A Review of Research in Levels, Profiles and Comparability

John F. Bell

and

Jackie Greatorex

Research and Evaluation Division
University of Cambridge Local Examinations Syndicate
1 Hills Road
Cambridge
CB1 2EU

5 May, 2000

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

Contents

1. Introduction
2. Review of research into qualification Levels
3. Review of research into describing profiles of contents and demands of individual qualifications
4. Review of research undertaken comparing qualifications
5. Summary and Conclusions

Appendix A: List of Comparability studies involving GCE, GCSE and GNVQ Qualifications
Appendix B: Reasons for using logistic and ordinal regression
Glossary
Bibliography

Introduction

QCA has a responsibility for creating a 'clear, coherent and well regulated framework of national qualifications' and has to be able to demonstrate where equivalence, similarity and difference exist between qualifications. QCA has formed a Qualifications Comparison Group to advise on this issue, and has commissioned this research review in order to identify relevant research findings and methodologies. The objectives of this research review are to provide information on ways in which qualifications across the framework can be compared and to provide evidence that will assist QCA and the Qualifications Comparison Group in the prioritising, planning and undertaking of future work in this area.

The objective of this report is to address the following research questions:

1. What research has been carried out investigating and/or comparing qualification levels? (e.g. what makes a 'level 3' across different qualification types?)
2. What work has been carried out on providing descriptive, and possibly comparative, profiles of the content and/or demands of individual qualifications?
3. What research has been undertaken comparing qualifications, including comparisons both of similar types of qualifications and those, which are dissimilar? (e.g. between different A level subjects, between A levels and GNVQs, or between A levels and NVQs)
4. In what ways can research be categorised so that it is possible to gain a clear overview of project size, methodology, duration and focus?
5. What methodologies have been used for qualifications comparison work? This will require critical commentary upon their strengths and limitations, and some evaluation of their potential for use for particular qualification types.

Concerns about comparability have a long history (Christie & Forrest, 1981). For example, in 1977, they noted that the Expenditure Committee of the House of Commons in its Tenth Report recommended 'that examining boards strive to ensure that standards be kept as similar as possible:

(i) between the various boards

(ii) between the various subjects

(iii) from one year to the next

(iv) between the various Mode II and Mode III schemes.'

(House of Commons Expenditure Committee, 1977, para. 116).

In order to understand how to evaluate comparability, it is necessary to consider the definition of comparability. Comparability is context related and different study designs are appropriate for different contexts and different purposes. For example, when there are different types of syllabus for the same subject and the same type of examination, then the following very strict definition of comparability can be used: Two examinations are comparable if pupils who demonstrate the same level of attainment obtain the same grade. This definition is too strict for other comparative purposes. These are based on fitness for purpose, which considers how well qualifications serve as preparation for particular pathways in later stages of learning. The purposes of the qualifications, which make up the national framework, are often different but in some circumstances it may be necessary to compare them. For example, NVQs are work-based qualifications for specific occupations that recognise what a person can do – there is little formal training. GNVQs are school- or college-based qualifications for broader vocational areas that involve a formal set of modules to attend. This does not mean that individuals with these different types of qualifications will not follow the same pathway later.

To consider the issue of the comparability of qualifications, it is necessary to understand how qualifications are used and how this use affects the definition of comparability. The main purposes of qualifications are to either decide on access to education and training or decide whether a person should be employed. A qualification can supply three types of information about a person:

(1) *Knowledge (which will be taken to include subject specific skills)*

(2) *Skills (i.e. transferable skills which can be acquired by studying one subject but can be applied to other subjects)*

(3) *Potential or aptitude (i.e. the ability to acquire new knowledge and skills).*

The importance of each of these types will depend on the circumstances. For example, if the objective is to recruit someone who can translate Spanish, then only a qualification in Spanish would be satisfactory and qualification in French would not. If, however, the objective was to identify someone with the skill and the potential to learn Italian, then qualifications in either Spanish or French could be considered comparable. In one context, the qualifications are comparable and in another they are not. This idea is particular relevant for vocational qualifications because they have 'an external referent derived from the world of employment' (Wolf, 1996). For example, a judgement of the validity of an accountant's final examinations will not be based on the syllabus that led to examination or test (as might be the case for many academic qualifications) but the behaviour and practice that follow it. However, a view of vocational qualifications that is solely limited to the particular career pathway can be damaging because it fails to consider the candidate's potential to follow a different career path. A qualified accountant may decide to change course and pursue an alternative career. This would need a broader consideration of the knowledge, skills and potential associated with the accountancy qualification. When strict comparability is not required; the objective of research is to identify the similarities and differences between different qualifications so that guidance can be given on their interpretation.

For a given set of circumstances, two qualifications can be comparable if they either indicate the same degree of knowledge, skills or potential, or the same degree of skills and potential, or if they indicate the same potential. When two qualifications are deemed to be comparable for a particular purpose they are described as equivalent. For example, at Sheffield University the undergraduate admissions office uses the following guidelines. Equivalents for the A-level grade B are given as an A/B on the Scottish Higher and a score of 6 on the International Baccalaureate. In psychometric theory, equivalence has a different definition. This involves the process of test equating. The purpose of equating is to obtain an effective equivalence between raw scores on two test forms. Lord (1980) argued that the scores on two tests, X and Y, are equated if the following four conditions are met:

1. Same ability – the two tests must both be measures of the same characteristic (latent trait, ability, or skill).
2. Equity – for every group of examinees of identical ability, the conditional frequency distribution of scores on test Y after transformation, is the same as the conditional frequency distribution of scores on test X.
3. Population invariance – the transformation is the same regardless of the group from which it is derived.
4. Symmetry – the transformation is invertible, that is, the mapping of scores from form X to form Y is the same as the mapping of scores from form Y to form X.

If all these conditions are met then the test forms are equivalent. Whilst the psychometric methods of test equating are too restrictive in application to address all aspects of comparability, the theory of equating methods is reviewed in the next section because there is much that is relevant to more general issues of comparability.

In addition to knowledge, skills and aptitude, there are also other criteria for comparing qualifications. Two important ones are economic value and parity of esteem. The economic value of a qualification for an individual is the additional income that can be expected for having the qualifications. Obviously this depends on the supply and demand for that qualification at a given moment in time. This means that it is not a sound basis for comparing qualifications. Parity of esteem relates to the perceived status of the qualifications. Although perceived status is an important issue, it may not be founded on actual evidence and is also not a sound basis for comparing qualifications. The most obvious example is this relates comparison of academic qualifications with vocational qualifications because the latter, apart from a few high-status professions, have been associated with lower social strata which was only offered that type of education (Max Planck Institute of Human Development, 1983; Wolf, 1996).

To investigate comparability, it is necessary to consider different situations. Three factors can be considered: the subject of the qualification, the level of the qualification, and the qualification system. Pairs of qualifications can be categorised in three ways by subject.

Firstly, they can be the same subject or subject area. Although two qualifications may be in the same subject area, it does not necessarily follow that exactly the same knowledge and skills will be covered. For example, in a subject such as History, the historical knowledge could come from a wide variety of periods and countries, and comparability is associated with the interpretative skills required by the subject.

Secondly they can be similar subjects, e.g. different modern foreign languages. For similar subjects, the knowledge and subject specific skills are different but are related. In modern foreign languages, the expected fluency in speaking for a particular grade can be made comparable e.g. all modern language qualifications in a system could require candidates to participate in the same kind of conversation. There are also situations when subjects have overlapping knowledge (e.g. accounting and business studies). For these situations comparability of the overlap can be investigated.

Finally, they can be completely different subjects. In this case, comparability relates to identifying the potential and the transferable skills associated with each qualification. In reality, the similarity between subjects can be conceptualised as a continuum but the classification of scenarios for comparability given above simplifies the interpretation.

The next factor is the level of the qualification, e.g. level 3 NVQ, undergraduate. The concept of level is complex but underlying it is the assumption that levels form a hierarchy. Firstly, there is the issue of deciding on what is appropriate at a particular level. Secondly, when comparing qualifications at different levels, the objective is to determine whether the attainment at the higher level guarantees performance at the lower level. This particular point is important when it is possible to obtain higher level qualifications without first obtaining lower level qualifications. Thirdly, levels for different frameworks may need to be compared to be deemed to be equivalent, e.g. CATS and NVQ levels.

Finally, there is the qualification system, e.g. A-levels, Scottish Highers or the International Baccalaureate. If the system is the same, then research is often concerned with maintaining comparability. In the case of different systems, the objectives and expectations may differ. In addition, the grading systems and standards will also be different. In this situation the objective of research into comparability is often to provide a description of similarities and differences.

Permutations of these three factors, subject, level and system, can be used to identify different types of comparability. In Table 1, the relationships between the permutations of these factors and the three types of information are presented. In all cases, it is possible to compare to a greater or lesser extent the potential of candidates. All the permutations have been included for completeness. Some of the types are not likely to be considered very often in practice.

Table 1: Examples of different types of comparability

Type	Subject	Level	Examination System	Knowledge	Skills	Potential
I	Same	Same	Same	Yes	Yes	Yes
II	Same	Same	Different	Yes	Yes	Yes
III	Same	Different	Same	Some	Some	Yes
IV	Same	Different	Different	Some	Some	Yes
V	Similar	Same	Same	Some	Yes	Yes
VI	Similar	Same	Different	Some	Yes	Yes
VII	Similar	Different	Same	Some	Some	Yes
VIII	Similar	Different	Different	Some	Yes	Yes
IX	Different	Same	Same	No	Sometimes	Yes
X	Different	Same	Different	No	Yes	Yes
XI	Different	Different	Same	No	Some	Yes
XII	Different	Different	Different	No	Sometimes	Yes

For comparability types I and II, the requirements for comparability can be required to be strict considering all three factors. Although it should be recognised that sometimes with different syllabuses and, more often with different systems the knowledge required need not be the same. It should also be noted that for some comparative purposes, interest will be focussed only on skills and/or potential.

For comparability types III and IV, comparability involves investigating whether the higher level qualification guarantees that an individual has the knowledge, skills and potential required at the lower level in situation when it is possible to obtain the higher level qualification without the lower one. The qualifications could be considered to be comparable, if only the lower level skills are required for a particular purpose.

Comparability types V-VIII relate to comparing similar subjects. Examples of these include modern foreign languages and different forms of Design and Technology. In some circumstances, there may be some overlap in terms of knowledge but this can be variable. It is assumed that the more general transferable skills are the same or very similar. An issue that can occur with similar subjects is that although the knowledge may differ, there is a concern that the degree of knowledge is the same. For these types of comparability, the objective will usually involve identifying levels of comparable skills.

Comparability types IX-XII relate to comparing different subjects. Type IX comparability, different subjects at the same level and the same system, has been a topic for debate (Fitz-Gibbon & Vincent, 1994; Goldstein & Cresswell, 1996; Newton, 1997; TaylorFitz-Gibbon & Vincent, 1997). It is sometimes described as subject difficulty. Difficulty is not a simple concept and can be thought of as having three aspects relating to perception, effort and demand. Perceived difficulty describes the attitude of candidates towards the subject and is not necessarily related to the other aspects of difficulty. There has been a great deal of research in this area (e.g., Clarke & Youngman, 1987; Watson, McEwen, & Dawson, 1994; Harvard, 1996; Robinson & Tayler, 1992). Secondly, there is the issue of difficulty regarding the effort required to obtain a qualification, e.g. having to provide more evidence for one qualification compared with another. This has important implications for the construction of syllabuses. Finally, there is difficulty relating to the demand of qualifications. Goldstein and Cresswell (1996) suggested that comparing grades from different subjects assumes that 'each syllabus and associated examination develops and assesses the same attribute'. They noted that this assumption is not true and argued that there is no sensible definition of subject difficulty and no sound basis for making adjustments. However, this issue of difficulty between different subjects must be considered because some qualifications have long been used as measures of aptitude. For example, Christie and Forrest (1981) noted that

> '...Selectors tend to use GCE and CSE examination results as indicative of aptitude rather than as required achievements and there is no reason to suppose that the results of the common system of examining at 16+ will receive a different treatment. They will still tend to ask for "five Ordinary-level passes" rather than "the successful completion of a five year secondary course in Chemistry." The notion of five unspecified passes depends upon a constant relationship being maintained, subject by subject, between aptitude and achievement...'

It is not only in the extreme case of an indicator of aptitude that two different subjects can be compared. Two subjects could both measure the same set of attributes but expect different levels of performance on each attribute. If the levels were higher for all attributes for one subject compared with another then it would be said that that subject was more difficult. Of course, since decisions about comparability are context based, for some purposes subjects will be comparable when the level expected of a small subset of attributes is the same (in the most extreme case, potential to do a particular task). This means, however, that lack of comparability, or subject difficulty is context dependent. One result of this is that some subjects are only partially comparable. In these cases, one qualification can be used to gain an exemption from part of another qualification. For example, the Credit Accumulation and Transfer Scheme (CATS) in higher education enables prospective candidates to claim academic credit for other qualifications (e.g. NVQs). To establish comparability a technique called mapping was used.

The remainder of this report consists of the following. Firstly, the research into qualification levels is reviewed. The issue of level frameworks for different types of qualifications is considered. A review of international frameworks has also been included in this section. This is followed by a review of research into describing profiles of contents and demands of individual qualifications. When a decision is made about an individual, the decision is not usually based on one qualification but on a range of qualifications that have been acquired as the individual has progressed through the educational system. When considering a range of qualifications, it is useful to consider the profiles of content and demands for each qualification. The work that has been carried out on providing descriptive, and comparative, profiles of the content and/or demands of individual qualifications is reviewed. The fourth chapter considers research undertaken in comparing qualifications is reviewed. This section consists of a literature review of the research into the comparability of qualifications. The research is evaluated in relation to the definitions of comparability.

Review of research into qualification levels

In this chapter, the issue of level frameworks for different types of qualifications is considered. In the introduction, the level of a qualification was identified as a factor that could be used to categorise qualifications. Different frameworks have often been developed for different types of qualifications, e.g., academic and vocational. Despite the differences in purposes of qualifications there is some work which compares level descriptors from different qualification frameworks. There are also some international frameworks that attempt to unify a wide variety of qualifications into a single framework. These are also reviewed in this chapter.

There has been a great deal of research into this issue in higher education. This literature review covers two main points:

- How educators view levels;
- Methods of finding an agreement between educators on descriptions of levels.

To understand these points it is necessary to consider the structure of qualification frameworks. They map the relationship between different qualifications and are often divided into a hierarchy of levels. For example, the NVQ framework has five levels. QCA (1999) have recently outlined the new proposals for the post 16 curriculum in the document 'Curriculum Guidance for 2000 – implementing the changes to 16-19 qualifications'. These proposals are designed to make the post 16 curriculum more flexible so that students can combine both academic and vocational qualifications. To achieve this greater flexibility qualifications are treated as building blocks that can be combined to form diverse types of programme that will increase opportunities. The terminology that will be used here is changing; A levels will be known as Advanced GCEs and GNVQs will be called vocational A levels. GNVQs have been slimmed down so that students can take the newer 6 and 3 unit awards or the older 12 unit awards. The 6 unit award is called the single award. The 6 unit block is comparable in size to an A-level. With the qualifications as similar size blocks, they can be mixed and matched so that students can study both vocational and academic qualifications together – bridging the academic/vocational divide. Other changes in the new curriculum from 2000 are the introduction of a Key Skills certificate (Key Skills are signposted within the GNVQ specifications) and Advanced Extension Awards (AEA). These later awards will be offered in 13 traditional academic subject areas, to replace S level qualifications and university admissions tests. They have been mapped against S level papers. Offering these awards has implications for stratification of the advanced level sector (Hodgson and Spours, 2000).

It is possible that although the advanced single award GNVQs will be awarded at the same level as A levels and that they are deemed to have parity of esteem with A levels, they may be associated with lower achieving students in the post 16 sector. Methods of comparing qualifications have been discussed in the previous chapter. In this chapter comparability will be addressed by considering the concept of level. A method of indicating whether, say, GNVQ and A level awards are comparable is to define what is required in general at a particular level and then justify awarding particular qualifications at particular levels.

This method has been considered in HE when credit frameworks have been introduced (Robertson, 1994). As changes are being made to qualifications and the curriculum for the post 16 sector educators are again considering the issue of what distinguishes one level from another. There is little literature about this issue for the post 16 sector, much of the research and debate has been confined to HE. This literature will be reviewed because the lessons and experience from the HE debate are relevant to the development of the changes in the post 16 sector.

Levels in Higher Education: Background and History

In Higher Education, qualifications are awarded at a particular level or students are awarded credit at particular levels that they accumulate to make a qualification (Higher Education Quality Council - HEQC, 1997). Until recently the meaning of these levels remained the tacit understanding of specialist communities and students. Course design and choice of subject matter indicated hints about the meaning of each level. HEQC (1997) found that there were different concepts of level: -

- A measure of intellectual demand or difficulty;
- A measure of progression through a curriculum or syllabus;
- A discriminator in the grading of academic performance.

In this chapter, level refers to qualification levels. Quality Assurance Agency - QAA (1999a) proposed that there should be 5 higher education levels associated with the most well know qualifications: -

1) HE Certificates;
2) HE Diplomas;
3) Bachelors degrees with honours;
4) Masters;
5) Doctorates.

Credit and qualification frameworks in HE tend to be synonymous (QAA, 1999c). Credit and qualification levels are usually associated with the first two concepts of levels. Greatorex (1998) reviewed the issue of the notion of levels in higher education and much of this section is an updated and revised extract from this source.

The notion of levels was introduced by the Council for National Academic Awards - CNAA. The meaning the CNAA (1989, 7) gave to each level was as follows:

'Level 1: the standard of a course unit in the first year of a full-time three-year Degree or Honours Degree

Level 2: a course unit will be assigned to level 2 if it is of the standard normally encountered in the second year of a full-time three year Degree or Honours Degree

Level 3: corresponds to the standard of course units normally encountered in the final year of a three-year full-time Degree course.

Level M: Course units in approved postgraduate programmes of study attract this rating even if they contain material from first Degree work for students lacking this background.'

The purpose of the CNAA was to formally establish and guard academic standards and oversee the quality of the polytechnics and colleges in the higher education system. The CNAA were invited by the government to encourage the development of the Credit Accumulation and Transfer Scheme (CATS) in institutions of higher education, the objective being to facilitate the transfer of credits from one institution to another at any stage in a student's study career. According to Robertson (1996) for this scheme to be feasible there needed to be confidence between institutions that the standards in different institutions in higher education were similar. He subsequently argued that the assumption that different institutions had the same standards was inherited from the CNAA. This need for confidence in national standards heightened the debate about the undergraduate levels 1, 2, 3 and postgraduate level M, because the CNAA (1989) viewed these levels as incorporating standards. For this assumption to be valid there needed to be:

- a common view of what levels were;
- an agreement on the attributes required at each level;
- an application of the same standards at the same level in different institutions and disciplines.

CNAA (1989), South East England Consortium for Credit Accumulation and Transfer - SEEC (1996), Inter Consortium Credit Agreement Project - InCCA (1998) and QAA (1999b) argued that the credit framework constituted a series of hierarchical levels and assumed that student progression through these levels was linear. QAA cite wide support within higher education in the UK:

'... for the development of national qualifications frameworks structured in terms of a series of levels (indicating progressively greater intellectual challenge...)' (QAA, 1999b, 3).

InCCA (1998) also defined the difference between levels as being the different intellectual challenge associated with each level. Of course intellectual challenge is not the only purpose of higher education. QAA (1999a) have modified this assumption and argue that there is a progression within as well as between levels. However, this assumption may not hold. Hirst and Peters (1970) and Reeves (1988) argued that the traditional hierarchical organisation of disciplinary knowledge mistakenly led some educators to believe that there was a particular order in which knowledge needed to be learnt. This belief gave legitimacy to the notion of learning as a hierarchical progression through disciplinary canons. These traditions were challenged by the Robbins Report (1963), which said that there was a need to change the type of education and training that was provided by HE.

In the 1980s and 1990s, political commitment created a 'new' higher education that offered more student places than the traditional élitist system. There were also attempts to make higher education more like a business. Knowledge became a marketable commodity and academics were accountable to their customers (students and employers). Accountability included making the commodity of knowledge explicit by writing level descriptors and learning outcomes. Level descriptors are a description of the qualities associated with a qualification or credit level. Generic descriptors or generic level descriptors are descriptors that apply to more than one subject or programme. Learning outcomes were a statement of the learning to be achieved. In particular, outcomes in professional and vocational training tended to be made explicit and there was an increase in professional and vocational courses in higher education (for example, the advent of midwifery degrees). Barnett (1994) believed that these changes reflected how society was searching for new ways of viewing knowledge, which incorporated knowledge and skills. He argued that the emerging discourses did not give further insights into knowledge and skills. Instead, they reflected a power struggle between the State, academics and employers.

The market principle and the introduction of vocationalism led to educators trying to define the attributes of graduates. Otter (1992) conducted a project to identify what graduates could do. With reference to levels she concluded that educators knew what was meant by graduate level, but that they did not have the language to articulate it. However, four years later educators were still searching for a description of the attributes shared by all graduates, for example, HEQC (1997) and Bell (1996). The Graduate Standards Project (HEQC, 1997) attempted to describe the attributes shared by all graduates. These attributes became known as 'graduateness'. There were other authors such as Winter (1993, 1994) who considered the differences between levels. The results of such investigations are discussed below.

The development of qualification and credit frameworks has not been confined to the UK. Other examples are found in New Zealand and South Africa whose qualification frameworks will be discussed. One of the reasons for developing qualification frameworks is to establish comparability or parity of esteem between qualifications. However there are still some qualifications that are in different frameworks where comparability studies have been undertaken. Examples of these studies are given and the methods of comparison discussed.

