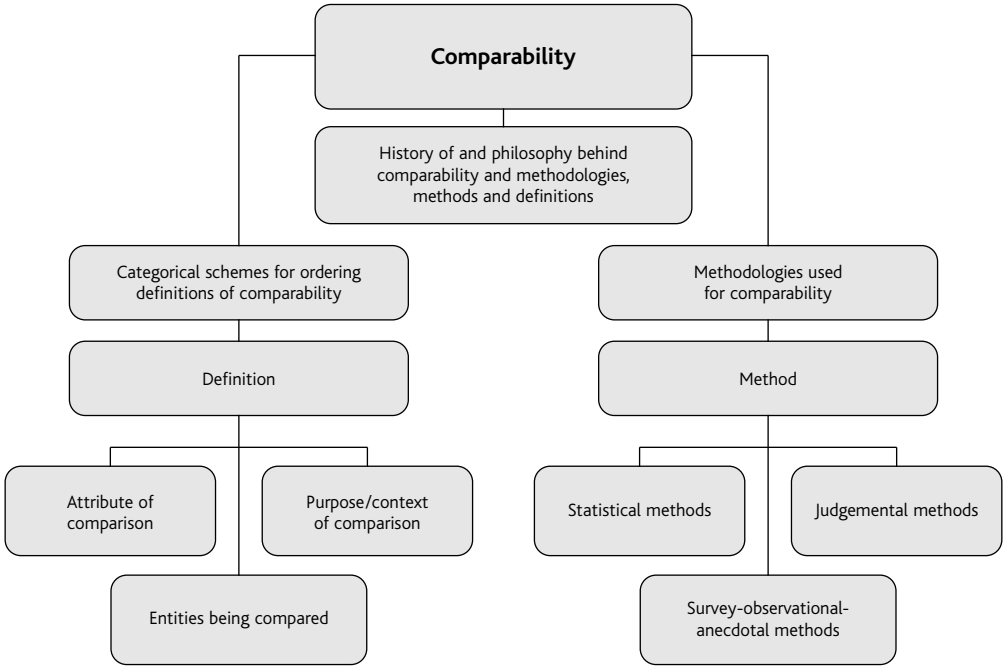

Figure 2: A framework for understanding the arguments in the literature

> *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.