Review of International frameworks

There are several international qualifications frameworks. Some of them have extensive websites that give a great deal of information about the frameworks and their regulatory bodies. It would take a substantial research project to assess these frameworks fully and this section aims to highlight some interesting features of them. Within any framework, the process that influences the allocation of qualification to a level is either the standard setting process for new qualifications or for an existing qualification the recognition of the standards that have been set. It is this aspect that this section will emphasise.

One of the most interesting is the New Zealand National Qualifications Framework (http://www.minedu.govt.nz/ngfwhitepaper/factsheet.htm). All qualifications in the framework are described consistently in terms of *learning outcomes, level, credit* and *detailed field.* These terms are defined as follows:

- Learning outcomes describe what people will know and can do when they complete a qualification.
- Levels describe the complexity of learning outcomes.
- Credits are a measure of the amount of learning and assessment required on average to gain a qualification or to complete a course.
- Detailed field is a standard set of subject classifications covering all qualifications and courses.

Further details can be found in the white paper at the following website: http://wwww.minedu.govt.nz/nqfwhitepaper/whitepaper.pdf. Having qualifications based on learning outcomes makes this framework different from systems that focus on outputs such as courses or inputs such as curricula or teaching hours. International bodies involved in funding education systems (e.g. the World Bank, Asian Development Bank and the OECD) have endorsed outcome models.

One of the suggested advantages of this framework is that when qualifications are registered people will be able to make meaningful comparisons between similar qualifications, which will enable them to make better decisions about which educational options to pursue. All qualifications currently registered on the framework are composed of registered unit standards. The statements relating to these standards describe what a learner knows or can do. Because they are nationally agreed, learners' achievements can be recognised in a number of different contexts and their knowledge and skills will be transferable between qualifications and providers.

A process of consensus is used to develop the framework. Initially, expert groups (engineers for engineering standards, etc.) draft the standards. Then stakeholders receive the draft standards for their comments and contributions. Once agreement is reached, the standards are registered. They are reviewed by stakeholders and experts on a regular basis.

The New Zealand framework has been criticised. For example, Peddie (1997) pointed out that difficulty was not always associated with level and the New Zealand Vice-Chancellors' Committee (1994) perceived that the framework was incompatible with the nature and aims of most university degree courses. They proposed an alternative dual structure covering university and non-university qualifications and the development of systems for dealing with movements between the two.

South Africa also has a qualifications framework managed by SAQA - South African Qualifications Authority (http://www.saqa.org.za/docs/index.html). This framework has eight levels ranging from a

General Education and Training Certificate (GETC) at level 1 to doctoral and post-doctoral research degrees at level 8. It was agreed by education and training stakeholders throughout South Africa.

SAQA has two arms: one associated with standards setting and the other with quality assurance. The standards setting is managed by twelve National Standards Bodies (NSBs) – one for each of twelve organising fields of learning. They oversee the work of SGB (Standard Generating Bodies). These bodies are created using the following process:

1. Advertise, through the media (certainly print, and possibly radio), the Government Gazette, and the SAQA web-site, the proposed establishment of an SGB and its brief, calling for nominations to serve on it.
2. Process and verify nominations and, after consulting affected constituencies, shortlist up to 25 nominations for publication in the Government Gazette.
3. Publish names of nominees for public comment.
4. Process responses and appoint SGB members (this involves consulting the affected constituencies). If funding is secured, a first meeting of the SGB is convened.

SGBs have several functions including standards generation, updating standards, and recommending criteria for the registration of assessors and moderators or moderating bodies.

Quality assurance is managed by Education and Training Quality Assurance bodies (ETQA). They may be established in a social sector, in an economic sector or in an education and training sub-system sector. ETQAs are accredited in respect of their primary focus, based upon their association with their sectors, so that its functions do not duplicate the functions of an existing ETQA. These functions include accrediting providers; promoting quality amongst providers; evaluating assessment and facilitating moderation among constituent providers; the certification of learners. They can also make recommendations relating to standards or qualifications to NSBs.

In addition, SAQA can appoint moderating bodies to ensure that assessment of learning outcomes is fair, valid and reliable across the framework.

Australia (http://www.csu.edu.au/acadman/l5m.htm) has a system with 12 levels ranging from Senior Secondary School certificate to Doctoral degree. Some examples of the distinguishing features between qualifications at different levels, which are only intended for guidance, are given. For example, one requirement for the Education certificate level IV is to "do the learning outcomes that enable an individual with the qualification to apply solutions to a defined range of unpredictable problems". For diploma (level 5), this requirement becomes "analyse and plan approaches to technical problems or management requirements", and for an Advanced Diploma (level 6) it is extended to become "analyse, diagnose, design and execute judgements across a broad range of technical or management functions."

There is an implementation handbook for this framework (AQF, 1998), which provides detailed guidelines defining each qualification to assist course development, course approval and assessment. The guidelines specify the typical learning outcomes, which help in defining the level of the qualification. It also includes a set of articulation principles. Articulation is the name given to a process of establishing connections between qualifications in the framework. This means that linkages between qualifications can be formally endorsed as part of the approval process.

In the European Union the Directorate General Education and Culture (http://europa.eu.int/comm/education/prog.htm) has a number of programmes that address issues relating to comparability. The SOCRATES programmes (European action programme for co-operation in the field of education) include EURYDICE (the information network on Education in Europe). This programme is responsible for a number of comparative studies on a wide variety of educational issues. The publications are listed at http://www.eurydice.org/Publication_List/En/FrameSet.html. These publications include a European Glossary on Education: Examinations, Qualifications and Titles (available from http://www.eurydice.org/documents/glossary/EN/Frameset.html). This document is a descriptive account that identifies the level of a qualification within each national system.

The second relevant programme is LEONARDO DA VINCI (the European action programme for co-operation in the field of vocational training) which is responsible for CEDEFOP (the European Centre for the Development of Vocational Training). CEDEFOP work includes projects on:

- Key Qualifications and curricular renewal of vocational education and training;
- Identification, assessment and recognition of non-formal learning;
- Transparency of qualifications.

CEDEFOP aims to be a point of reference providing policy-makers and practitioners, at all levels in the EU, with information to promote a clearer understanding of developments in vocational education and training, and so enable them to take informed decisions for future action.

What distinguishes one level from another?

This section reviews the studies and discussions that have offered insights into the distinguishing characteristics of each level. Many of the descriptions of what distinguishes one level from another can be found in level descriptors. A critique of how theories like Bloom's taxonomy have been used as the basis for descriptors is also given.

Robertson (1994) argued that there was little to be gained from developing level descriptors. Despite his reservations, he suggested that work should continue to find level descriptors, which were important if agreements were to be made between higher education and NCVQ on the comparability of higher levels. But his main aversion to descriptors was that:

'It may need to be accepted that *levels of achievement*[1] are indeed arbitrary conventions; they may be rendered less arbitrary wherever possible, but to search for the ultimate precision and fairness might cause

[1] Robertson's (1994) use of a change of font to emphasise the term 'levels of achievement' has been followed in this quotation.

the entire edifice of qualifications and progression to unravel in a fruitless intellectual search for the impossible' (Robertson, 1994, 131).

Peddie (1997) also argued that levels are a convenient way of signalling the conventional sequence of learning rather than indicating differences in difficulty between different levels.

Otter (1992) argued that as the sectional boundaries between different parts of education and training were becoming more permeable it was important to understand how level was being defined across these boundaries and to come to a consensus on this. This is similar to the argument presented by Greatorex (1998). Otter (1995) described how 'The Unit for the Development of Adult Continuing Education' (UDACE) developed learning outcomes without reference to levels. The levels were considered to be stages of a course rather than levels of attainment, which is like the CNAA view of levels. Otter (1995, 281) says:

'Overall there were two views. One was that learning outcomes should be described at the graduate level. Some learning outcomes, might, depending on the construction of the course, be demonstrated at earlier stages than others. Some learning outcomes might therefore act as progression requirements at different stages. An alternative was that learning outcomes might be described at the graduate level and have incremental assessment criteria, which reflected the increasing autonomy of the student. In practice it appeared that the construction of most courses meant that some learning activities, particularly project work, were only offered at later stages, and that some learning outcomes could not actually be demonstrated until the student had completed that part of the course. The construction of the course effectively determined when the learning opportunities were provided, and therefore when some learning outcomes could be demonstrated'.

This means that the construction of courses can affect views of levels.

Davis and Burnard (1992) considered how nurses distinguished between diplomas and degrees and their relationship between nursing practice and educational levels. They argued that there is a complex relationship between the different kinds of awards. Also, there are difficulties with equating professional experience with academic study to award academic credit through schemes like the Accreditation of Prior Learning (APL). They raised questions about how different higher degrees relate to one another. For example, is a MPhil 'higher' than other master's degrees? There has also been debate about whether Scottish degrees are 'worthy' of the label 'Masters' when they are undergraduate qualifications. QAA (1999c) argued that the press have portrayed this debate as being a much bigger issue than it really is. Questions about the nature of postgraduate degrees are still not fully resolved as QAA are consulting HE practitioners about how many levels there should be in a qualification framework and what the benchmarks should be for each level (QAA, 1999c).

To illustrate the problem with levels, an example based on the classification of nursing degrees will be considered. Davis and Burnard (1992) reported that nurse educators suggested that the differences between the levels were that diplomas were introductions to topics, bachelors' degrees were more in-depth and MPhils and PhDs were specialised and in-depth research based studies. However, some degrees are more specialist than others at the same level are. Given the limitations with levels, they suggested that we should think of qualifications in terms of their fitness for purpose to the student and others. Davis and Burnard (1992) used the Delphi technique to identify what Nursing Professors considered to be the difference between levels. In their Delphi survey, they asked Nursing Professors to list five items that they thought were applicable to each level of study. The responses were content analysed. They also asked Master's Nursing students from another European country to identify key characteristics of the same academic courses. The content analysis revealed that there were four factors that were perceived to affect levels.

Table 1: Some of the characteristic differences between diploma, bachelor's, Master's and Doctoral courses listed by Davis and Burnard (1992)

	Diploma studies	Bachelor degree studies	Master's degree studies	Doctoral degree studies
Characteristic of knowledge	Broad base; discrete categories	Broad and deep; integrated across categories	Narrow and deep; specific	Very specific; generation of new knowledge; dissemination of knowledge
Nature of research studies	Introduction: critical reviewing skills	Utilisation of research findings	Competence in research methods	Advanced research skills
Generalist courses Courses move from tutor control towards negotiated and independent learning: field of study is very broad		Specialist courses Courses are characterised by more equal relationship between tutor and student: field of study is very specific		

Davis and Burnard (1992) reported that the professors took an academic approach and the Dutch master's students took a practice based approach. They suggested that rather than the simple hierarchy that is often used as a basis for educational levels, learning is actually a spiral. People go from novice to expert in each domain and at the same time they go through a cycle of learning, gaining experience which consolidates the knowledge which they have acquired. To come to their conclusion they drew upon Dewey (1938) and Kolb's (1985) learning cycles and Benner's (1984) work about nurses' expertise. They ended with the note that due to this difference between learning as portrayed by learning theories and the levels hierarchy it is difficult to establish parity between achievement through life experience and educational levels.

Winter (1993, 1994) offerered an in-depth discussion of the problems associated with 'educational levels' and described the results of some associated research. The CATS framework aims to offer credit for learning however and wherever the learning took place, thereby bridging the gap between academic and vocational learning. Winter (1993), like Greatorex (1998), mentioned that different educators in HE and FE use different concepts as the basis of level descriptors. Bloom's taxonomy is used in the Northern Ireland Credit Accumulation and Transfer Scheme (NICATS) descriptors (NICATS, 1998). Bloom's taxonomy is also used in some descriptors in nursing course documentation (James and Redfern, 1995). A continuum of autonomy has been identified in level descriptors reviewed by Winter (1993) and James and Redfern

(1995). Mager (1991) reported that the Open College Network level descriptors are based on continuums – which involve judgmental autonomy, independent learning and application. The FEU aligns academic awards, e.g., A levels with employment roles and a hierarchy of responsibility for other peoples' work (FEU, 1992). Winter (1993) argued that the latter dimension is related to Steinaker and Bell's taxonomy (1979). This experiential taxonomy is a hierarchy from 'exposure to a learning experience' to the level at which the learner takes responsibility for disseminating their knowledge. In addition, Winter (1993) argued that we are left with the problem of developing a CATS framework with levels that incorporate a plethora of judgmental hierarchies academic, cognitive, managerial and experiential. This raises the question, of whether there is a coherent theory that supports a levels framework and incorporates the many dimensions which have been associated with levels.

Swann (1992) criticised the educational levels in the English National Curriculum document as learners follow a wide variety of individual learning pathways. O'Reilly (1990) argued that hierarchies of difficulties in school mathematics are not universal but are the results of particular teaching methods. There are differences between the National Curriculum levels and qualification levels, for example, the National Curriculum level descriptors are used as an assessment tool and qualification levels are not. However, there are similarities between them. For example, they both assume linear progression. So criticisms that Swann and O'Reilly have made about National Curriculum levels are also applicable to qualification levels. Winter (1993) reviewed the theories and metaphors that have been forwarded as a basis for a hierarchical structure of levels. He concluded that the metaphors are questionable and that there are limitations to applying theories like Piaget's and Bloom's taxonomy in this context.

Greatorex (1999a) reviewed novice to expert theories which applied to a variety of higher order skills, for example, chess (Kahney, 1993), medical diagnosis (Schmidt et al., 1990), nursing (Benner, 1984) and geometry (Kahney, 1993). As the theories applied across such a variety of activities it seemed likely that the novice to expert theories would be able to account for student development. However, novice to expert theories can neglect the issue of domain specific learning and they were simplified to write level descriptors. Therefore the descriptors did not reflect the complexity of the novice to expert theories. Patel and Greon (1992) identified that novice to expert development was not necessarily linear. This does not support the notion of levels as used by CATS which assumes linear progression (Inter Consortium Credit Agreement [InCCA], 1998).

Winter (1994, 94) analysed 'the categories used to assess the work of students engaged in A level, undergraduate and higher degree studies'. He listed the 'general' rather than subject specific terms that were used to express examiners' expectations of achievement at A level. These were taken from examination boards' syllabus documentation. He found that many of the terms which were used to distinguish between performance at A level were also the words that HE tutors used to describe the differences between achievement at undergraduate and postgraduate levels. Therefore HE tutors are not defining a different stage of educational development. He undertook a content analysis of the comments that HE tutors made about students' work, e.g., essays and accreditation of prior learning portfolios, at different levels (Foundation, Honours and Postgraduate). He ignored words like 'good' and 'excellent' that have a circular logic. He noted that there were some terms that suggested a different set of expectations, which were different to tutors' expectations of students at lower levels. In summary, the distinctive feature of HE (Foundation) level work was tutors noted that students were taking a personal stance. 85% of the words that were used to describe honours level work were also used to describe A level work. Winter concluded that the characteristics which distinguished the honours degree work from the A level work were:

'merely extending the fundamental emphasis on personal involvement already expected in the foundation year data'.

This has two implications:

- There is no basis for a distinction between foundation and final year undergraduate work;
- It is not intellectual process that distinguishes, say, A level from undergraduate work; rather it is a shift in the role of the learner who develops a personal stance through their studies.

He concluded that we need two types of learning outcomes for HE level, learning outcomes about: (1) autonomous learning; and (2) personal synthesis derived from practical experience and reading.

In relation to postgraduate levels, Winter (1994) reported that tutors expect students to take on a new role. Students are expected to have a commitment to a specialism and their work is expected to have external value. The external value of the work is defined in terms of it being both scholarly (publishable) and of value to the institution in which the study is carried out, e.g., 'valuable to the Local Education Authority'. Winter (1994) refers to this external value as consultancy. These differences are very unlike the QAA definition of level, which is that they are stages of increasing intellectual challenge.

The limitation of content analysis (and Winter's conclusions) is that the same words are used in different contexts to mean different things. So content analysis may count a word twice when in each context it has a different meaning. The other limitations of content analysis are that it involves the imposition of codes by the researcher (Holsti, 1969) and can result in abstraction (Weber, 1985).

Wolf et al. (1997) compared the language that was used to describe A levels and higher levels of study. Like Winter (1994), Wolf et al. (1997) found that generic descriptors for degrees contained similar language to that which was used to describe A level attainment. Therefore they argued that generic descriptors did not provide a secure description of the achievement of undergraduates, which should be different to the achievement of A level students. They undertook this work as part of the Graduateness project. They argued that a common threshold could not be set for all graduates but that there was scope for subject based approaches. Wolf (1995) argued that it is not words that communicate the level of achievement, rather this is illustrated by examples of work.

James and Redfern (1995) qualitatively analysed a random sample of level descriptors from institutions in England who offered the Higher Award (a nursing qualification). They identified and categorised the ways in which the level descriptors varied. They found that there were a variety of ways in which Nurse educators described (and viewed) levels, for example, different educational features were used differently in different descriptors. One institute had written:

'(Level 2)...involves the study of theory and practice in greater depth'

and another:

'Level 2 courses take account of the participant's possible lack of confidence in academic study' (James and Redfern, 1995, 313).

They noted that some descriptors reflected theories like Bloom's taxonomy, novice to expert theories and Schön's reflective practitioner. Like Winter (1994), they found that some level descriptors referred to autonomy. They noted that the nature of progression is complex. There are quantitative and qualitative models of progression – the quantitative model is that each level consists of more of the same thing and the qualitative model is where higher levels are characterised by the emergence of different qualities. The bulk of the descriptors that were analysed used both models of progression. They concluded that there is a variety of models used as the basis of level descriptors, there is a diversity of views of levels, and level descriptors should be based on an epistemology of practice. The limitation of James and Redfern's (1995) methodology is that the researchers may have imposed their own interpretations and categories on the descriptors. They acknowledged that their analysis was at an exploratory stage.

Do educators agree about the characteristics associated with each level?

In this section studies which have aimed to identify whether educators can agree on the characteristics associated with each level are outlined. The limitations of the agreements and the methodologies used to identify agreement are also considered.

For CATS and the QAA frameworks to be operationalised there must be an agreement on generic level descriptors and a common view of levels. One way of establishing whether there is an agreement amongst educators on level descriptors is to use the Delphi technique (Greatorex, 1999a). The Delphi technique is a method that is used to identify whether there is an agreement between experts on a topic or problem. There are a series of stages to the Delphi study. Initially, there is an open-ended survey where experts give their views and suggest answers to the problem. The Delphi researcher collates these answers and summarises them into a second survey. In response to this survey, experts vote about views they agree with. In the third survey, the experts are again presented with the different views but they are also presented with the responses of the other participants. The experts are asked to vote again in the light of the group feedback. The third round can be repeated as a fourth round. In the third and fourth rounds experts can be asked to justify why they are voting differently to the majority of the experts, which provides an insight into why some experts do not agree with one another. The experts are considered to have reached a consensus when agreement has stabilised (Linstone and Turoff, 1975). The mean as a measure of central tendency, represents the group opinion of the panel. The standard deviation, as a measure of spread, represents the amount of agreement in the panel (Greatorex and Dexter, 1998).

Greatorex (1999a) used an adapted version of the Delphi technique to identify whether there was an agreement between healthcare educators on the appropriate levels for descriptors. To write the descriptors for the Delphi survey descriptors that were already used for healthcare courses and learning theories were consulted. The educators were presented with each descriptor three times and asked to state which level (from undergraduate levels 1 to 3 and postgraduate level M) they thought was appropriate for the descriptor. The educators agreed on the appropriate levels for some of the descriptors. It could be argued that this result was a product of the method as the Delphi technique may be construed as forcing experts into agreement. Arguing that the agreement is a product of the method is an understandable stance as Wolf (1995) pointed out that if a national curriculum attainment target is separated from examples of work, teachers cannot agree on the appropriate level. This is a very similar task to the one the experts undertook in Greatorex's survey. Although the educators could come to an agreement on the appropriate levels for some of the descriptors some of the level descriptors were so vague that they could be meaningless. This is why Greatorex and Nyatanaga (1994) advocated using generic and specific level descriptors, as the specific descriptors add meaning to the generic descriptors. Wolf et al. (1997) suggested that there was room for subject specific approaches to describing the qualities associated with graduates but that a generic approach might not be helpful. QAA (1999c) are developing subject specific level descriptors and interdisciplinary level descriptors for degrees involving interdisciplinary study. This suggests that generic descriptors such as those used for NVQs may not be useful for covering both academic and vocational qualifications in the post 16 sector. However, there are examples such as NICATS and the Derbyshire Regional Network which use level descriptors for both academic and vocational qualifications in the HE and FE sector.

During the Delphi study, it became evident that some of the descriptors that the educators agreed upon suggested that progression in HE was linear but there were other characteristics associated with particular levels that were not progressive. It was thought that the progressive and non-progressive level descriptors might be an artefact of the method, which was used. However, James and Redfern (1995) also found that there were some characteristics in levels that are not necessarily progressive. That is, there may be qualitatively different characteristics at each level. This suggests that there are both progressive and non-progressive characteristics associated with levels, which challenges the assumptions of CATS, adopted by InCCA (1998), that levels reflect linear progression, so the levels model is limited.

The most often quoted limitations of the Delphi method are the definition of experts and consensus. These definitions will vary from study to study. Many comparability studies use experts e.g. Coles and Matthews (1996) and Gray (1999) and therefore defining expertise is a limitation familiar to these types of studies. Greatorex and Dexter (1998) suggested a method for analysing Delphi data which gives insights into what happens between the rounds which affects the nature of the consensus. They pointed out that other authors have discussed other limitations of the Delphi technique:

- Sackman (1975) and Bardecki (1984) have outlined the different reasons for experts dropping out of Delphi studies;
- Erffmeyer et al. (1986) have discussed the optimum number of rounds;
- Kastein et al.(1993) considered the reliability of the outcomes of Delphi studies;

- Mulgrave and Ducanis (1975) and Taylor et al. (1990) examined the characteristics of experts which might affect their behaviour;
- Schiebe et al. (1975) and Martino (1993) scrutinised the stability of judgements across rounds.

Greatorex (1998) pointed out that the limitation of the Delphi study as an exercise in comparability is that it does not look at the use of the descriptors in practice i.e. in relation to students' work.

The NICATS (1999) Project was set up to develop a single credit framework across FE and HE in Northern Ireland. The articulation of level descriptors for both FE and HE was at the heart of the NICATS project. The only other example of one single continuum of levels for both the FE and HE sector was the Derbyshire Regional Network. However, the level descriptors associated with this continuum did not illustrate a consistent approach to describing the characteristics associated with each level in FE and HE. The FE level descriptors were generic and the HE descriptors were subject specific. As part of the project a research study was undertaken to develop level descriptors for the credit framework. A task group undertook a literature review of existing level descriptors. Using these and theoretical models (e.g. Bloom's taxonomy (Bloom et al., 1971) and Steinaker and Bell's (1979) Model of Experiential Learning) and the professional experience of the task group, draft level descriptors were developed. Davis and Burnard (1992), Winter (1993) and James and Redfern (1995) pointed out that using Bloom's taxonomy as the basis of level descriptors might not be appropriate. For example, the taxonomy was not intended as a description of the development of knowledge. James and Redfern (1995) found that different cognitive descriptors e.g. 'synthesis' which is part of Bloom's taxonomy was associated with different levels by different Nurse educators. For more details about Bloom's taxonomy, see Kreitzer and Madaus (1994). The level descriptors were presented in a consultation document. FE and HE institutions and organisations like Further Education Development Agency responded to the consultation document. The responses were analysed and NICATS responded to the responses. In the light of the consultation the level descriptors were revised and guidance notes for the interpretation of the descriptors were developed. Curriculum specialists were seconded from HE and FE one day a week for nine months to consult on the applicability of descriptors in different subject areas. The curriculum specialists chose as far as possible a range of programmes across a variety of institutions and in some cases professional bodies were also consulted. A representative of these programmes and professional bodies was forwarded the NICATS literature and then interviewed in confidence by the curriculum specialists about their understanding of CATS, NICATS and the implications for their programme. They were also asked about their definition of a level, the NICATS level descriptors, what their course documentation said about levels and whether this corresponded with the NICATS level descriptors. The relevant course documentation was also examined for indicators of level. The subject specialists each produced a subject specific report and the reports were summarised in a second report. The NICATS level descriptors were revised in the light of this consultation to produce level descriptors upon which HE and FE representatives agreed (NICATS, 1999). The NICATS project is now in the process of operationalising the credit framework and the level descriptors. The project is undergoing a continual process of evaluation.

QAA and its qualifications and credit framework

QAA (1999c) are developing a new qualifications and credit framework for HE due to be finished in 2000. The credit and qualifications framework assumes that all qualifications that are awarded at a particular level are comparable. In the CNAA levels framework, there were 4 levels that did not incorporate MPhil and PhD study (CNAA, 1989). QAA are still in consultation with HE practitioners to decide how many levels there should be in the QAA framework. This includes developing level descriptors that will be used as criteria for determining whether a programme is worthy of being labelled, say, a master's programme. These level descriptors for undergraduate levels will be developed from benchmark standards for each subject. However, the method for developing the level descriptors from the benchmarks has not yet been made public. Therefore it is difficult to establish the utility of the QAA's method of developing level descriptors from benchmarks. Level descriptors for postgraduate levels cannot be developed in the same way as for undergraduate qualifications as there are no QAA benchmarks for postgraduate qualifications. Both the undergraduate and postgraduate level descriptors will also be developed from level descriptors that have been written by InCCA(1998) and South East England Consortium - SEEC (1996). The level descriptors will generally be subject specific. However, QAA will also develop interdisciplinary level descriptors to judge the quality of programmes. These level descriptors cannot be developed until the number of levels and method for developing the descriptors has been agreed.

Comparison of qualifications at the same level but from different qualification frameworks

It is difficult to compare qualifications at the same level in different frameworks or qualifications that have different purposes. In this section the limitations of comparing qualifications for different purposes are noted and the methods of comparing the qualifications are discussed.

NVQ/SVQs and other occupational qualifications are located in a national framework for vocational qualifications. More recently, GNVQs have also been incorporated in this framework. GNVQs can be seen as pre-vocational qualifications - they are preparation for the workplace. In contrast, NVQs are vocational qualifications, they are recognition that someone is competent to practice in a particular occupation. The framework extends from the most basic occupation through to 'professional' competence. It therefore extends beyond the scope of higher education. The Department of Employment (1995) say that NVQ level 4 is widely believed to be equivalent to degree level. But Debling (1995) suggested that some graduates may remain in level 3 jobs for a decade before they reach level 4 and 5 NVQ professional award status. There has been a steady increase in the number of HE students and students taking and passing A levels. This has led to qualification inflation, which could result in individuals being employed in positions for which they are over qualified and becoming unmotivated.

Degrees and vocational qualifications are awarded within different qualification frameworks as they are for different purposes. It is not a simple matter to equate the vocational levels with the HE levels, as 'level' in NVQ terms describes the complexity of occupational role and level in HE is about intellectual difficulty or academic autonomy (Mitchell, 1993). Another reason why it is hard to compare NVQs and degrees is that the level frameworks operate differently. In HE, modules in the CATS framework are each mapped to a level, then students pick the modules that they would like to undertake to form their own programmes. Raggatt and Williams (1999) explain that in the NVQ framework, NVQs are assigned to levels but that some units in the same qualification may be more suitable for different levels. It is the combination of these skills and knowledge that are associated with a particular level of occupational competence.

Jessup (1995) argued that it was difficult to compare GNVQs with GCSEs and GCE A levels as they were for different purposes. Taking this argument further it would be even more difficult to compare NVQs and A levels, GCSEs or university degrees in academic subjects. In the NVQ framework level 3 is notionally equivalent to A level (Wolf, 1995). Despite the protest of Jessup about the dangers of comparing qualifications with different purposes there are examples of comparability studies between vocational and academic qualifications (Coles and Matthews, 1995). These studies are important as there are some qualifications where students are encouraged to study for a degree and gain an NVQ/SVQ simultaneously, e.g., Museum Conservation and Heritage training (Hunt, 1996), and also because both vocational and academic qualifications can be used as different pathways to higher levels of education.

There is some work that maps NVQ criteria to HE learning outcomes within comparable qualifications (Waterhouse and Crook, 1998). Mapping is undertaken by listing the learning outcomes for similar qualifications and indicating which learning outcomes from each qualification are the equivalents of the learning outcomes from the other qualification. Mapping is really a matching exercise based on academic judgement. Although this mapping process can be useful it relies on a common understanding between educationalists about what is meant by the learning outcomes that are being compared. Wolf (1995) argued that a consensus understanding is developed through discussion about examples of work and educator networks. This suggests that educators should not undertake the mapping exercise in isolation. It should involve discussion about exemplars of work to facilitate the comparison of learning outcomes. Mapping is also a time consuming, labour intensive and tedious process.

The Department of Employment (1995) undertook an exercise matching HE degree learning outcomes and NVQ/SVQ competencies in Art and Design. It is widely believed that level 4 is equivalent to degree level learning, but a comparison between level 4 competencies and degree level learning outcomes illustrated that few similarities were found. Therefore the project focused on matching level 3 (below degree level) competencies and learning outcomes. Similarities were found between level 3 competencies and learning outcomes, covering the creative function of design and visual arts practice. The NVQ/SVQ requirements were tested in a HE environment. There were 15 Art and Design undergraduates who took part. The students did not have the employment experience to meet the NVQ/SVQ competencies and the written assignments and seminar presentations that they offered as evidence did not fit the assessment evidence requirements for NVQ/SVQ (Employment Department, 1995). So although mapping exercises may illustrate similarities between outcomes they do not take account of the context in which the outcomes are reached. The degree students were studying for a qualification with the purpose of educational progression and intellectual challenge; this affected the extent to which they could provide evidence of meeting the NVQ/SVQ competencies.

Coles and Matthews (1995) have considered using fitness for purpose as a method for comparing qualifications in terms of their fitness for purpose for progression into employment or higher levels of education. The methodology was tested by comparing GCE A levels in science subjects and Advanced Science GNVQ. In this method, different qualifications are evaluated in terms of an external comparator, as their standards are too different to compare directly. Briefly, the method consists of subject specialists from HE and employers describing the level of knowledge and skills that they require in a qualification at a particular level. Qualifications are scrutinised to identify whether they meet the criteria that have been developed. The method is flexible as it covers all qualifications – vocational and academic - and all levels. It can be used to compare qualifications at the same level and to compare qualifications 'vertically'. Comparability in methodology is based on the perception of some of the 'users': subject specialists, tutors and employers. But candidates who are possibility the most important group of qualification 'users' are not consulted and candidate's achievement is not used to compare the qualifications. Coles and Matthews (1995) acknowledged that the subject specific experts were not necessarily representative of the scientific community and that the method does not involve consideration of students' work. Fitness for purpose methodology could be very important in the early stages of developing a national framework for qualifications. It was found in Science that employer and Higher Education tutors had more in common than they had different in terms of their requirements of advanced qualifications. GCE and GNVQ met their user requirements to the same degree. GNVQ matched more general skill components than GCE A levels. It was found that the users wanted students to do basic things well rather than study a wide range of topics in depth.

Within levels there is not just the issue of how educators view levels and whether agreement exists: there is also the issue of parity of esteem. Dearing (1996) wanted to develop a qualifications framework where vocational and academic qualifications enjoyed parity of esteem. However this may not become a reality. For example, OfSTED and FEFC (1999) have found that despite the hard work of teachers and lecturers many students and parents still do not accept the claim that GNVQ enjoys parity of esteem with A levels. So whilst the achievements in GNVQs and A levels may be comparable the qualifications are not equally valued by students and parents.

Conclusions

The literature revealed discrepancies in the way levels were conceptualised and what constituted each level. This threw doubt on the assumption built into CATS and other qualification frameworks, namely, that there was agreement on these two points. The other limitations of qualification frameworks that use a levels model is that the concept of levels is too simple to represent the complex process of learning and progression. Generic level descriptors can be so ambiguous that they are meaningless which makes subject specific approaches attractive. One of the major limitations of level descriptors is that there has been no systematic attempt to match students' work to level descriptors. In this scenario the development of level descriptors and learning outcomes can at worst be an extensive paper chase.

Generic level descriptors do not provide a basis for explaining why an NVQ and an A-level have been included at the same level in the same framework. It is difficult to compare qualifications from different frameworks as they are often for different purposes. Matching learning outcomes and competencies, etc., may suggest equivalencies. However, it is limited in the same way as the development of generic level descriptors is limited. There should be some attempt to link the learning outcomes or competencies of different programmes to students' work for these programmes or to compare work from the different programmes. That is, comparability studies would benefit from a multi-method approach where comparisons are made between outcomes and students' work or achievements. Otherwise invalid statements about comparability would be made. Of course, in some cases it is difficult to meaningfully compare students' work from different programmes. But if meaningful comparisons cannot be made about students' work then why is comparing statements of learning outcomes considered to be more meaningful?

The issue of 'levelness' will not disappear whilst qualification frameworks are maintained. For example, Curriculum 2000 includes GNVQ single awards, AS levels and A2. The AS level is the first half of an old A level and the A2 is the second half. So there is progression between the two qualifications. However the single award GNVQ is not divided into two halves and so is it of the same level as the A level, the AS level or the A2? This is a study for further research, which could be undertaken using the fitness for purpose methodology 'vertically'.

Review of research into describing profiles of contents and demands of individual qualifications

Initially, this chapter introduces the some of the ways in which the profiles of demands and contents are described in the United Kingdom. Then the background to the review of research into profiles of contents and demands of qualifications is given. This includes a brief discussion of the outcomes approach, criterion referencing and a summary of their limitations. Once the background information is considered, a series of research methods for articulating profiles of contents and demands is discussed. Where possible, methods are discussed in relation to examples of how they have been used.

Descriptions of profiles of contents and demands used in the United Kingdom

This chapter describes methods of articulation of profiles of contents and demands of individual qualifications. There are several ways in which descriptions of profiles and contents are given. Some of the major UK schemes are outlined here. In the National Curriculum for England there are 4 key stages.

Key Stage	Pupils' Ages
1	5-7
2	7-11
3	11-14
4	14-16

For each subject and key stage, programmes of study set out what students should be taught using attainment targets (which are short names for the domains that will be taught and assessed). At the end of Key Stages 1, 2, and 3 for all subjects except art, music and physical education, standards of students performance are set out in eight level descriptors. The level descriptors are different for each attainment target. Each level descriptor is more difficult than the descriptor below. A Level 2 descriptor was initially set at the attainment of the 'average' 7 year old and level 4 the attainment of the 'average' 11 year old. The level descriptors are used as an assessment tool to assess the attainment of each student (DFE, 1995).

The national tests for 16 year olds are General Certificates of Secondary Education (GCSEs). When a child is awarded a GCSE they are awarded a grade. When GCSEs were initially developed there were attempts to develop grade related criteria. These were the criteria that a student has to reach to be awarded a particular grade (Kingdom and Stobart, 1988). Subsequently the grade criteria were dropped. Grade descriptors have been used as an indicator of the qualities that might be expected to be found in a performance at a particular grade since before grade criteria were developed (Gipps, 1990). Level descriptors and grades do not tend to be explicitly related to one another.

The structure used to describe NVQs is different. NVQs are based on groupings of occupational standards. The occupational standards have 'three aspects:

- Elements of competence which state the functions which are needed in particular occupational areas;
- Performance criteria, which are attached to each element, describe the quality of the outcomes of successful performance;
- The indicators of range describe the potential dimensions or parameters of the function - what is included in the coverage of the element and performance criteria and what is not' (Mitchell, 1995, 96).

The range statements have now been dropped. After the occupational standards had been grouped together in units and qualifications, the qualifications were assigned a level. The NVQ level descriptors have been articulated and they are generic descriptors that cover all NVQs.

There is a different structure to describe qualifications for higher education in the UK. HE qualifications tend to be divided into modules or units, which are short courses that can be assessed on their own. The requirements to be awarded the academic credit for a module are often expressed as learning outcomes. Learning outcomes indicate the standards (often minimum standards) that students are required to reach in order to pass a module. Each module is allocated a level and this is the level at which credit is awarded.

The characteristics associated with levels are described in level descriptors. Unlike National Curriculum level descriptors, HE level descriptors are not used as assessment tools but are used to allocate levels to modules (Moon, 1999). In this chapter, one of the main references is Winter and Maisch (1996) who developed social work and engineering programmes through a research project. The programmes were structured within a modular scheme and CATS (Credit Accumulation and Transfer Scheme). Each module consisted of a set of competence statements (elements of competence) and the set of competencies constituted a unit of competence. Students needed to achieve the competencies to be awarded the credit for a module. In this way, they are similar to the learning outcomes. In addition, there was a second dimension of competence: the core assessment criteria, which embodied the general requirements of the professional role, whereas the competencies refer to the specifics of the occupational role. The general criteria were applied through the assessment of each module. Although there is a general trend within HE to use a structure of level descriptors and learning outcomes or competencies, there are a variety of ways to structure and articulate these level descriptors and learning outcomes.

Randall (1998), the Chief Executive of the Quality Assurance Agency (QAA), explained that QAA aim to establish whether there is comparability between degree standards to hold Higher Education accountable to stakeholders and that to facilitate this there will be:

- A qualifications framework – to ensure that qualifications that share a title are of a common level and nature (this involves defining both levels and the criteria by which a degree can be considered to be worthy of the title Bachelors etc.)
- Guidance on developing programme specifications so institutions can clearly set out the outcomes of their programmes;
- Nationally agreed subject specific benchmark standards against which all programmes in a particular subject can be judged.

This process requires a great deal of articulation but the details of the methods have not yet been made public (QAA, 1999).

Another aspect of describing profiles of the content and demands of qualifications is to describe the demands of the tasks that the students are expected to tackle. The demands of a task are the cognitive skills and processes and their application of these which are required by a task. As an example, Hughes et al. (1998) developed a scale of demands for GCSE and GCE examination questions. For Geography questions, a lower demand was 'There are a number of simple steps in the question' and the corresponding higher demand is 'the question is not broken down' (Hughes et al., 1998, 12).

In summary, there is a jargon jungle associated with performance assessment, curriculum, qualifications, their contents and profiles. Indeed not only is there a variety of jargon which is used to describe the qualifications, their contents and profiles but also some of the jargon is used to mean more than one type of description. For example, a National Curriculum level descriptor is different to a higher education level descriptor in content and purpose.

A Review of Research into describing the Profiles and Contents of Qualifications

It is beyond the scope of this report to give a full history of the trend towards describing the profiles of demands and contents of qualifications. This summary focuses upon national developments and traces the ideas behind criterion referencing and the outcomes approach to developments in performance assessment and educational measurement in the United States of America. Brown (1980), Wolf (1995) and Burke (1995) give fuller accounts of this area.

In the 1980s, vocational qualifications were rationalised into a national vocational qualifications framework (Raggatt and Williams, 1999). There have been similar trends in higher education with developments like the Credit Accumulation and Transfer Scheme (CATS) framework and the Dearing proposals in the National Committee of Inquiry into Higher Education (NCIHE) (1997) for a new HE qualification framework. These attempted to 'tidy up the landscape' or to map out how qualifications relate to one another, i.e., which qualifications were comparable and provide appropriate progression routes to higher levels of education? These moves were accompanied by the pressure to describe achievement.

The result of the trend towards increased description of programmes and qualifications at all levels is to try and secure standards by making them explicit and to explain to students, teachers and employers (customers) what is required of students to be awarded a qualification. It is also linked to the general shift away from norm referencing to criterion referencing. As early as 1936 Hartog and Rhodes supported the use of clearly specified criteria because when experienced teachers do not have these criteria, there is considerable disagreement between them about the work of individual pupils. More recent examples of this trend are National Vocational Qualifications (NVQs) - which are based on explicit national standards (Raggatt and Williams, 1999), the development of grade related criteria for GCSE (Kingdon and Stobart, 1988) and the movement towards learning outcomes in higher education (HE) (Race, 1998).

Cresswell and Houston (1991) explained that attainment has been described in many ways including grade descriptors, grade criteria, level descriptors, assessment objectives, attainment targets and statements of attainment. They all purport to describe levels of attainment in behavioural terms, the descriptions tend to ignore context and can be very generalised. Such tools have derived from criterion referenced testing. As reported above there are still many different ways of describing attainment. Brown (1980) defines criterion-referenced assessment as 'Assessment that provides information about the specific knowledge and abilities of pupils through their performances on various kinds of tasks that are interpretable in terms of what the pupils know or can do, without reference to the performance of others'. The argument for criterion referenced assessment is that it can be used to facilitate learning. It is different from norm-referenced assessment, which compares students' performance against one another. The term criterion referencing did not emerge until Glaser and Klaus used it in 1962 although the ideas behind it date much further back. For example, in 1913 Thorndike was concerned that marks tend to have relative meanings and that, a score of 60% is better than one of 50% but no-one knows what either student knows and can do. By 1931 Tyler had developed a generalised technique for test construction based on the principle that objectives should be defined in behavioural terms. Tyler has been related to the outcomes approach as well as the criterion referencing approach (see below). There are strong parallels; both approaches are dedicated to stating what learners know and can do. Glaser and Klaus (1962, 421) argued that 'criterion-referenced measures depend on an absolute standard of quality' and on page 422 they continued that 'knowledge of an individual's score on a criterion referenced measure provides explicit information as to what the individual can or cannot do'. Popham and Husek (1969) challenged psychometricians to develop the technology to take criterion referenced assessment forward. Brown (1980) explains that criterion referencing has been defined in many different ways. In Great Britain National Vocational Qualifications were based upon an outcomes or criterion referencing approach to assessment. Wolf (1995) explained that a problem with NVQs had been that they had attempted to provide very specific criteria to be met. But this increase in description did not necessarily lead to better assessment.

Arguably descriptions of expectations or demands are useful in suggesting to teachers and students what is required for a particular syllabus. For example, A level and GCSE syllabuses sometimes include grade descriptors which indicate what candidates are expected to do to achieve a particular grade. Race (1998) explained that learning outcomes give details of syllabus content and what students are expected to achieve, and they are also an indication of standards. This is an important point because two people can study the same content and achieve different standards of achievement. The learning outcomes can help students make informed decisions about choosing a programme or module and make explicit the targets that they are expected to reach. Race (1998) also explained that learning outcomes indicate to external stakeholders of HE (often employers) the standard, level and relevance of modules and programmes. Statements about attainment that are useful to students and employers are not easily useable as assessment criteria because attainment is context dependent and the statements are so general that they are meaningless. Cresswell and Houston (1991) argued that the goal of using the same descriptions for users and assessors is unobtainable. There are some other uses for descriptions of the standards/requirements of programmes. The following are some examples:

- level descriptors from the National Curriculum are used as an assessment tool. The progress of students is identified by awarding a level usually on a holistic basis (Cresswell and Houston, 1991);

- Coles and Matthews (1996) developed a method of comparing qualifications and evaluating their fitness for purpose. To compare the qualifications they profiled the knowledge and skills that tutors and employers required of people with the qualifications;
- In the case of NVQs, national occupational standards were developed to move towards job opportunities being based on competence rather than time served. It was hoped that this would increase opportunities for people to be recognised as competent and hold a qualification that had currency in the job market. Similarly, changing to a learning outcomes approach in HE opens opportunities for candidates to claim academic credits for their achievements through experience, e.g., 'on the job' learning rather than by attending a programme and doing an assignment to achieve the credit when they may have already met the programme requirements.

The limitation of these approaches is that standards cannot be maintained through making them explicit – the standards do not reside in the statements of standards; they are maintained in the tacit knowledge of professional and academic communities (Wolf, 1995). She explained that when teachers are given National Curriculum Attainment Targets they disagree about the levels for which they are written. However when the attainment targets are presented with examples there is little disagreement between the teachers about the appropriate levels. This illustrates that standards are conveyed by how teachers interpret descriptions of standards and demands in relation to examples of work.

There have been criticisms of these moves towards an outcomes or competency based approach. Burke (1995) argued that if the theoretical basis of the outcomes approach is flawed then the whole enterprise is in jeopardy. Outcomes are the linchpin of the NVQ approach to learning and assessment. Jessup's competency based or outcomes model was used as the foundation for NVQs. Burke (1995) argued that Jessup's outcomes model is a form of objectives theory. For more details of the outcomes model, see Jessup (1991). The outcomes or objectives approach can be traced through authors such as Taylor (1912), Bobbitt (1918), Charters (1924), Tyler (1949) and Bloom (1956) who applied their theories in education, training and the workplace. In the literature the outcomes approach is linked to behaviourism, for example, Gagné (1965). Burke (1995) argued that this is not the behaviourist psychology of Skinner. The difference is that the outcomes approach is flexible enough to go beyond the observable, whereas behaviourist psychology is confined to that which can be observed. Also the outcomes approach is viewed as a way of empowering the learner rather than modifying behaviour which is associated with behaviourist psychology.

Stenhouse (1975) argued that the objectives approach is not suitable for initiating people into knowledge, although he was more at ease with it being used in training and instruction. The outcomes approach is considered to stifle creativity. It is not a good way to improve practice. But Jessup (1991) argued that outcomes which refer to personal and intellectual development can be operationalised. The difference between Stenhouse's approach and Jessup's is that Jessup's model incorporated all kinds of learning, including learning from experience, in contrast to Stenhouse's work which referred to the educational curriculum. Atkins et al. (1993) were concerned that an outcomes approach in HE would encourage students to take a minimalist approach where they would only try to achieve the benchmark standards that they needed to pass rather than striving for excellence which is one of the purposes of HE. Another concern was that students would all become too similar as they all achieved the same benchmark objectives. However, Burke (1995) pointed out that there are intended and unintended outcomes from the outcomes approach and therefore not all students are the same.

Cresswell and Houston (1991) explained that profiles of the knowledge and skills which candidates achieve cannot be context free as attainment is context and task related (they consider these profiles to be more useful than a single grade). This suggests that when statements like competencies are developed they are not entirely independent of context and content. One of the criticisms that has been levelled at the outcomes approach is that it has divorced outcomes from content. Many people follow Stenhouse's (1975) view that the curriculum is a process and therefore reject the very notion of outcomes. Jessup (1991) argued the outcomes approach ignores the process of reaching an outcome in so much as an outcome is awarded credit how ever or where ever the learning was acquired. However, as all aspects of the curriculum (teaching, learning, subject contents and assessment) are linked and if one is affected they are all affected. It follows that outcomes cannot be divorced from the learning process. The outcomes approach can lead to the phenomenon of *teaching to the test* where what is taught in the curriculum is driven by assessment requirements (Broadfoot, 1996). These assessment requirements are increasingly expressed as outcomes and the curriculum is therefore outcomes driven. This is a problem if the defined learning outcomes are incomplete or unsatisfactory. Although there is a range of views about this issue it seems that it is difficult to divorce outcomes from content and context.

The purpose of this chapter is to consider methods for articulating profiles of contents and demands of qualifications including, grade descriptors, level descriptors, competencies and similar indicators or criteria. The review covers academic and vocational qualifications. Some of the examples consider the articulation of statements about individual qualifications and others are about groups of qualifications. The methods may be valid for individual qualifications and groups of qualifications. This chapter also includes methods which authors suggested for articulating outcomes.

Methods of Articulation

This next section is a discussion of the methods that have been used or recommended for articulating the contents and profiles of qualifications. The table below is an attempt to summarise the methods that are discussed and to give an indication of what they have been used to articulate or what they could be used to articulate. Some methods are appropriate when starting from scratch (carte blanche scenario) and the other methods are for improving an existing articulation. There are some methods that can be applied to both. These methods are described in more detail in the remainder of this section. It can also been seen from the table that the same method can be used to develop different types of descriptions, e.g., KRG has been used to develop grade descriptors, descriptions of the demands of examination questions, core assessment criteria, competencies for jobs and behavioural indicators. Different methods can be used individually or together to develop the same type of descriptions, e.g., mastery levels analysis and Kelly's Repertory Grid have both been used to describe the different attainments associated with different grades. This suggests that a research method should be chosen on the basis of its suitability to develop a particular type of description.

Table 1: Summary of Methods of Articulation

Method of Articulation		This method was used to gain
Methods from a carte blanch scenario	**Methods where something has already been articulated**	
Delphi Technique	Delphi Technique	Level descriptors A consensus on the results of a functional analysis.
	Angoff	Cut scores
Mastery levels		Grade descriptors
First stage of KRG		Grade descriptors
First stages of KRG	Second Stage of KRG	Demands of examination questions Core Assessment Criteria Competencies for Jobs Behavioural Indicators
Working Parties		Grade related criteria Subject benchmark standards
Domain approach		Grade related criteria
	Q sort	Grade related criteria
Functional Analysis		Competencies Occupational Standards
Fitness for Purpose		Describing knowledge, skills and understanding associated with different qualifications
Critical Incidence analysis		Gain a description of people's characteristics or psychological factors relating to performance
		This method could be used to
Nominal Group Technique		Gain a consensus on which competencies were the most important Gain a description of the nature of professional practice
Focus Groups		Gain a description of the tasks that are undertaken as part of a job role
	Content Analysis	
Network Analysis		Gain a description of the content and organisation of a person's knowledge of a domain
	Position Analysis Questionnaire	Cluster jobs into families of similar knowledge and skills
	Work Profiling Questionnaire	Develop job descriptions, profile jobs and human attitudes

Delphi Technique

The Delphi technique is a method that is used to identify whether there is an agreement between experts on a topic or problem. A Delphi study consists of a series of stages. Initially there is an open-ended survey where experts give their views and suggest answers to the problem. The Delphi researcher collates these answers and summarises them into a second survey. In response to this survey experts vote to indicate which views they agree with. In the third survey the experts are again presented with the different views, but they are also presented with the responses of the other participants. The experts are asked to vote again in the light of the group feedback. The third round can be repeated as a fourth round. In the third and fourth rounds experts can be asked to justify why they are voting differently to the majority of the experts, which provides an insight into why some experts do not agree with one another. The experts are considered to have reached a consensus when the agreement has stabilised (Linstone and Turoff, 1975).

Davis and Burnard (1992) used the Delphi technique to identify the differences between different qualification levels for nursing (Diploma, Bachelor, Masters and Doctoral). Greatorex (1998) also used this method for identifying the qualities educators associated with levels in the Credit Accumulation and Transfer Scheme (CATS). The Delphi technique was used in a comparative context as the level descriptors developed in the study were appropriate for the field of Healthcare including Nursing, Midwifery, Radiography and Physiotherapy. Greatorex (1998) pointed out that the limitation of the Delphi study as an exercise in comparability is that it does not look at the use of the descriptors in practice, i.e., in relation to students' work. The details of these studies given in chapter 2. A more general discussion of the Delphi technique is given below.

The most often quoted limitations of the Delphi method are the definition of *experts* and *consensus*. The definitions of consensus and expert will vary from study to study. Many comparability studies utilise experts, e.g., Coles and Matthews (1996) and Gray (1999), and therefore defining expertise is a limitation familiar to these types of studies. However the issue of how to define consensus is not necessarily such an issue when other methods are used. Greatorex and Dexter (1998) suggested a method for analysing Delphi data which gives insights into what happens between the rounds which affects the nature of the consensus. They noted that the mean as a measure of central tendency represents the group opinion of the panel. The standard deviation as a measure of spread, represents the amount of agreement in the panel. For a summary of how other authors have discussed the other limitations of the Delphi technique see chapter 2.

In the context of developing profiles of contents and demands the first round of the Delphi study would be the time when educators would initially articulate their views about what constituted the profiles and or content. The later rounds would be a way of identifying whether the educators agreed on the profiles and content. This is essentially how Davis and Burnard (1992) used the Delphi technique. Greatorex (1998) did not have the first stage of the Delphi technique in her study, as in a pilot study the educators said that it was too difficult a task; they needed something to begin the process of articulation and defining levels. Rather than the experts offering ideas from which level descriptors could be written, level descriptors which already existed were used. Winter and Maisch (1996) also omitted the brainstorming stage of the Delphi process in their survey to identify whether there was a consensus amongst social work practitioners on the contents of a functional document (or functional map), which are described later in this chapter.

Angoff Procedure

The Angoff procedure is a method of setting cutscores. Initially judges predict the likelihood of a pupil of a certain level of attainment answering items correctly. Then the judges are provided with more information, e.g., facility values indicating the difficulty of each question. At this stage the judges can revise their judgements. Finally some judges explain their reasons for the standards that they have set. Judges can at this point revise their judgements. Jaeger (1978) pointed out that the Angoff technique is similar to the Delphi technique. For example, the judges used in both procedures are deemed to be experts and judges reconsider their opinion (in the case of the Angoff procedure, this is a cutscore) in the light of new information. The Angoff methodology is used to derive cut scores on tests from criteria. The method does not provide a way of developing or articulating the criteria and will not therefore be considered in detail.

The advantage of the Angoff technique, like Delphi is that it encourages discussion, which can lead to more homogenous assessor judgement. An example of the Angoff technique being used to derive cut scores from criteria is Morrison et al. (1994). Their criteria were based on the concept of a pupil who would just gain a level 5 from the National Curriculum.

Mastery levels analysis

Massey (1982) set out to describe to examiners what students *actually* achieved, so examiners could consider how this differed from what they expected candidates to do. This approach to describing attainment is based on the notion of mastery. Candidates are divided into grade groups (groups according to the grade they achieved). A grade group is considered to have mastered a question part if the marks gained by the grade group meet particular statistical criteria. A grade group is considered to have reached mastery level for a given question part when:

- the average mark for that component grade group is 75% or more;
- the average mark for all component grade groups below the original component grade group is less than 75%;
- the average mark for the component grade group above is more than 75%;
- there is a significant difference ($p<=0.05$) between the marks of the original component grade group and the component grade group below. t-tests were used to identify significance. The use of the 75% is quite arbitrary. Other percentages could also be used. Greatorex (1999b) suggested that 50% is arguably the lowest percentage that can be used as a mastery level as a level lower than 50% would mean that the candidates had mastered less than they had achieved.

For the questions where the candidates achieved mastery the specific knowledge and skills that they need to master the questions can be described. This method of mastery levels analysis has restricted use as it is only suitable for describing performance in relation to tests, examinations or exercises. However, this is a useful method for identifying tasks that discriminate between higher and lower achieving candidates. This method is suitable for use in a single qualification or examination.

Kelly's Repertory Grid

There are two examples of studies where Kelly's Repertory Grid (KRG) has been used along with other research methods to develop grade descriptors:

- Pollitt and Murray (1996) combined Thurstone pairs analysis and KRG;
- Greatorex (1999b) combined a mastery levels analysis with KRG.

Kelly (1955) used a grid to identify individuals' personal constructs (mental representations or concepts which individuals use to think about or view people). The grid method has been refined and used in a variety of contexts besides personal constructs. It can be used to identify interviewees' perceptions and views (Cohen and Manion, 1994). There are two parts to KRG. The first part involves eliciting participants' personal constructs (views) by asking them to compare objects or people and explaining how they are similar and different. The second part is when participants are supplied with dichotomous constructs, e.g., gentle and aggressive, to which they respond, indicating how each construct applies to an object or person, e.g., indicate on the Likert scale whether a particular person is gentle (1) or aggressive (5).

In Pollitt and Murray's (1996) study based on performance in a Cambridge Certificate of Proficiency in English examination, the Thurstone pairs method was used as a technique for developing a scale upon which stimuli can be located. In this case it was used to identify which scripts showed higher and lower performance. In this method two performances were compared by a number of expert judges and the results were analysed to establish the ranking of the performances from high to low performance. KRG was then used to help the judges describe the characteristics that were evident in the performances. After the judges had indicated which performance in each comparison pair was the higher performance, they were asked to talk about how the performances were similar and different. From these qualitative descriptions, Pollitt and Murray (1996) listed the qualities that judges paid attention to when they distinguished between high and low performance. They concluded that this combination of methods did not generate insights very effectively. However they argued that the comparison format did facilitate the articulation of the performance characteristics which appeared salient to the judges. Fulcher (1993) argued that many descriptors in current rating scales for language proficiency are thought up by the scale designers and have little empirical basis. So any research method that has been used to develop outcomes based on evidence is an improvement on this situation.

Greatorex (1999b) built on work by Massey (1982) (who used a *mastery levels analysis* - strict statistical criteria) and Pollitt and Murray (1996), who used KRG, both suggested that an effective way of writing grade descriptors is to focus on what the distinguishing characteristics of performance at different levels are. Greatorex (1999b) used a mastery levels analysis to identify which questions in an Accounting A level discriminated between performance at adjacent grades. In other words, which questions were *mastered* by A grade candidates but not by B grade candidates. This suggested where in candidates' scripts Accounting Experts (AEs) should look to identify the qualities that distinguished A from B grade candidates. Scripts were compared in triads, for example, two A grade answers to a particular question were compared with one another and a B grade answer. The AEs explained how the A grade answers were similar to one another and how these differed from the B grade script. This could be considered to be a biased way of using KRG as the AEs were focusing on the similarities between A grade achievement and how this differed from B grade achievement. However, this was precisely the purpose of the exercise in an attempt to elicit the AEs tacit knowledge about the distinguishing characteristics of performance at each grade. This method is different from that used by Pollitt and Murray (1996) who used pairs of scripts rather than triads. Both methods are valid ways of conducting a KRG study (Cohen and Manion, 1994).

The grade descriptors from Greatorex (1999b) were validated by using them at Award Meetings for two Accounting A levels. Grade descriptors were developed based on two different years of the assessment and a mapping exercise was undertaken to identify whether there was comparability from year to year. This method was successful in developing grade descriptors. The advantages of this method of developing grade descriptors are that the descriptors:

- are grounded in candidates' responses (this is also a disadvantage as the descriptors were developed post hoc rather than deciding what candidates are expected to do and then developing the assessments to test the desired knowledge and skills);
- are focused on the characteristics which distinguished achievement at one grade from achievement at another grade.

However the method is comparatively labour intensive and the Professional Officer, who managed the syllabus under consideration, commented that the resulting grade descriptors were not significantly different from descriptors that were not based on empirical evidence. One of the questions that was raised in this research was whether the average/strong/weak grade A performance should be described. Average grade performance was used as this was considered to be more typical of performance at each grade. This fits with Kandola and Pearn's (1992, 36) assertion that when writing competencies developers need to decide whether to describe 'jobs as they are done or as they might be done...whether to focus on job performed at an average level or concentrate solely on performance of above–average performers'.

Both the Greatorex (1999b) and the Pollitt and Murray (1996) studies were based around a single qualification. However the methods could well be applicable to other qualifications. These studies also used used the first part of the KRG procedure. Other researchers, e.g.,Winter and Maisch (1996) and Hughes et al. (1998), used both the first and second parts of the procedure.

Winter and Maisch (1996) were involved in the development of competence referenced professional education programmes in engineering and social work. This is called the ASSET program. They used KRG along with other methods to develop core assessment criteria for social work. (Their use of functional analysis as their primary method will be discussed later). Twelve social workers were asked to list ten colleagues who were engaged in similar work to themselves. They considered triads of three names and wrote down a quality possessed by two people but not the third. This was repeated until all combinations of names were exhausted and or the ten qualities had been listed. Winter and Maisch (1996) also used the second part of the KRG procedure. They asked the participants to give the quality that was the opposite to each quality that they had listed. Then the participants ranked the pairs of qualities in order of importance to social work activities. For example, practical as opposed to impractical was ranked as the most important pair of qualities.

Winter and Maisch (1996), Pollitt and Murray (1996) and Greatorex (1999b) have all used the first part of the KRG procedure to elicit *qualities*. These qualities may need some editing and rewording before they can become part of statements of outcomes like grade descriptors.

Hughes et al. (1998) used KRG to gauge the demands of A level and GCSE questions. KRG has the power to make what is implicit explicit. This is useful because SRAC (1990) found that examiners' concepts of demands are implicit rather than explicit. In the Hughes et al. (1998) study examiners were presented with triads of questions, and they described how two were similar to each other and different from the third. The examiners then rated each question on a scale of 1 to 5 for each of the constructs that they had elicited. The results were factor analysed to group similar constructs together. The results were compared with other scales of cognitive demand. As an example, the GCSE Geography results suggested that there were three different types of demand:

- the provision and use of resources and information;
- the links between different aspects of the questions;
- the core difficulty of the content of Geography.

The demands for A level Geography and Chemistry and History GCSE and A level were also profiled. Hughes et al. (1998) argued that their scales need validating and the interrater reliability of the KRG scales needs to be checked. They suggested that a single scale may be developed for gauging the demand of different subject areas, however this scale may be more useful for some subjects than others. They concluded that the scales provide a language for examiners to articulate and share discussion about demands.

The studies that have been described above have all used the first stage of KRG and some studies have used the second part. It is also possible to conduct just the second part of KRG if participants are provided with constructs (Cohen and Manion, 1994). However, this raises the question of how were the constructs were chosen?

It can be difficult to decide how to analyse the data from KRG which involve scales and ratings. As mentioned earlier, Hughes et al. (1998) used factor analysis. Cohen and Manion (1994) gave an account of the methods that are used to analyse data from the rankings and scales that are used in KRG. They also gave an overview of the software that is available to analyse the data. Stewart (1998) also offered information about grid analysis and the computer assisted analysis of grids.

The studies that have been described above are examples of when KRG has been used, but other authors recommend it for other purposes. Shackleton (1992) suggested repertory grids for eliciting competencies for future job roles and for comparing good and poorer performers. Kandola and Pearn (1992) suggested that KRG is useful for differentiating between good and less good performers and the development of behavioural indicators. However they warned that the repertory grid does not give a detailed picture of the tasks to be carried out and the objectives to be met. They suggested that the results of the first part of the KRG procedure can be content analysed or developed into a series of scales. They recommend that KRG is used along with other methodologies, which was the case in studies that have been described here.

Working parties

Many of the methods of articulation that are described here are research data collection and analysis methods. But Working Parties do not necessarily use a research method for data collection or analysis, they might use almost any approach that they prefer depending on the autonomy of the working party within the project. Two examples of projects involving working parties have been given below.

Gipps (1990) explained that Sir Keith Joseph's announcement of the GCSE in 1984 included a reference to 'grade related criteria' – the requirements that students must reach to be awarded a grade. The grade related criteria were proposed because there were concerns that there was a lack of comparability between Examination Boards. The Secondary Examinations Council (SEC) set up working parties in each of the main subjects to develop the criteria. They 'identified domains - coherent and defined areas of knowledge, understanding and skills within each subject... The groups then broke the domains down into abilities and produced definitions of performance, or criteria, required for achievement at different levels' (Gipps, 1990, 83). Cresswell (1987) pointed out that to formulate grade related criteria in subject specific terms the working parties often divided domains into subdomains which may have added to the complexity of the criteria. Gipps (1990) reported that this method was not successful as the resulting criteria were too complex. This made assessment unmanageable and they were not useful for employers. Cresswell (1987) pointed to another limitation of the domain approach - the grade related criteria could be interpreted in different ways. This is a problem with any criterion referencing or outcomes approach to assessment. He explained that the domain approach can be used so that a grade translates into reaching the requirements of a particular combination of domains. There are problems with assigning some performances to different grades as performances fall between grades. This is why for the National Curriculum, teachers use a rule of best fit so performance is assigned to the level which best describes the performance. Cresswell (1987) argued that giving a profile of a candidate's achievement in each domain along with a grade is a meaningful way of articulating the competence and knowledge expected of those who have achieved a particular grade. The problem is then how to aggregate the domains to form a particular grade to be awarded. He added that there is little interest from employers and educators in having the information that grade related criteria would provide. Gipps (1990) explained that after consideration and experimentation the grade related criteria were dropped and new approaches to making assessment objective were explored.

Cresswell (1987) reported that SEC argued that attempts to deduce grade related criteria from current or past CSE or GCE scripts had failed because candidates achieved the same grade by choosing different questions. However this did not prevent Greatorex (1999b) from developing grade descriptors from a question paper involving question choice. This was due to one of the differences between grade-related criteria and descriptors; criteria are *requirements* and descriptors are *qualities that are likely to be found in performance*. When candidates are awarded a grade using descriptors, they do not need to exhibit all the qualities listed for a grade to be awarded because grade descriptors are only one of the indicators used to award grades. It may be more difficult to find the same characteristics of performance in all permutations of questions than it is to find distinguishing features of performance at a particular grade and look for some but not all these qualities in a performance.

Another set of profiles of contents and demands developed by working parties will be nationally agreed subject benchmark standards for degrees. Randall (1998) in the Times Higher Educational Supplement explained that nationally agreed Quality Assurance Agency (QAA) subject benchmarks would be developed for 42 subjects. Subject benchmarks were a recommendation made by NCIHE (1997). Subject benchmark groups have the job of developing the benchmarks. Once the benchmarks have been developed reviewers will ask HE practitioners how their learning outcomes and programme specifications reflect the subject benchmarks to cover all degrees in that subject. The national subject benchmarks will be reviewed once they have been used and tested. This serves as a validation phase of the development of the benchmarks. The benchmarks will be used along with the level descriptors developed by SEEC (1996) and InCCA (1998) as the basis for a set of QAA level descriptors. The method for articulating the benchmarks has not been made public, nor has the method for using the benchmarks and existing level descriptors to develop new level descriptors (QAA, 1999).

These two examples are methods that were used to develop a profile of contents and demands for groups of qualifications rather than individual qualifications. However the methods may also be appropriate for individual qualifications. The examples provide an insight into the limitations of the domain approach to developing grade descriptors rather than the advantages and disadvantages of working parties. However this is not the only approach that can be utilised by working parties. What can be taken from this review is that working parties should be encouraged to follow a principle of evidence-based practice, which may involve using appropriate research methods that are discussed here or elsewhere.

Q Sort

Burroughs (1971) described Q sort as a method of construct validation. Participants are presented with a number of items (statements written on cards) which they sort into a scoring scale in accordance with some criteria, e.g., 'liking', 'description of home', 'description of performance'. The items are placed in piles along a scale, for the criterion of liking, a ten point scale might be used with 0 as dislike and the scores in between representing increasing amounts of liking to 10 which represents liking very much. The researcher might limit the number of items in each pile. Burroughs (1971) also suggested that factor analysis could be used to analyse the data from the Q sort. This combination of data collection and analysis is known as Q methodology. Burroughs (1971) argued that Q sort is a better method than ranking when a large number of items are involved (60 to 100 items). He also argued that the validity and reliability of the Q sort methodology are sound. Q sort is a method that can also be used to aggregating or summarising data by allocating similar or related data to a category. For instance, it could be used after the brain storming stage of a Delphi study to reduce the information in the open ended questionnaire into a few statements to which the panellists can indicate their level of agreement.

Hadfield (1980) attempted to identify subject specific grade criteria for a CSE Art Examination. There were six assessors in the study and each had three exhibits; one of CSE grade 1 quality, another of CSE grade 2-3 quality and another of CSE grade 4-5 quality. The assessors were presented with 60 statements, which could apply to any art exhibit. They were asked to sort the statements into 7 piles ranging from pile 1 which was characteristic of the exhibit to pile 7 which was uncharacteristic of the exhibit. Pile 4 was for irrelevant and inapplicable statements. This was repeated by all assessors for each of their exhibits. Each statement was judged on a present/absent dichotomy. Analysis of variance, regression and factor analysis were used to analyse the data. The analysis suggested that there were many routes out of the lower grades and fewer routes into the higher grades. The Q sort technique provided a simple procedure for feeding back to examiners their own operational grade definitions.

Q sort could also be used to identify the extent of the agreement between examiners or judges on the statements that are appropriate to different performances or to develop summary statements about performance from a range of statements about performance. In the later case, judges would put similar statements about performance into the same pile as a method of data reduction. The statements that were grouped together would need to be summarised. However, Q sort needs to be used with another method as it can only be used when some assessment criteria or statements about qualities in candidates' work have already been articulated.

Functional Analysis

Functional analysis is recommended by the DfEE and the National Council for Vocational Qualifications (NCVQ) for developing occupational standards. It is a particular type of interviewing that focuses on the function of a work role, and the actions and tasks that are required to be undertaken. The method was developed by Mansfield and Horton (1986) as an alternative to other forms of occupational analysis which did not lend themselves well to developing standards.

Mitchell (1995) explained that in National Vocational Qualifications a candidate's performance is judged against the occupational standards. Functional analysis is used to develop these standards. This method focuses on occupational roles and the expectations of people performing the role. It is an iterative process, which can involve, for example, retracing the analysis process and adjusting the framework to give a more logical or acceptable structure to the functional map. Mitchell (1995) added that functional analysis produces a functional map describing the purpose of the occupational roles. It does not rank the standards in importance or in a hierarchy. Once the standards have been grouped into occupations the qualifications are allocated a level. Mitchell (1995) pointed out that there has been discussion about whether the functional analysis and outcomes approach that has been adopted at the lower levels could be applied to higher levels. At the time she was writing this debate had not been resolved. The literature reviewed here suggests that this method has been used in the context of developing occupational standards for quite specific work roles. However, it may be that the method can be used to provide standards that can form the basis of more general or academic qualifications. For a detailed and authoritative discussion of functional analysis, and its origin and role in the development of NVQs, see Mansfield and Mitchell (1996).

Winter and Maisch (1996) used functional analysis to develop competence statements that were acceptable to their social work practitioner candidates. They point to a number of advantages of this method:

- it involves a large number of practitioners;
- it is learner centred;
- it focuses on what the practitioners do;
- it avoids researchers' interpretative categories being too influential, thereby avoiding abstraction;
- practitioners have the opportunity to order their own responses.

The advantage of the way Winter and Maisch (1996) used functional analysis is that it listened to students' voices. Other methods, e.g., KRG (Pollitt and Murray, 1996; Greatorex, 1999b) and fitness for purpose (Coles and Matthews, 1996), have been used with experts or employers. In such cases, students' voices have been omitted. Winter and Maisch (1996) pointed out that functional analysis is similar to KRG as they both aim to elicit the views of participants without imposing predetermined concepts from the investigator. However a difference between the methods is that KRG can be used to list *qualities* or participants' interpretative categories, whereas functional analysis describes the *actions* that people are engaged in.

Functional analysis begins by clarifying the key purpose of the occupation then a sequential analysis of the details of the required activities. To develop this statement Winter and Maisch (1996) asked groups of practitioners with three to eight members to 'think of one phrase that sums up the focus of the work that you do?' and 'what is the key purpose of the work that you do?'. This initial summary phase was refined by asking the questions 'What new developments are there in the field?' and 'Do these new developments affect the original summary phrase'. These phrases were elaborated by asking 'What has to happen for the key purpose to be achieved' or 'What is involved in doing this?' So the tasks and roles which are needed to fulfil the occupational purpose were listed. These statements were refined by asking 'What do you have to do to achieve this?' The data from the interviews with a total of 27 practitioners were organised into a functional document where the detailed elements make possible the achievement of more general elements at higher levels in the map. Whilst the ordering of the functional document is guided by the relationships between purpose and action as perceived by the practitioners' the researchers also play a role in clarifying these relationships. Winter and Maisch (1996) used a functional document rather than a functional map as they thought it was more readable. The constituent details can be used as performance criteria. A functional map provides an interpretative description of what is required rather than an objective description. Whilst the method results in performance criteria it should be recognised that occupational practice is a holistic action. Therefore it is important that the initial statement from which the performance criteria are generated is a functional statement of the entire purpose of the occupational role. It should be noted that whilst performance criteria can be developed from the initial key purpose of the occupational role, the key purpose of the occupational role cannot be developed from a reverse process as the whole is greater than the sum of the parts. Winter and Maisch (1996) said that the participants found that the functional analysis was beneficial as they had not reflected on their job roles in this way before. Winter and Maisch (1996) identified that there was a consensus amongst a wider group of practitioners by using the Delphi technique. There were 125 responses to the postal survey including the original 27 practitioners and two collaborative responses. 81% of the responses to the document were very positive but the other comments were taken into account in other areas of their work. The functional document was reorganised and /or rephrased to develop elements of competence for CATS modules.

Mansfield and Mitchell (1996) pointed out that once the functional analysis is complete and the occupational standards have been developed modules and qualifications then need to be developed from the standards. The qualification may not incorporate all the occupational standards it may only cover some. So functional analysis is used to develop profile of an occupation rather than a qualification. However, as illustrated by the work by Winter and Maisch (1996), it can be used to develop standards for the basis of a qualification.

A disadvantage of functional analysis, which has not as yet been mentioned, might be that it focuses on the tasks that need to be undertaken to fulfil a role and this may be perceived to omit the knowledge that is needed to undertake the role effectively.

Fitness for Purpose

Fitness for purpose methodology developed and applied by Coles and Matthews (1996) includes the possibility of comparing the demands of one qualification with another. In designing the project to apply the fitness for purpose methodology, Coles and Matthews (1996) built upon earlier work in this area by Lord et al. (1995). The method consisted of subject specialists describing the level of knowledge and skills that they required in a qualification at a particular level. Qualifications were scrutinised to identify whether they meet the criteria that have been developed. The method is flexible as it covers all qualifications – vocational and academic, and all levels. It could be used to compare qualifications at the same level and to compare qualifications 'vertically'. Comparability in this case was based on the perception of some of the 'users': subject specialists, tutors and employers. But candidates who were possibly the most important group of qualification 'users' were not consulted and candidates' achievement was not used to compare the qualifications. Coles and Matthews (1995) acknowledged that the subject specific experts were not necessarily representative of the scientific community and that the method does not involve consideration of students' work. This is one of the few methods for articulating the demands or outcomes of a qualification and comparing it with another qualification. In studies where other methods for articulation have been used, it is necessary to undertaken further studies to identify comparability. Such additional studies can add to the cost of research. However a large study, with more than one phase, like the fitness for purpose type of studies may also be costly. Winter and Maisch (1996) say that qualitative analysis may lead to researchers using their own interpretative categories and over abstracting the data. It is possible that this disadvantage of some types of qualitative research may apply to the fitness for purpose methodology. Also the methodology does not include the consideration of students' work. This was a conscious decision by the researchers. They argue that students work varies between students and years and that the process of comparing students work can be time consuming. However students' work could be used to validate the descriptions of the qualifications and as another method of comparison to complement using the descriptions as a comparator. This method can be used to develop a profile of demands and contents of one or more qualifications.

Critical Incidence Analysis

In the method jobholders and supervisors or others describe incidents which did or did not meet a job objective. The incidence must be observed in some way and there should be little reasonable doubt that the incident is relevant to effective performance. They are asked to describe what led to the incident and to describe the behaviour displayed and why that was or was not effective. (This comparison of good and poor performance is similar to the KRG approach). Indications of place and time should be given. This is repeated until the participants can no longer think of incidents. This can be a lengthy process. The interview responses are content analysed by experts in an attempt to reach a consensus about trends and clusters. Spencer (1983) described using critical incidence analysis in the context of 'soft skill competencies'. Kandola and Pearn (1992) suggested this method for identifying worker characteristics or psychological factors that contribute to effective job performance. It will not identify the precise tasks and activities to be performed. This is similar to Kelly's Repertory grid, which can be used to elicit qualities but not behaviours. Winter and Maisch's (1996) reservations about content analysis have been recorded above. However there are other methods of analysis that can be used on the data. Kandola and Pearn (1992) explained that a list of activities and behaviours can be developed from the interview data into a rating form. A larger sample of participants then rates the items on a number of criteria. This is a more rigorous approach than content analysis. However, the validity of this approach to the method is determined by how well the list of activities and behaviours represent the accounts of the incidents that are given by the participants.

Behavioural event interviewing is a version of critical incidence analysis. In this method fewer events are considered but in much greater depth, to the extent where the incident may be recreated (Kandola and

Pearn, 1992). So critical incidence analysis can be said to be more reliable whereas behavioural event interviewing is more valid. Kandola and Pearn (1992) added that behavioural event interviewing might be better than critical incident analysis for developing management competencies. Indeed Boyatzis (1982) used behavioural event interviewing to develop a model of the competent manager for the American Management Association.

Whilst these methods identify the behaviours that take place during critical events and the behaviours that are associated with good performance these lists of behaviours need to be developed into competence statements. Kandola and Pearn (1992) suggested that experts categorise the incidents and discuss them to develop agreed competencies. If a more objective approach is required then the results from the rating lists can be factor analysed.

Suggested Methods for Articulation

Thus far the different methods which have been explored have already been used to articulate demands and outcomes. There are other methods that have been suggested for contributing to the development of outcomes e.g. assessment criteria. Many of these methods (nominal group technique, focus groups, content analysis, network analysis) were suggested by Winter and Maisch (1996). Other methods (position analysis questionnaire, work profiling system and job analysis) were suggested in Boam and Sparrow (1992) for analysing occupations and writing competencies. These methods could be used to establish the competencies that should be assessed by both professional and academic qualifications. In the case of professional qualifications this would reflect the approach that was used when developing NVQs and the ASSET programme - identifying competencies from work based practice and using them as the basis for professional qualifications.

Nominal group technique

The procedure for the nominal group technique consisted of two rounds as described below:

Round 1

- each panel member individually wrote down their views
- each panel member contributed one idea which was written on a flip chart by the facilitator
- suggestions were discussed for clarification and evaluation and like suggestions were grouped together by the panel members
- the ideas were ranked privately by each panel member
- the ranking was tabulated

Round 2

- the overall ranking was calculated by the facilitator and fed back to the panel members

The two rounds could be completed as one meeting or the first round could be done by post and the second round conducted as a meeting (Jones and Hunter, 1995).

O'Neill and Jackson (1983) used the nominal group technique to initiate curriculum development. Winter and Maisch (1996) considered it as a method which could be used to articulate the nature of social work practice but decided against it on the basis that from an early stage some contributions are dismissed as the objective is to reach a consensus. The nominal group technique has been used in health and social services, education, industry and government for group data gathering and decision making. It originated from social psychology studies, research into group decision making, aggregating group judgements and citizen participation in group planning. It is useful for gaining a wide input of views, finding whether a consensus existed and ranking and rating problems and solutions (Roth, 1994). The participants can be from any background required by the study (Debold, 1996).

It can be deduced from the above description of nominal group technique that the results of the technique consist of establishing a range of ideas, identifying the best or most pressing points and prioritising these points as a panel. The method is a systematic way of identifying any agreement on the priorities for a situation. Jones and Hunter (1995) contended that the nominal group technique should be used to identify whether a consensus opinion existed, and that as an optional extra, the group discussion could be recorded and used as a source of qualitative data.

The usefulness of the nominal group technique relied on the ability of the facilitator to ensure that:

- professional and personal interests that could overshadow group discussion did not dominate the nominal group;
- all participants had their say.

The quality of a study using the nominal group technique would be affected by the sample of participants who were used and the abilities of the facilitator. The nominal group technique provides a forum for structured discussion and voting. However, it is not designed specifically to make what is implicit explicit as, for instance KRG is, and therefore may not aid the articulation of demands and /or outcomes. It may serve as a method for generating discussion about which competencies are most central to a subject or profession. The nominal group technique is similar to the Delphi technique and they are often discussed together, e.g., Jones and Hunter (1995) and Delbecq et al. (1975). The similarities are that both aim to identify a group consensus through an iterative process using a system of voting. The differences are that the Delphi technique always uses an expert panel, which is not always the case with the nominal group technique, and the Delphi method does not involve face to face interaction which can be a part of the nominal group technique. Jones and Hunter (1995) argued that based on their experience of using both the nominal group technique and the Delphi technique, they preferred the nominal group technique.

Focus groups

Focus group interviewing is often utilised by market researchers and medical researchers but has also been used by social researchers (Gibbs, 1998). Powell et al. (1996: 499) defined a focus group as 'a group of individuals selected and assembled by researchers to discuss and comment on, from personal experience, the topic that is the subject of the research'. The participants in the focus group do not need to be experts, although of course who participates in the study depends on the purposes of the research. The difference between a group interview and a focus group is that the group interview emphasises questions and responses between the researcher and participants. But focus groups utilise the interaction within the group to produce data about topics introduced by the researcher (Gibbs, 1998). This is different to the Delphi technique which seeks to overcome the problems that may occur through a face to face discussion, e.g., personality affecting group dynamics and discussion, by offering a system of voting and feedback which does not involve face to face contact. Focus groups do not necessarily need to be conducted face to face. It can be conducted via email or another electronic conferencing facility. Obviously there may also be problems with this type of interaction.

The focus group method was considered by Winter and Maisch (1996) for gathering data about what social work practitioners actually do. The advantage of focus groups is that a range of ideas and views can be collected from a group of participants in a short space of time (Gibbs, 1998, Catterall and Maclaran, 1997). They can also be used to identify which ideas are salient (Gibbs, 1998). Catterall and Maclaran (1997) noted that there are a variety of methods of analysis that can be used to analyse focus group data, ranging from 'cut and paste' to annotating scripts and content analysis. To understand the data properly and analyse them to best effect, group dynamics and communication must be taken into account. Using a 'snapshot' approach where segments of text are brought together as a report is limited because the 'moving picture' of the focus group (the story of the discussion) is lost. This is better captured by annotating the scripts (Caterall and Maclaran, 1997). Therefore Caterall and Maclaran (1997) suggested that data should be analysed off screen for the 'moving picture' by tracing participants' contributions using highlighter pens. They suggested that researchers should also work on screen to code data to identify 'snapshots' or the range of views which were expressed. There is less control over the data than in an interviewing situation and it is difficult to identify the individual messages as opposed to the group message. Also the group can suffer from group think following an unfruitful tangent (Gibbs, 1998). This discussion about the analysis of focus group data clearly involves researchers imposing their own interpretative framework on the data, which can result in loss of data and over abstraction. This is the reason why Winter and Maisch (1996) rejected the focus group method for developing competencies in their study.

A focus group is similar to a nominal group in that there is a structured group discussion about a specific topic that is chaired or facilitated. The aim of a nominal group is to identify a consensus (Jones and Hunter, 1995). This is not necessarily the aim of a focus group, which is useful for gaining a range of views (Catterall and Maclaran, 1997). Gibbs (1999) argued that it may be useful to identify salient ideas. This is why the nominal group technique involves a system of voting that is omitted in the focus group.

Content Analysis

Content analysis is a qualitative method of data analysis. It consists of reading the documents to be analysed, categorising words together (researchers must be vigilant in how they group words together) and undertaking frequency counts for each category (Cohen and Manion, 1994). Morgan (1988) says that focus groups (mentioned above) can involve a phase of 'content analysis'. This is the part of the focus group procedure which Winter and Maisch (1996) were concerned about as it would involve the researcher developing an interpretative framework to analyse the data. Therefore they rejected it as a method for developing competencies for social work. Despite these doubts about content analysis Winter (1993, 1994) used it to analyse tutors' comments about students' work to identify what educators believed to distinguish one level from another. This study incorporated a variety of qualifications. The analysis was not used to produce grade descriptors. This process was described in the previous chapter. The same method could be applied to identifying the distinguishing characteristics of grades or other types of levels of attainment.

Network Analysis

Network analysis is a qualitative method for analysing social episodes, unlike many of the methods described in this chapter it is not a method of data collection. An episode may be a child entering and later leaving primary school or people describing their experiences of receiving a job interview. Such accounts explain our past, present and future actions. In this type of analysis a relational network is used to represent the content and organisation of a person's knowledge of a domain. Qualitative data are classified in a manner that retains the complexity of the phenomena under investigation. The interdependency of the categories is represented as a network showing the relationships between different categories. There should be some kind of consensus about the network so that it is reliable and it should also be valid. The network should be clear, complete and consistent. It is useful for dealing with a great deal of complex qualitative data (Cohen and Manion, 1994). Winter and Maisch (1996) considered and rejected network analysis as a method of capturing the details of social work practice. They argued that the logic of network analysis is cyclical. Once the network analysis is completed competencies would still need to be derived from the results. This approach is for eliciting relational knowledge and the reasons behind actions. This means that it may be appropriate for profiling the knowledge content of qualifications, or qualifications which are for testing knowledge e.g. writing an essay about Hitler, rather than actions or skill based competencies, e.g., building a staircase. (Of course this does not mean that action or skill-orientated qualifications do not involve a knowledge component, e.g., there is a great deal of knowledge involved in building a staircase.) Network analysis is not suitable for giving an account of behaviours and tasks to be undertaken. In this way it is different from functional analysis which makes statements about tasks and actions.

Position Analysis Questionnaire

The position analysis questionnaire can be task or worker focused and there is a potential for quantitative analysis. But in standardised questionnaires, like the position analysis questionnaire, there can be an abstraction from the jobs that are undertaken, which does not occur with some other methods. There are 194 job elements in the position analysis questionnaire that a trained interviewer uses to rate different jobs. In the UK the questionnaire can be obtained from Oxford Psychologists Press. The information provided by job holders is compared with the similar analyses of most jobs in the economy. This can be used to compare a particular job with other jobs. The questionnaire and analysis are based on underlying job dimensions. For example, 'Job context

1 Stressful/unpleasant

2 Personally demanding situations

3 Hazardous job situations' (Kandola and Pearn, 1992, 47).

Kandola and Pearn (1992) did not give details of the questions that are asked of job holders. Although this questionnaire has a strong statistical grounding it is unlikely to identify unique, unusual and /or new competencies. However, it can be used to cluster diverse jobs into job families based on the human and psychological demands of the occupations (Kandola and Pearn, 1992).

Work Profiling System

The Work Profiling System was developed by Saville and Holdsworth Ltd. It constitutes three different questionnaires for different types of jobs

1) Managerial/professional

2) Service/administrative

3) Manual/technical

This procedure involves a self-completion questionnaire and a validation interview with an analyst. But Kandola and Pearn (1992) did not give an indication of the type of questions that are in the questionnaires. It can be used to produce job descriptions and to profile jobs and human attributes. It provides a framework in which defined competencies can be developed. It is recommended that it is used alongside less structured methods of job analysis (Kandola and Pearn, 1992).

Conclusions

This chapter has illustrated that there are a variety of methods available for articulating profiles of content and demands of qualifications. In some cases it may be appropriate to describe actions, in which case functional analysis may be useful, whereas in other situations it may be useful to describe the qualities that are exhibited by individuals and in this scenario KRG may be an appropriate method. The choice of method is determined by the objectives of the exercise. This review has illustrated that not only is the method used determined by the type of profile/outcomes to be written but that studies often benefit from a combination of methods to articulate the profiles/outcomes. Kandola and Pearn (1992) and Winter and Maisch (1996) also advocate a combination of research methods.

This literature review of 'methods of articulation' includes methods for data collection and analysis and methods that are really a combination of data collection and analysis. The Delphi technique, KRG, Q sort, Nominal group technique, focus group and critical incidence analysis are methods of data collection. Mastery levels analysis, content analysis and network analysis are methods of data analysis. Obviously appropriate data collection and analysis methods need to be used together. For example, functional analysis is both a data collection and analysis method. Data is collected by asking practitioners questions and this data is analysed to develop a functional map. Working parties are difficult to classify as either a method of data collection or analysis as the experts in the working groups can use any data collection or analysis methods that they prefer or they can adopt a non research based approach to the articulation of contents and profiles.

In response to any description of content and or outcomes we can ask the question 'whose content and outcomes are they?' For example, Winter and Maisch (1996) used social work practitioners to develop competencies about social work, but are the practitioners the best people to write the competencies? In developing any profiles of content or outcomes the developers should consider who should be consulted to develop the descriptions. In response to any proposal to articulate the contents and demands of qualifications we can also ask 'Is the articulation of contents and demands worth the time and resources spent in this activity?' In a THES article about Wolf's views about learning outcomes competencies, etc., she points out that in Higher Education more time is being spent writing specifications than making sure that users understand them and also that academic judgement cannot be reduced to list (Leon, 2000). It is hoped that this inappropriate balance in time and resources is not experienced in future projects to develop profiles of contents and demands.

There are four stages to the articulation of profiles of the demands and content and /or outcomes:

- describing the profile of contents and demands from a cart blanche situation (e.g. first stage of KRG);
- coming to a consensus or summarising data (e.g. second stage of KRG);
- developing profiles of demands and contents from the research results;
- validating the profile of demands and contents.

Some literature does not make explicit from where some of the information used to develop profiles of demands and contents originated. For example, where did Hadfield's (1980) statements about CSE art attainment originate? Some of the studies reviewed here include a validation stage, e.g., Greatorex (1999b), but others do not, e.g., Greatorex (1998). Some authors like Hughes et al. (1998) acknowledge the advantages of a validation stage. Studies that do not have a validation stage might benefit from this final stage in the process.

In many of these studies where research methods are used to develop outcomes, the results of the study need refining before they can be considered to be competencies, learning outcomes or grade descriptors, etc. Some suggestions have been made for refining the results of studies. There are methods like Q sort which are useful for developing outcome statements but which need some kind of statement to begin the procedure. It may be useful to consider methods of developing statements to be used in Q sort, e.g., the first stage of KRG.

A range of methods has been considered for articulating outcomes. This may be due to the trend towards using outcomes. This suggests that there is room for more consideration of methods to describe the contents and demands of qualifications.

Review of research undertaken comparing qualifications

In this section, research undertaken comparing qualifications will be described. Comparative research has two main strands. In the first strand, which is usually restricted to Type I comparability, the objective is to investigate whether two qualification are equivalent. In the second strand, interest is focused on describing the similarities and differences of two qualifications and it is not necessarily expected that the qualifications are equivalent. The research in this chapter relates more to the first strand rather than the second. There has been a great deal of research into Type I comparability for UK examinations. Most of this chapter will use terminology appropriate for academic school examinations. This does not mean that the techniques described here are restricted solely to these qualifications.

Type I comparability of academic school examinations has been reviewed twice before. Bardell and Shoesmith (1978) reviewed that comparability studies undertaken during 1964-1977. This review covered 30 studies. Forrest and Shoesmith (1985) extended the review by considering the changes that had affected the examination boards and reviewed the 20 studies that have taken place between 1975 and 1983. Both these reports concentrated exclusively on studies relating to within subject comparability undertaken by the boards. The objective of this report is much broader and it is not the intention to review comparability studies with the same level of detail. A list of the comparability studies that have been undertaken by the awarding bodies since then has been included as an appendix to this report.

Since these reports there have been a number of changes that are having a considerable impact on the conduct of comparability studies. Firstly, the level of regulation has increased with syllabuses having to obtain official approval and examinations being subject to a mandatory code of practice. This has the led to a reduction into the number and diversity of syllabuses within a subject. Secondly, the move to greater accountability in education, including national curriculum testing and performance tables, has led to the development of linked databases of test and examination results. This means that the resources to carry out some forms of statistically-based comparability studies are readily available.

There are five generic approaches to this problem:

1. Using measures of prior outcomes
2. Using measures of concurrent outcomes
3. Using measures of subsequent outcomes
4. Comparing performance of candidates who have attempted both qualifications *at the same time.*
5. Expert judgement of the qualifications

Although the first three are related and the same methods of analysis can be used to investigate the problem, they have been separated because the advantages and disadvantages are different. The word outcomes has been deliberately chosen so that it covers the results of a wide range of measures including tests of aptitude, achievement, subsequent job performance.

In the next section, three generic methods using prior measures will be described. This is followed by a section on comparing candidates who have attempted both qualifications at the same time. Methods involving the judgement of experts are considered in the next section. This is followed by a discussion of the problem that arise when the subject or subject area is not the same. Most of this research relates to the controversial area of subject difficulty.

Methods involving measures of outcomes

For these methods, the statistical methods used to analyse the data can be the same but the interpretation of the results is different. It is useful to consider what the exact wording of the study results is and how inferences about comparability can be made from each of the first three methods. Assuming that there is no difference between two qualifications, then the results of each study and the assumptions needed to make a valid inference are given in Table 1. These issues have also been considered by Jones (1997).

Table 1: Results and assumptions of generic approaches to comparing qualifications

Method	Strict meaning of results	Some assumptions required for comparability
Using prior outcomes	The measures of prior outcomes are, on average, the same for the candidates who have obtained both qualifications at a particular level/grade.	If assessing knowledge and skills, then relative/absolute* progress in obtaining them must be the same for both qualifications. If assessing potential, it must be stable over the period between obtaining the prior outcome measure and obtaining the qualification.
Using concurrent outcomes	The measures of concurrent outcomes are, on average, the same for candidates who have obtained both qualifications at a particular level or grade.	The attainment is the same only for the skills, knowledge and/or potential assessed by the concurrent measure. The qualifications could differ on other aspects of attainment.
Using subsequent outcomes	The measures of subsequent outcomes are, on average, the same for candidates who have obtained both qualifications at a particular level or grade.	There is a causal relationship between achievement required to obtain the qualification and subsequent outcomes. The subsequent outcomes have not been influenced differentially by subsequent events (e.g. the holders of one qualification getting additional training courses).

*for absolute progress, the measure of prior outcomes produces a score on the same scale as the qualifications.

Comparability studies using a common outcome measure are usually carried out by considering how the relationship between examination performance and the common measure of attainment varies by syllabus (for the purposes of analysis it is sensible to separate syllabuses within boards and possibly options within syllabuses). The simplest models for this type of data involve converting the examination grade into a score and analysing it as a continuous variable using linear regression or, more recently, a linear multilevel model. Although this leads to the simplest models, this approach is unsatisfactory because of the presence of ceiling and floor effects. Further technical details are given in Appendix B. This means that the values of the residuals and regressors will be correlated which can result in biased estimates of the regression coefficients (McKelvey and Zavoina, 1975). In addition, Winship and Mare (1984) noted that the advantage of ordinal regression models in accounting for ceiling and floor effects of the dependent variable is most critical when the dependent variable is highly skewed, or when groups defined by different covariate values (e.g. dummy (0,1) variables for each syllabuses) are compared which have widely varying skewness in the dependent variable. This situation does occur in examination databases because some syllabuses attract entries that are, on average, more able than other syllabuses. Goldstein and Cresswell (1996) criticised Fitz-Gibbon and Vincent, (1994a) for fitting a straight line to the relationship between total A-level score and mean GCSE score. They noted that (Goldstein and Thomas, 1995) found that the relationship was non-linear.

For an individual examination, it is more appropriate to consider an examination grade as an ordinal variable (Fielding, 1999). These can be fitted using the proportional odds model. Although this model solves the problem of floor and ceiling effects by fitting a series of s-shaped logistic curves, there are a number of disadvantages. The proportional odds model is one example of a generalised linear model. There are some problems in using this model. Firstly, parameter estimation in generalised linear models is more complicated than in linear models. This is a particular problem when the multilevel structure is included in the model. However, the main problem with the proportional odds model is not computation but the assumption of identical log-odds ratio for each grade i.e. the relationship between probability of obtaining a grade and the outcome measure is the same for each grade. Violation of this assumption could lead to the formulation of an incorrect or mis-specified model.

To test the assumptions associated with a proportional odds model, the dichotomised response variables based upon cumulative probabilities (e.g. obtaining at least a grade C and not obtaining a grade C) should be analysed separately. Due to the stringent model assumption the proportional odds model is the wrong method to *start* a valid data analysis of this type of data. The basic assumption of all logistic models that the logits and the covariates are linearly related can only be checked by using dichotomized responses. It is only after linear relationships between the logits and covariates have been established in the separate binary logistic models that it is reasonable to fit a proportional odds model.

Because separate binary regression models are required for model checking and model building in any case, is there any need for the proportional odds model? The use of separate binary logistic regression models does have some disadvantages. This approach can lead to final models with different sets of covariates for different grades making interpretation difficult (e.g. a sex difference at one grade but not at another though this finding is inherently interesting). Categories at the ends of the scale may have very low or very high probabilities, and parameter estimates may not be statistically significant due to less power, i.e., the ability of a statistical test to detect differences (usually this will not be a problem because the data sets used in comparability studies are large). The proportional odds model, when it fits, is superior because few parameters are fitted. This is better because the standard errors of these parameters are smaller. However, if the data set is large, the standard errors for the separate binary logistic models are likely to be small (meaning that there is sufficient power to detect small differences between syllabuses). In comparison with the proportional odds models the presentation of the results of binary logistic regressions is much simpler and more interpretable for less experienced users.

All data involving qualifications has a multilevel structure in which candidates are nested within centres. For most educational research this is a feature of interest but for the purposes of comparability it is a nuisance. The nested structure adds to the complexity of the data. There would be considerable advantages if it were reasonable to ignore this structure because standard statistical software could be used and the computational requirement would be lower. This is a serious issue for analyses involving the entire entry of some qualifications. For linear regression there has been a substantial amount of research into the effect of ignoring the multilevel structure of data. For example, Tate and Wongbundhit (1983) confirmed by simulation research that the estimates of regression coefficients are unbiased but have a much smaller sampling variance than the parameter estimates for a multilevel model. Ignoring the multilevel structure of the data would produce spuriously significant effects. Kreft (1996), reanalyzing results from Kim (1990), estimated that ordinary least square estimates are about 90% as efficient as ML estimates. Kreft concluded that if researchers in the social sciences are interested in the estimates of the regression parameters, the results of multilevel analysis will be close to the results obtained with more traditional regression techniques. In both cases the fixed effects estimates are unbiased. The main difference is in the standard errors of these parameters, which are estimated too small if intra-class correlation is present in traditional regression analyses. This fact makes the random coefficient model more conservative than the traditional regression. If researchers are interested in co-variance components, the random coefficient model, and the available software for the analyses of hierarchically nested data, will be a good choice. Unfortunately there are two problems with Kreft's conclusions. Firstly, they depend on the intra-centre correlation being fairly small (this is usually the case with educational data; candidates within a centre usually differ in performance). Secondly, ignoring the structure would not be acceptable to proponents of multilevel models who would object to the simplification. Note that the argument that it may be safe to ignore the structure applies only to the type of analysis used in comparability studies and not for the type of analysis used to investigate the value-added by individual schools where the multilevel structure is obviously of interest.

It is feasible to use a series of binary multilevel logistic models to investigate comparability although this would be time consuming. Further research comparing single level logistic regression models and multilevel models is necessary. For large data sets, a sensible strategy would be to fit single level models first and then if important differences are found, to investigate further using multilevel models. This means that it is necessary to decide on how large a difference is important. For examinations with large entries, the results of this type of analysis include statistically significant results that are not in practice important. This issue can be investigated by considering the mark distributions of the examination and considering how the percentage pass rate changes with each mark around the boundary under consideration. For some of the studies with outcome data, new data sets are generated every year. It would be sensible to consider data from more than one year to investigate whether the differences were consistent.

All methods involving such measures can potentially result in the situation that the relationship between examination performance and the measure can vary for differing groups within the examination entry. For example, it is possible that male candidates outperform female candidates for one syllabus after allowing the measure, but that the reverse can occur for another syllabus. If this occurs then the conclusion is that the syllabuses do differ but it is not solely a difference in grading standard but something related to the syllabus and the style of assessment.

Given that most qualifications use ordinal outcomes, it is clear that logistic regression as described above should be used in preference to ordinary linear regression. In much of the research reviewed below, ordinary linear regression was used because logistic regression had not become generally available as an analysis technique. In the next three subsections, research involving the three different types of outcome measures will be considered.

Using measures of prior attainment

As a result of the greater interest in school effectiveness and improvement, there has been an increase in the use of value-added measures, this has had lead to the development of linked databases of educational test/examinations results (e.g. KS3/GCSE database, 16+/18+ database). These databases enable the relative progress of candidates (in most cases, it only possible to measure relative progress and not absolute progress because the two measures will not be on the same scale) to be assessed by various groups of individuals. One problems is that the progress made by different identifiable groups within an entry may have different rates of progress (e.g. Haque and Bell, 2000) and if the composition of the entries of syllabuses vary then it is possible that difference between syllabuses may be exaggerated.

Initially analyses were based on databases compiled by providers of value-added information. For example, Tymms and Vincent (1995) used data from the A-level Information System (ALIS) (Fitz-Gibbon, 1992a and b). Although the details of their statistical analysis can be criticised, they found that there was no case of a Board being significantly severe or lenient for all three years for which considered. Taverner and Wright (1997) considered the relationship between GCSE results and modular and linear A-level mathematics. They found that A-level candidates who sat the modular syllabuses tended to obtain better results than candidates who sat the linear options. They explained this in terms of the modular structure leading to more effective learning and higher attainment rather than in terms of differences in achievement.

Since the matched data for Key Stage 3 and GCSE has become available, it has been used in the more recent inter-board comparability studies and there are proposals for using this data for identifying potential differences in the major GCSE subjects. Because there is no way of distinguishing between a difference being caused by differing grading standard or by a syllabus facilitating greater attainment. This generic method is best used for screening purposes to identify syllabuses that are performing differently for further research.

Using measures of concurrent achievement

Historically, concurrent measure of outcomes have been extensively used in comparability studies (e.g., Schools Council, 1966; Nuttall, 1971; Willmott, 1977). In more formal psychometric equating, reference tests are referred to as anchor tests. The problem with this approach is that the outcome depends on the relationship between the examinations and the test. A reference test would penalise a board that did not include some of the content of the reference test. This could, of course, be regarded as a valid outcome if it indicated that the examination did not meet a particular subject specification. Christie and Forrest (1981) pointed out that there are three ways of measuring concurrent performance have been used to assess comparability within subjects. Firstly, there are references tests that measure 'general ability', 'aptitude', or 'calibre'. Secondly, there are studies using subject-based reference tests. Finally, a common element can be included as part of all examinations. The choice of reference test could reflect the type of comparability as described in the introduction.

They reported the results of a comparison of the effects of using an external aptitude test and an external achievement test. This came from an Advanced-level Physics comparability study (University of London Entrance and School Examinations Council, 1972). Although the authors note that the results are not definitive, they suggested that different conclusions could be drawn depending which test was used. There findings were supported by a study of comparability between Mode I and Mode III examinations in Geography and Biology in two CSE boards (Bloomfield, Dobby, and Duckworth, 1977). They used a subject reference test and an aptitude test, Test 100, which was a specially devised ability test. They report estimates of severity or leniency for each school based on each of the two tests separately. The correlation coefficient for the two sets of estimates was 0.73 for Biology and 0.82 for geography.

External tests have also been used to monitor changes over time. Murphy et al. (1996) describe a study of conducted by the NFER using 'Test100' to monitor the standards over time. This research considered CSE and GCE examinations during the period 1968-1973 ((Willmott, 1977; Willmott, 1980). This research was heavily criticised (e.g., Wood, 1976) with most criticism relating to the nature of the test. An important argument put forward related to the fact that any test which succeeds in measuring general ability is bound to relate variably to different subjects which obviously make different cognitive demands. Using general ability reference tests to investigate the comparability of various public examinations makes the assumption that public examinations are just measure general ability (Cresswell, 1996). The researchers, aware of these concerns, included in their report an account of differences between schools, the limitations of Test 100, and the unsuitability of their statistical model for the data that they collected. Attempts were made to improve and update Test 100 in 1975 and 1976 but these were considered to be a failure (Forrest and Shoesmith, 1985).

Despite this history, some researchers have persisted with this approach. Recently Fitz-Gibbon and Vincent (1994) and Tymms and Vincent (1995) have looked again at the reference test approach. They used The International Test of Developed Abilities (IDTA). This was a test developed at the Educational Testing Service in the US as a measure suitable for the selection of college entrants around the world.

A good illustration of the problem of how different reference tests can lead to different conclusions is the research of Flynn (1986). He considered the changes in three American tests, SAT-V, WAIS, and the armed forces mental test, over a long period of time. He noted that when subjects take a mental test, it measures their problem-solving ability through a vehicle. If the vehicle is academic skills, and if the skills are in decline, problem solving gains must overcome academic skill losses. For example, a subject simply cannot perform well on the SAT-Verbal (SAT-V - the American examination) sub-scale without the advanced academic skills taught in high school English courses. However, other tests avoid the use of academic skills. The Wechsler IQ tests require no more than elementary academic skills, and some performance subtests minimise the need even for these. The content of the armed forces tests includes simple arithmetic, both presented as such and verbally, word knowledge, and paragraph comprehension, all of which are on an elementary school level (Korb, 1982).

Between 1963 and 1981, trends on these three mental tests were as follows. The SAT-V score declined by about 4.32 IQ points. The WAIS data for young adults gave a rise 3.33 IQ points. The armed forces data suggest no gain during the same period. Thus there was a gain on a test of problem-solving ability with a moderate reliance on reliance on elementary academic skills, no gains on a test with a heavy reliance on elementary academic skills, and a loss with a test with a heavy reliance on advanced academic skills. This means that great care should be taken in interpreting the relationship between changes over time and different tests.

More recently, UCLES Research and Development Division has developed a reference to investigate the comparability of international and UK examinations at the same level (Dexter and Massey, 2000). This research demonstrates that the relationship between the examination performance of different groups within examinations makes simplistic interpretations misleading and inappropriate.

Studies that involve the use of a subject based reference test tended to be rarer. This is an inevitable consequence of higher costs of constructing tailored subject based assessment instruments. These studies include a study of the standards in CSE and GCE O-level English and mathematics (Wrigley, Sparrow, and Inglis, 1967), A-level Physics, (Newbould and Shoesmith, 1974) and Biology and geography item banks (Bloomfield et al., 1977). Newbould and Massey (1979) argue that it is likely that subject-biased tests may be more likely than aptitude tests to prove to be biased. For example, Bloomfield et al. (1977) removed from their item banks any item that was not covered in all the syllabuses. This procedure is not satisfactory because a reference test constructed from such items may cover most of a 'narrow' syllabus but only a small portion of a 'wide' one. Another problem is that the set of items may reflect the main emphasis of one syllabus and a peripheral part of another. It should also be recognised that any given examination represents a represent selection from the syllabus. This means that the relationship between the examination and reference test could change from year to year.

A solution to the above problem would be to use a comprehensive reference test. All tests and examinations must because of time constraints represent a selection of the content of their syllabus because of time constraints. If an extremely long test were developed that allowed a profile of scores to be generated, it would be possible to compare how the entries of two examinations differed on these profiles (e.g. whether one syllabus was 'narrow' and another 'wide'). The problem with doing this is that such a reference test would take too long to administer. A solution to this problem would be to use a survey design such as that used in the Assessment of Performance Unit's surveys. The APU survey used multiple matrix sampling (Johnson, 1989; Johnson and Bell, 1985). The components of the reference test are administered to different random samples from the entry. The second difficulty with this approach would relate to the generation of the reference tests, which would obviously be time-consuming and expensive. There is another advantage with this approach in that if it were to be repeated it would provide a means of monitoring standards over time. It also has the advantage that it can identify areas of weakness in the profile of performance and provide useful feedback to the teaching profession.

A special case of concurrent achievement is when a common paper is used in different boards' examinations especially for a comparability study. Shoesmith, Newbould, and Harrison (1977) attempted this by setting a common prose test in A-level French. They found that this was a more complex task than using an external subject-based monitor test. Newbould and Massey (1979) summarised the experience of the Test Development and Research Unit (TDRU) of using information from 'common tests' forming a part of different examinations. They described four comparability studies, two concerned with A-level economics and to with O-level French and reviewed the complexity of the method.

Although qualifications are rarely considered directly, research in comparative education raises some appropriate issues in the use of reference tests. This is because achievement is been considered to be important since Marc-Antoine de Jullien first proposed his comprehensive schema for study foreign educational systems in 1817 (Fraser, 1964). The first large-scale international comparative study was organised by the International Association for the Evaluation of Educational Achievement. In addition to the achievement testing, the studies organised by the IEA include in-depth analysis of curricula, school organisation, teacher and student attitudes and background and socio-economic indicators. The first quarter century of the IEA's work is described by (Postlethwaite, 1987). Some of the problems that have been identified in these international studies, such differences in syllabus coverage and time spent on each subject, which are relevant when considering qualifications from different systems. For example, Freudenthal (1975) observed that the amount of variation in the mathematics curriculum and the way content is approached differs substantially between countries. Differences in coverage have been shown to be the strongest factor explaining national differences in mathematics surveys (Husén, 1967; Burstein, 1992).

The problems associated with use of reference tests are as follows. Firstly, an appropriate test has to be chosen or created. The choice of the test determines the way in which the qualifications can be said to be the same or different. In addition, there are problems in administering such tests at a time when centres may be busy with other high stakes assessment.

Subsequent achievement

For vocational qualifications, subsequent achievement is the main criteria for assessing validity. For academic achievements, there has been some research on this issue. Some research has been carried out into the adequacy of the preparation of science GCSE for A-level (Dearing, 1996). In addition there has been a small amount of relevant research has been conducted on relationship between entry qualifications and degree success. This has been carried out primarily into the correlation between A-level grades and degree classification (Most departments would entries that were too small to consider the effect of different syllabuses). Early studies found an overall correlation of around 0.3 in the 'old' universities (Bligh, Caves, and Settle, 1980; Sear, 1983). Bourner and Hamed (1987), after collating the data from 1983 for all graduates from universities, polytechnics and colleges, obtained correlations of 0.2 or lower for disciplines offered. In a meta-analytic study covering research up to 1983, Peers and Johnston (1994) found an overall correlation of 0.28. Given that A-level grade is used a selection criteria for universities, the low correlations are likely to be due to the low variation in the A-level score within each departmental entry.

There has been some research into the relationship between type of qualification and degree performance. This research is relatively recent because a growing number of university entrants have qualifications other than A-levels. Bourner and Hamed (1987) found that there was little difference in the final degree performance of those with qualifications other than A-levels compared to those with A-levels. There were differences between candidates with different types of non-A-level qualifications. The performance of graduates with HNCs, City and Guilds, and ONC's being associated with higher degree performance than HNDs and Scottish Highers. Clearly, the range and uptake of non-A-level qualifications has changed since this study. More recently, this issue has been investigated by Hoskins, Newstead, and Dennis (1997) who used the computerised records of a large university. They found that students with A-levels tend to obtain better degree results compared with other qualifications but that the difference was marginal. However, they found a significant interaction effect between age and entrance qualification. Students aged less than 25 tended to do worse when they had other qualifications but students over 25 did better. They suggested that this finding was the result of differences in motivation. It does, however, illustrate the importance of considering the interactions with other factors to avoid making simplistic and erroneous conclusions about comparability.

There are a number of problems with this type of research. Firstly, by its very nature, the results describe difference in comparability some time in the past. Secondly, there is a catch-22 situation resulting from differences in parity of esteem. Selection means that candidates can only demonstrate their qualifications are comparable if the are allowed to take the course and will only be allowed on the course if their qualifications are comparable. Thirdly, the qualifications can be different but the candidates with some candidates might have to work harder and/or receive additional support to make up for the differences. Finally, it could give to much emphasis to one particular learning pathway, e.g., not all A-level mathematics candidates go on to study mathematics at university.

This approach should be considered in three circumstances. Firstly, when there are concerns about how good particular qualifications are for a particular course. Secondly, when the objective is to investigate the potential of various qualifications. Finally, for vocational qualifications, subsequent performance in the particular vocation determines the validity of the qualification so this approach is the obvious method (particularly, for different versions of a qualification in the same area).

Comparing performance of candidates who have attempted both qualifications at the same time.

This section will be restricted to a discussion of the situation where the subject or subject area is the same. When the subjects are different, then the nature of the problem changes and this is described in a letter section. In most cases, individuals are unlikely to deliberately attempt to obtain two qualifications in the same area at the same time (for some qualifications, timetabling considerations would make this impossible).

Bardell, Forrest and Shoesmith (1978) describe research carried out on 1966 summer examination results which involved a comparison was made of candidates who sat the same subject with two different boards and found that there were no major differences in standards. This type of study was not repeated because at the time it was considered that the candidates who entered the same subject were atypical and were often on the borderline. They perceived that this approach had many limitations. One major problem is that, in the first instance, they argue that the candidates with a dual entry are atypical. Another problem is that candidates are not necessarily consistent in their performance. As part of the investigation, examiners were asked to scrutinise the scripts when the results were discrepant. They found that several candidates performed quite differently on their dual entries.

This means that it may be necessary to use an experimental design. It is likely that the candidates have been prepared for one qualification (the primary qualification) and not for the other (the secondary qualification).

This generates the problem that they are less likely to perform well on the qualification that that they have not been prepared for (e.g. because of motivation, familiarity with assessment style, differences in content). To overcome this it is necessary to use a sample of candidates from both primary qualifications. If the subsequent results are consistent with both samples tend to obtain better results for one of the qualifications then this strong evidence of lack of comparability. If the results tend to be same for both samples, then the qualifications are comparable. If the results tend to be the same for one qualification and one samples but better or worse for the other for other sample, then there is a lack of comparability (presumably related to coverage). If the results are inconsistent with each sample doing better on its primary qualification then it is difficult to assess the results. Finally, there is logical, but hopefully very unlikely in practice, result that both samples tended to perform better on the secondary qualifications.

There are major difficulties with the fourth generic method. Firstly, both assessments have to be made at the end of the course and candidates would be probably be unwilling to take additional assessments at that time. If the assessment were not made at the end of the course, then differences could be generated by differences in syllabus structure. Secondly, this approach could only be applied to the examination-based components and not to coursework.

Expert judgement about qualifications

The final generic method is based on the judgement of experts. This type of study is based on the assumption that experts can be found who, on the basis of their professional judgements and despite the numerous differences between examinations, can decide from a scrutiny of scripts, whether comparable grades are being awarded by boards to candidates of comparable levels of attainment. In their review, Bardell et al. (1978) identified to two kinds of this type of study. In the first, identification of 'borderlines', the examiners of one board are provided with a range of scripts covering the grades on which decisions have to be made. They then have to decide where to draw each grade boundary. This is; in effect, replicating their role in an awarding meeting. In the second kind the examiners are given a sample of scripts on or about each borderline and asked to judge whether the original decision was lenient or severe, or correct.

The problem with both these kinds of studies is that they require that the judges have some notion of what the standard should be. This means that the judges have to be examiners or have to be trained so that they know what the standard is. They have to have criteria by which the grade boundaries are to be judged. There are three different alternative sets of criteria. Firstly, judges can use their own criteria. Secondly, they can use those of the board whose scripts they are considering. Finally, they can use special criteria for the exercise itself. Experience from several of these types of study indicates the formulation of such criteria is difficult and applying them is even more so.

A problem with all studies involving expert judges is that it is time-consuming and expensive since the moderators have to be paid and reading scripts is a lengthy process. This means that the sample of scripts in any one exercise has to be small. There are also problems when the assessments based on coursework form part of the assessment. This could be addressed by using pairs of moderators from more than one board. These methods do have an important incidental advantage. Discussing one another's examinations techniques, marking and awarding procedures, and the criteria for grading is a useful exercise for practising examiners. This type of study has not escaped without criticism. Johnson and Cohen (1983) pointed out that the method of cross moderation has come in for much criticism for lack of rigour and, in particular, failure to come up with conclusive results.

More recently, Thurstone paired comparison methodology has been used to investigate comparability. This method is described in the following papers which describe the methodology applied to monitoring standards over time rather than comparability (Bell, Bramley, and Raikes, 1997a and b; Bramley, Bell, and Pollitt, 1998). In this approach, the judges are asked to decide which of a pair of scripts is better. This has the advantage that differences in the judges' notional standards are less important. Differences in severity and leniency of notional standards have no effect on the outcome of paired comparison. This means that external judges who are familiar with subject area but not with the particular notional standard used for the qualifications can be used in this study. This adds to the credibility of the study. Thurstone paired comparison methodology has been used for comparability studies carried out by the awarding bodies (Forster and Gray, 2000).

An important issue with this type of study relates to the choice of scripts. In the recent comparability studies, the scripts used were all chosen to lie on the borderline. In the monitoring over time study, the scripts were chosen to cover a range of marks, which allowed the magnitude of the difference to be considered. The design process for this type of studies is complex and has been discussed in greater detail in Bell (2000).

For comparing essays, one methodology that could be considered is the use of computerised essay marking. Although this has never been applied to a comparability study, this methodology does have some potential advantages. It would be possible to compare many more scripts because the costs would involve data entry rather than expert judgement. There are the issues of how the expert system is calibrated to mark the scripts and the problem of credibility. However, it could be used as part of a screening process to identify situations that could be investigated further by expert judges.

In addition to studying scripts, the expert judges can also be used to identify differences in the structure of syllabuses, question papers and mark schemes. Obviously a report on this aspect can be generated by convening appropriate meetings and have appropriate discussions. There are, however, more structured approaches. For example, Kelly's Repertory Grid could be used (described in chapter 2). Gray (1999) used this approach in a comparability study of Science GCSE. Scrutineers were asked to compare pairs of syllabuses, question papers and mark schemes (as a package) and to explain how they were similar and different. In a group discussion, the examiners brought constructs together into bipolar statements. Then a questionnaire was developed from the constructs which corresponded to how the syllabuses, question papers and mark schemes differed from one another. The questionnaire was sent to a wider group of scrutineers who were asked to rate the syllabuses, question papers and mark schemes on the bipolar 7 point scales. The ratings were averaged over scrutineers. This study illustrated in what ways the scrutineers considered the syllabuses, question papers and mark schemes to be different. It was coupled with a cross moderation exercise. Kelly's Repertory Grid was used to give these measures a context by describing the characteristics of the different syllabus packages and giving a numerical estimate of the relative extent to which each syllabus package was associated with these characteristics.

It is important to recognise that studies based solely on a comparison of the syllabuses and examination papers are unsatisfactory. It is necessary to consider examples of candidates' work. For example, one examination may seem to contain some harder questions than another examination but it may be the case that few candidates who enter examinations are able to answer the harder questions.

Of the five generic methods, expert judgement, if properly executed, is likely to be the most useful in determining the comparability of different syllabuses. Unfortunately they are too costly to be used extensively.

Between subject comparability UK Examinations

So far in this chapter, the discussion has largely been focussed on comparability between subjects of the same type. There has also been research into the comparability when the subjects are not the same. This research has usually been initiated because of concerns about subject difficulty. Nuttall, Backhouse, and Willmott (1974) suggest that this is because the concept comparability is least definable here. However, they go on to suggest that they "can see no logical reason why, if a *large group of candidates representative of the population* took, for example, both English and mathematics, there average grades should not be the same." They further argued that, in the case of English and mathematics, their importance for future opportunities mean that, as a whole, candidates would try equally hard in both subjects. This might be debatable for other pairs of subjects. They also argued that there is no reason to suppose that teachers in one subject are better than teachers in another (and it would require strong evidence before such a justification could be used in practice). They also proposed that although some individual will be better at mathematics than they are at English it can be expected that they will be balanced by individuals who are better at English than mathematics.

In some instances, subject difficulty was addressed in the initial setting of the standards. For CSE, grade 4 was defined as 'a 16-year-old pupil of average *ability* who has applied himself (sic) to a course of study regarded by teachers of the subject as appropriate to his age, ability and aptitude, may reasonably expect to secure grade 4.' For each subject, the standards are related to the attainment of the average 16-year-old. A similar definition was adopted for the development of National Curriculum levels. The main difficulty with such a definition is that most subjects are only taken by a subset of the population, the average attainment in the subject cannot be determined directly.

Nuttall et al. (1974) argued that differences between standards in subjects matter. They supported the argument with the following example:

'If an employer were faced with two boys (sic), each with five O-level passes, he would have no basis on which to choose between them in terms of attainment. If, however, both had passes in English and mathematics, and one had passes in English Literature, biology and French while the other had passes in history, physics and Latin, it is highly probably that the latter's general level of achievement would be higher than the former's, on the basis of the results presented later in this report.'

Obviously, in a real situation, the general level of attainment might be less important than the performance in specific subjects. More recently, the issue of subject difficulty has increased relevance with the greater emphasis on standard setting and performance tables for schools. It is questionable that is educational desirable that candidates could be channelled into particular subjects on the basis of perceived subject difficulty. Using the results of subject pairs comparisons as an index of severity, Christie and Forrest (1981) investigated the effect of severity on subject uptake between 1966 and 1973. They found that for every one of the subjects that the number of entries had fallen was classed as severe by the subject pairs comparison. Similarly those subjects with increased uptake tended to be lenient. More recently, subject difficulty has also been raised in the context of effect on measures of value added by teachers and performance tables (TaylorFitz-Gibbon and Vincent, 1997). There is an obvious tension between adjusting for differences in subject difficulty and maintaining standards over time. If a difference in difficulty is accepted, then it is difficult to correct it without changing the grading standard of the examination. Goldstein and Cresswell (1996), however, argued that making the adjustments would be undesirable.

One of the methods that have been used to address this problem is subject pairs analysis. This technique is described in (Forrest and Vickerman, 1982). They argue that subject pairs data provides a firm base for discussion into the standards of entry of subjects. The limitation of subject pairs data is that for any given boundary the data used is from the central region of the Figure 1 (highlighted in bold). The numbers in the remaining cells of the diagram cannot be known.

			Examination Y		
		Unable to pass Does not enter	Unable to pass Enters	Able to pass Enters	Able to pass Does not enter
	Unable to pass Does not enter	No results for X and Y	No result for X Fails Y	No result for X Passes Y	No results for X and Y
Examination	Unable to pass Enters	Fails X No result for Y	Fails X Fails Y	Fails X Passes Y	Fails X No result for Y
X	Able to pass Enters	Passes X No result for Y	Passes X Fails Y	Passes X Passes Y	Pass X No result for Y
	Able to pass Does not enter	No results for X and Y	No result for X Fails Y	No result for X Passes Y	No results for X and Y

Figure 1: Choice and examination success matrix for two examinations

Although subject pairs have their limitations in absolutely determining subject difficulty, it can be argued that they have a use in monitoring the system. If patterns from subject pairs analyses were to change then it would indicate a need for further investigation.

Kelly (1976) described and used an iterative procedure to investigate differences in subject difficulty in Scottish Highers. This involves repeated calculations of a correction factor based on equating the mean pass grade in one subject with the mean pass grades obtained by the same candidates in all other subjects. Although this approach was only used to describe differences, Kelly noted that in Australia standardisation methods of this type were used by the states of Victoria and New South Wales. The reason for this is the university and college scholarships were awarded on the basis of aggregated marks in different matriculation subjects.

Reference tests have also been used to investigate subject difficulty as well as within subject comparability (e.g. Nuttall et al., (1974) and Fitz-Gibbon and Vincent, (1994a). The use of reference tests for this purpose has many of the same problems associated with use of reference tests for within-subject comparability. The main problem is that performance across different subjects is multidimensional and subjects will require different levels of performance on different dimensions. Given that the score on single reference test measures a single dimension, the differences in 'difficulty' identified by this test does not relate to that dimension. If candidates who pass subject X tend to obtain a higher score on a reference test than candidates who pass subject Y, it does not necessarily mean that another reference test measuring another dimension may not show the reverse pattern. Note that for one subject to be more difficult than another it is not necessary to assume uni-dimensionality, only that required performance required on all dimensions covered by both subjects is the same and in at least one case higher for one of subjects. The problem with assessing difficulty is generating a suitable profile of reference test results. Given that the Key Stage 3 tests results form a profile of three different scores, it would be interesting to investigate how the relationship between these score and GCSE performance varies from subject to subject.

The situation is easier when comparisons are restricted to similar subjects. It is reasonable to consider *appropriate* reference test results. While comparing the difficulty of English and mathematics seems to be questionable, concern about the relative difficulty of science subjects is more legitimate. If this argument is extended, it becomes apparent the problem is considering a network of relationships between subjects.

Although subject difficulty is a complex and controversial issue, it cannot be ignored entirely. For example, small differences in prior attainment as measure by average GCSE for different A-level subjects might be tolerable, if the differences increased in size then there might be some grounds for concern.

Conclusions

The five generic methods of comparability identified in this document can be used for different purposes. For UK examinations Type I comparability can be addressed by use the linked databases to carry out a screening process using statistical techniques. One way of investigating the problem further would be to use a study involving expert judgement. If a difference is found other evidence could be considered to all decisions about the action that could be taken. There are difficulties in deciding what this action should be. For example, if two syllabuses differ should the standard be changed on one, or both, or at all?

References tests have two useful purposes. They can be used when there are no existing measures of prior or concurrent achievement common to all candidates entered for the qualifications under consideration. Secondly, if there is interest in the overlap of particular knowledge or skills, then a suitable reference test would be useful.

Considering subsequent performance is very important is the emphasis on comparability is based on fitness for purpose. Studies of this type are also useful for addressing issues of parity of esteem.

Undertaking research into taking two different versions of the same qualification at the same time is difficult. This methodology has a very important role in test and examination development.

Although they are costly, studies involving judges are very important. Judges have the advantage that they can consider the importance of differences in observed attainment. Studies involving Thurstone paired comparison methodology can be designed to have results that have a high degree of credibility (external judges can be used and if judges disagree this can be identified in the results).

Finally there is the vexed area of subject difficulty. There are no easy solutions to this problem but it cannot be ignored completely. Large differences in difficulty (as measured by prior or concurrent attainment) for qualifications within the same level of framework would not necessarily be acceptable. This is particularly the case if the subject areas are similar. There is a need for further research into the multidimensional nature of qualifications. This research could supply valuable insights into the nature of various subjects.

Summary and conclusions

This report considers and reviews various aspects of research into comparability. In the introduction, it was noted that comparability is context dependent and that it depends on the use to which the qualifications are being put. For one set of circumstances, two qualifications can be considered to be comparable while for another set they are not. This has two important implications. Firstly, some of the methods that a reviewed in this report can only be used to assess some types of comparability. Secondly, when comparability is being investigated by the use of experts, the choice of expert should relate to the context of the comparability.

This report has been structured so that comparability is considered in descending level of detail. In the second chapter, research into the nature of levels and qualifications was considered. In the third chapter, research into profile and demands of a qualification was reviewed. The fourth chapter considered research undertaken in comparing qualifications, which is frequently based on the results of a particular administration of the qualification.

In the second chapter of this report, the research into qualification levels was described. Much of the research into this issue has been in higher education. The literature reviewed in the report indicated that levels could be conceptualised in different ways. Some authors argue that the complex process of learning cannot be represented by a simple level framework. In addition, progression in a framework can sometimes represent traditional programme structure and not the level of demand. The research also identified that it was important that there should be systematic attempts to match students work with level descriptors.

One important aspect that has been identified with some level frameworks is that level is determined by the changing relationship between tutor and the student with the student taking more responsibility for learning as the level increases. Whilst this can be applied in higher education, it is more difficult to apply to high stakes assessments administered by external bodies that have no control over the tuition process. There are also difficulties in assigning levels to learning outcomes as level is not just determined by content and learning process it is also influenced by context. So, for example, descriptive statistics may be taught to A-level mathematics students and also to Masters Research Methods students.

Some international frameworks have been reviewed in this report. A major reason for the development of such frameworks has been concern about life-long learning and flexible career pathways. For example, SAQA cite a conference paper by Sir Christopher Ball (1996). He argued that the kind of learner profile that is suited to the 21st century, would be for 'flexible generalists'. These are "people equipped with necessary knowledge, skills and values to adjust readily to multiple career changes and make, through their own personal development a significant contribution to the life of this country and the world." He went on to argue that this requires a shift in thinking from education for employment (i.e., developing the ability to do a specific job) to education for employability (i.e., developing the ability to adapt acquired skills to new working environments). This type of argument inevitably leads to issues about comparability of qualifications in different subject areas.

The international qualification frameworks reviewed cover the whole range of qualifications and not separate sectors of education (although this had led to criticisms). An important feature of these frameworks is importance of obtaining a consensus when developing a framework. In the UK, there are sectoral boundaries between frameworks and this has had unfortunate consequences in terms of definitions. For example, levels are defined differently for different types of qualifications. The word level in NVQ terminology indicates the complexity of an occupational role while in higher education it is associated with intellectual demand and/or autonomy.

Research from higher education has shown that considering level frameworks is time consuming and there can be severe problems if too much is attempted in too short a period. In addition, there are arguments that generic level descriptors can be too vague to be useful. Research is needed to identify whether level descriptors developed without using examples of students' work have a satisfactory match with the qualities exhibited in students' work.

There is no uniformly best method for describing the profiles of content and demands of a qualification. The appropriateness of the method is related to the purpose of the description. For example, different methods may be required for describing the action of students as opposed to the qualities exhibited by students. In addition, research has also demonstrated that using a combination of methods can lead to a more useful description of the profile.

Some of the methods (e.g., Delphi technique and fitness for purpose methodology) have the advantage that they can be simultaneously used to compare qualifications. This has obvious cost advantages. It should be recognised that many of the methods identified in this report can be used for more than one purpose. Because of the resource implications involved in applying the methods of articulation, examples of their use in practice are limited. Therefore, a degree of caution is necessary when assessing their effectiveness. A procedure that has proved useful and effective in one context might prove to be a disappointment in another. Similarly, a methodology that has been reported as being not successful in one context might be useful in another (particularly if it has been modified). It is also interesting to note that many of the researchers who have worked in this area adopted an approach that used more than one method. This would suggest that a combination of methods is appropriate so that the limitations of one method can be compensated for by the advantages of another method.

The literature review in chapter 3 suggests that the power of Kelly Repertory Grid (KRG) is that it aims to make what is implicit explicit. For this reason it has been used by a number of different authors to describe the qualities of different groups or people. It may be worth undertaking further research to refine the use of KRG as a method of articulation. For instance, KRG has been used with a mastery levels analysis to develop grade descriptors, but this approach has not as yet been used for essay questions. Different types of questions may effect the utility of the method. Also the amount of information that needs to be compared in KRG may effect the utility of the method. So it may have been harder for examiners in Gray's (1999) study to compare syllabuses than for examiners in Greatorex's (1999) study to compare candidates' performance. KRG can be used to elicit qualities not behaviours. This is where functional analysis can be used. This method appears to be effective and is recommended by reputable organisations like NCVQ and DfEE. The use of this method is limited to situations were the aim is to describe the behaviours required to fulfil a specific function or role. This is why it may be more useful in the vocational, professional or specialist areas rather than for describing academic qualifications. Fitness for Purpose is a flexible method for describing and comparing qualifications at the same level and qualifications at different levels. This versatility may be important in investigating progression between new qualifications and in the development of qualification frameworks.

There is one important general conclusion from the research in this area. It is important that the process of articulation and description is securely grounded in examples of candidate's work either drawn from assessments or actual practice. This increases the validity of the descriptions. A related issue is that a great deal of time and effort can be spent writing descriptions. But the understanding and use of the descriptors is possibly more important. This suggests that assessment that involves the use of tools like National Curriculum level descriptors would benefit from research about how assessors use the levels and how they understand them. Similarly, as mentioned above HE may benefit from research to relate level descriptors to students' work.

It is important to recognise that there were two stages to the process of articulation of profiles of content and/or outcomes. Firstly, it is necessary to obtain a description of the profile, and secondly, the profile that has been developed should be validated. Most of the research about profiles described in this report was concerned with the development of learning outcomes. There is a need for more research investigating the demand associated with qualifications.

In the fourth chapter, the research undertaken in comparing qualifications has been reviewed. For type I comparability, investigating two qualifications in same subject at the same level from the same system, there has been a great deal of research. These can be grouped into five generic approaches. The first three are largely statistical and are based on the use of prior outcomes, concurrent outcomes and subsequent outcomes. There have been two major influences on these generic approaches. Firstly, there has been an increase in the use of linked databases, which give greater access to each of three types of outcome.

Secondly, there has been an increase in the sophistication of statistical techniques (together with greater availability of suitable software) for the application of these generic methods. The availability of prior outcomes resulting from the linked databases means that comparability studies using these methods do not require very high levels of resources. An important advantage means that is possible to repeat the analysis over a number of years so that consistent patterns can be identified.

There is a disadvantage with these methods. The results of the study are not necessarily the result of differential performance of the students. This can only be investigated by considering the actual work of the candidates in relation to the objectives of the assessment. This means that these relatively cheap statistical methods should be considered for use as a screening process that could be used to identify potential problems for further study or detailed consideration of other evidence. A possible exception relates to the use of subsequent outcomes. For vocational qualifications, subsequent performance is often the main criteria for assessing the qualification.

The fourth generic method identified in this chapter involves comparing performance of candidates who have attempted both qualifications at the same time. Given that there are problems relating the preparation of candidates, their motivation, and implementing of the studies described in the chapter, this method cannot be recommended for general use. Although it has the potential to be very effective if the circumstances are right.

Finally, there is the judgement of experts. Research has shown that studies using the Thurstone paired comparison methodology can provide highly credible results. Studies of this type have the major disadvantage of being boring for participants, time consuming and expensive if a reasonable amount of evidence and a reasonable number of judges are used. There is also a need for further research into design of studies to ensure the effective use of resources.

Also in the fourth chapter, the difficult issue of between subject comparability was considered. Although there are severe conceptual problems associated with this issue. In practice, it is an issue that cannot be ignored entirely. The effects of choosing and channelling mean the perceived difference in between subject difficulty may have serious consequences for the uptake of particular qualification. There is a very strong case when subjects are similar.

Many aspects of research into comparability described in the three chapters require judgements to be made by experts. This inevitably leads to the problem of deciding who should be classed as an expert. It is conceivable that using different definitions in selecting experts might lead to different results in a comparability study. Throughout this report, there have been references to stakeholders. Serious consideration should be given to the nature and role of stakeholders for particular qualifications. The list of potential stakeholders can include the instructors, students, assessors, employers and representatives of society in general. Obviously not all stakeholders have a role in the application of some methods but they do have a role in the overall process. It could be reasonably argued that some of problems with qualifications, e.g. with parity of esteem, could be lessened by ensuring the involvement of appropriate stakeholders at the various stages of the comparability process. The reliance on judgement means that it is unlikely that there will be a complete consensus about any aspect of comparability.

There is one final consideration. This report has been about comparing qualifications. In real world situations involving decisions about recruitment, people are compared. These people will have different combinations of qualifications and different experiences (for example, there is a great deal of variety at GCSE level, e.g. Bell (1998). There is a danger that if research concentrates too narrowly on a few qualifications inappropriate inferences will be made. When comparing qualifications, it is necessary to consider the different learning pathways followed by candidates taking different qualifications. So for instance it is as important to know that GNVQ is comparable with A level as it is to know that OCR and AEB French GCSEs are comparable. With the development of linked databases, it is possible to investigate the different qualifications acquired with the qualifications under active consideration. This is an area that requires further research.

Appendix A: List of comparability studies involving GCE, GCSE and GNVQ qualifications

This is a list of 27 reports describing comparability studies carried out by the examination boards/awarding bodies and regulatory between 1984 and 1997. The reports have been listed in chronological order of publication. Note that some of the organisations changed name or merged with other organisations over the period.

1. Kingdon, J M, Wilmut, J, Davidson, K, Atkins, S B (1984) *Report of the Inter-Board Comparability Study of Grading Standards in Advanced Level English*, University of London School Examinations Board on behalf of the GCE Examining Boards.
2. Adams, R, Walker, N and Phillips, E (1989) *Inter-Group Comparability Study: 1988 Mathematics*, WJEC/NISEC/IGRC.
3. Alton, A (1989) *GCSE Inter-Group Comparability Study: 1988 Physics*, SEG/IGRC.
4. Fowles, D and Forrest, G (1989) *GCSE Inter-Group Comparability Study: 1988 French*, NEA/IGRC.
5. Stobart, G (1989) *GCSE Inter-Group Comparability Study: 1988 History*, LEAG/IGRC.
6. Adams, R, Phillips, E and Walker, A (1990) *GCSE Inter-Group Comparability Study 1989: Music*, WJEC/NISEC/IGRC.
7. AEB (1990) *Standards in Advanced Level Mathematics: Report of Study 4. A statistical study of double-subject mathematics candidates at the June 1986 examinations*, AEB/SRAC.
8. Cron, N and Houston, J (1990) *GCSE Inter-Group Comparability Study: 1989 Chemistry*, SEG/IGRC.
9. Fearnley, A (1990) *GCSE Inter-Group Comparability Study 1989: English Literature*, NEA/IGRC.
10. NISEAC (1990) *Standards in Advanced Level Mathematics: Study 5 Reports of: A Cross-Moderation exercise based on scripts from the June 1987 examinations in Advanced Level Mathematics and Study 6: A statistical survey based on the June 1987 examinations in Advanced Level Mathematics*, NISEAC/SRAC.

11. Patrick, H and McLone, R (1990) *GCSE Inter-Group Comparability Study 1989: CDT Technology*, MEG/IGRC.

12. Stobart, G, Elwood, J, Fordham, R and Mwanza, J (1990) *GCSE Inter-Group Comparability Study 1989: Geography*, LEAG/IGRC.

13. UCLES (1990) *Standards in Advanced Level Mathematics: Report of Study 1. A study of the demands made by the two approaches to 'Double Mathematics'*, UCLES/SRAC.

14. Thomson, D G (1992) *Grading Modular Curricula: Final Report of the GCSE Modular Aggregation Research and Comparability Study*, UCLES/MEG.

15. Taylor, M (1993) *GCSE Inter-Group Cross-Moderation Study 1992. Science: A study based on the Summer 1992 examinations*, SEG/IGRC.

16. NEAB (1994) *A Comparability Study in GCSE Geography: A study based on the Summer 1993 examinations*, NEAB/IGRC.

17. ULEAC (1994) *A Comparability Study in GCSE History: A study based on the Summer 1993 examinations*, ULEAC/IGRC.

18. Phillips, E and Adams, R (1995) *A Comparability Study in GCSE Mathematics: A study based on the Summer 1994 examinations*, WJEC/IGRC.

19. Alton, A (1995) *A Comparability Study in GCSE Science: A study based on the Summer 1994 examinations*, SEG/IGRC.

20. Francis, L, Stobart, G and Greig, A (1995) *Report of a Comparability Exercise into GCE and GNVQ Business Studie*s, SCAA.

21. Gray, E (1995) *A Comparability Study in GCSE English: A study based on the Summer 1994 examinations*, MEG/IGRC.

22. NEAB (1995) *A Comparability Study in Advanced Level Physics: A study based on the Summer 1994 and 1990 examinations*, NEAB/SRAC.

23. Quinlan, M (1995) *A Comparability Study in Advanced Level Mathematics: A study based on the Summer 1994 and 1989 examinations*, ULEAC/SRAC.

24. 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*, SCR/Joint Forum.

25. Jones, B (ed.) (1997) *A review and evaluation of the methods used in the 1996 GCSE and GCE comparability studies*, SCR/Joint Forum.

26. Jones, B, Baird, J and Arlett, S (1997) *A Comparability Study in GCSE Art and Design (Unendorsed): A study based on the Summer 1996 examinations*, NEAB/Joint Forum for the GCSE and GCE.

27. SCAA (c.1997) *Report of Comparability Exercise: GCE/GNVQ Science Advanced Level*, SCAA.

Appendix B: Reasons for using logistic and ordinal regression

In conventional regression analysis, it is assumed that the dependent variable is linearly related to the independent variable. For modelling examination grades, this is unsatisfactory. This can be illustrated by figure 1 in which a binary (pass-fail) dependent variable is used. Since the variable only takes the values 0 and 1, it is difficult to visualise the relationship between the two variables (particularly, since most of points on the graph represent overlapping points). To model this kind data, logistic regression is used. In logistic regression, the relationship between the probability of passing and the independent variable X. Since a probability cannot be greater than 1 or less than zero, a linear relationship is not usually appropriate. For a positive relationship, as in the figure, for very high values of X the probability is 1. This is a ceiling effect. A logistic curve has been proposed for this type of data because the S-shape includes floor and ceiling effects.

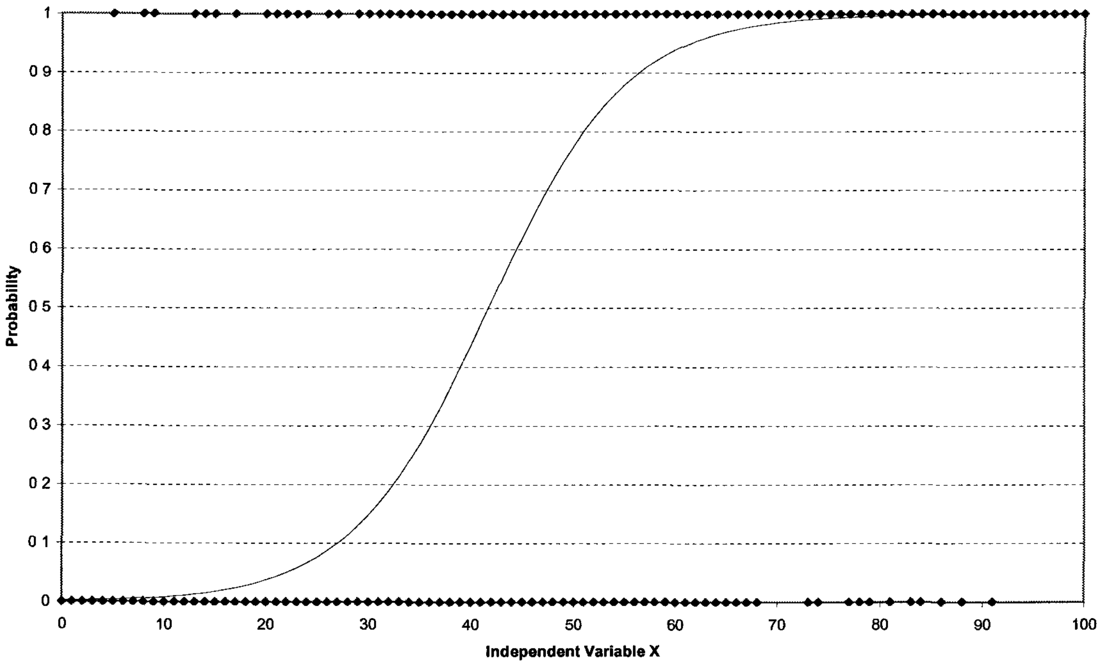


Figure 1: Example of the relationship of a binary independent variable with a continuous independent variable X

The usefulness of this curve can be demonstrated by using a decile plot. This was done by taking by dividing the range independent variable into ten equally-sized segments and counting the total number of individuals in each segment and the total number of individuals in segment who passed. Dividing the latter two quanitities gives an estimate of probability of passing in each decile. This probability can be converted into a logit ($\log(p/1-p)$

where p is the estimated probability). Using the logit transformation will results in a linear relationship between the logits and the independent variable if the logistic assumption holds.

In Table 1, the results of the calculations for the hypothetical data used in Figure 1 have been presented. The probabilities and the logits can be plotted against the midpoints of the deciles. In theory, the probabilities should follow the s-shaped logistic curve and the logits a straight line.

Table 1: Estimated probabilities for each decile

Midpoint of the decile	Mean y-value (Est. prob.)	Logit
4.5	0.08	-1.06
14.5	0.10	-0.95
24.5	0.26	-0.45
34.5	0.44	-0.10
44.5	0.44	-0.10
54.5	0.5	0.00
64.5	0.64	0.25
74.5	0.88	0.87
84.5	0.88	0.87
94.5	0.98	1.73

In Figure 2, the logits have been plotted against the midpoints. There is no strong evidence that a linear fit is inappropriate.

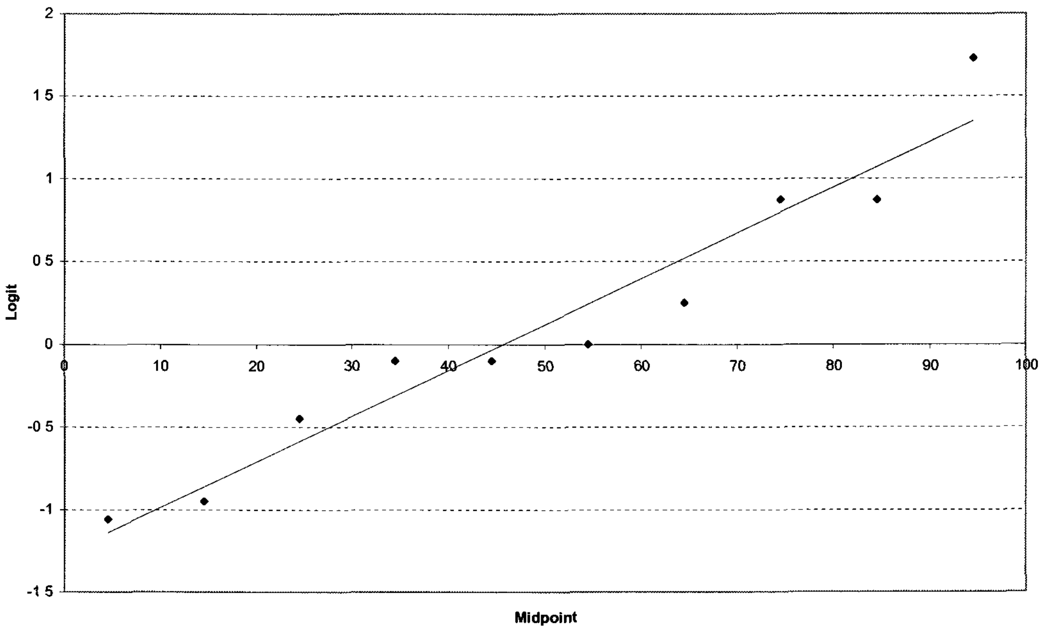
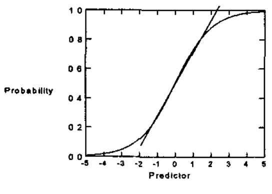


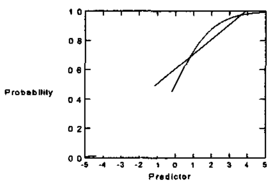
Figure 2: Logits against midpoints for the hypothetical data

The above method is only for illustrative purposes. In practice, a technique call called maximum likelihood estimation is used to obtain estimates of the slope and intercept for such data.

The reason why floor and ceiling effects cannot be ignored in comparability studies can be illustrated by considering the following hypothetical example illustrated in Figure 3. For qualification A, the relationship between the probability of obtaining a particular grade and a predictor is represented by the s-shaped logistic curve. If most of the candidates had values for the predictor of plus or minus 2 then a linear regression fit would be reasonable. For qualification B, the relationship between the probability and the predictor is the same but the entry consisted of candidates with high values for the predictor. It is clear from the sketch diagram that a linear fit would not be a good approximation. In addition, using linear fits would suggest a spurious difference in the relationship between the probability and the predictor, i.e. that the two qualifications are not comparable when they really are.



Qualification A



Qualification B

Figure 3: An example of the problem of using linear fits when floor and ceiling effects exist

When there are a number of grades awarded rather than a pass-fail, then dependent variable is ordinal. Various models have been proposed for regression modelling with ordinal dependent variables. One of the commonest is the proportional odds model. For this type of model, the cumulative probabilities are used, e.g. the probability of obtaining at least a grade C. In a proportional odds model, the shape of the relationship between the cumulative probability and the independent variable is assumed to be the same for each probability as in Figure 4 (for the cumulative logits, the relationship is assumed to be a series of parallel lines). In Figure 4, for an individual with a score of 35 on the independent variable, the probability of getting a grade A is effectively 0 and the probability of getting at least a grade F is approximately 0.75.

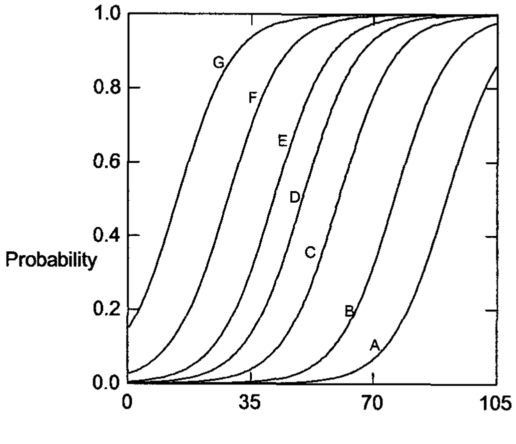


Figure 4: Probability curves resulting from fitting a proportional odds model

The problem with fitting a proportional odds model is that the assumption that the curves are same shape. This has to be verified and the easiest way of doing this is to carry out a series of separate logistic regressions at each grade boundary (e.g. modelling the probability of getting at least a grade D with the probability of getting a grade E or less). If the slope parameters were the same for each regression, then a proportional odd model could be fitted. Although a proportional odds model may be more efficient because fewer parameters are estimated, this advantage is minimal for the large data sets available for statistical comparability studies. It is for this reason that a series of separate logistic regressions can be carried out instead of a proportional odds model.

Glossary

Ability test: A test that measures the current performance or estimates performance in some cognitive, psychomotor, or physical functioning domain.

Accreditation of prior learning (also known as Accreditation of Prior Achievement): certificating competence on the basis of evidence from past achievements, often supplemented by current assessments. Sometimes used in a wider sense to include counselling, helping people to recognise the significance of their experience as a prelude to assessment and accreditation, and providing guidance and action planning following such accreditation also referred to as 'accreditation of prior achievement' (Burke, 1995).

Achievement test: A test that measures the extent of a certain body of knowledge or skill usually after instruction has been received.

APU: Assessment of Performance Unit. The Assessment of Performance Unit (APU) was set up in 1975 within the Department of Education and Science (DES) to promote the development of methods of assessing and monitoring achievement of children at school and to seek to identify the incidence of underachievement.

AQF: Australian Qualifications framework

Attainment targets: 'the broad objectives specified a National Curriculum subject, setting out the knowledge, skills and understanding pupils are expected to acquire' (Burke, xi).

AWA: Advanced Extension Award.

Benchmark: a point of reference.

CATS: Credit Accumulation and Transfer Scheme

CNAA: 'Council for National Academic Awards; before the polytechnics became universities, the CNAA had an important validating role for degrees certificates and diplomas; now defunct '(Burke, 1995, viii).

Comparable scores: Scores on different test expressed on the same scale and having the same relative meaning with some common reference group.

Competence: The ability to perform to recognised standards (Burke, 1995). The vocational and training sector defines *competency* as the possession and application of both knowledge and skills to defined standards, expressed as outcomes, that correspond to relevant workplace requirements and other vocational needs (AQFB, 1998).

Content domain: A clearly defined body of knowledge and skills that identifies what is being assessed.

Core Skills: 'skills (or facet of skill) which underpin, and are common to, a wide range of competent performance. The acquisition of such skills is believed to facilitate transfer to performance in a wide range of functions and situations' (Burke, 1995, xi).

Credit: Credit can be conceptualised as an accounting or points system where achievements i.e. qualifications and parts of qualifications each have a credit rating.

Credit Accumulation: 'the general process by which separate components of a qualification system can be separately achieved and certificated, allowing the accumulation of such achievements over time' (Burke, 1995, xi).

Credit Transfer: 'the recognition of a credit gained in one qualifications, or system of qualifications, as satisfying some or all of the requirements of a different qualification, or system of qualifications. It alleviates the need for repeating assessments (and possibly training) for the award, or that part of the award, for which recognition is given in the second qualification or system' (Burke, 1995, xi).

Criterion: An indicator of the accepted level of performance.

Criterion-referenced test: A test that allows its users to make score interpretations in relation to a functional

CSE: Certificate of Secondary Education.

Demands: The cognitive skills and processes and the application of these that are required by a task.

DfEE: Department for Education and Employment.

Domain: a specified area of study or knowledge. So in English domains might be oral communication, Reading and Writing.

Equated forms: Two or more test forms that yield equivalent parallel scores for specified groups of test takers.

Equating method: A process used to convert the score scale of one form of a test to the score scale of another from so that the scores are equivalent or parallel.

FEU: Further Education Unit.

GCE A levels: General Certificate of Education Advanced Level.

GCSE: General Certificate of Secondary Education.

General National Vocational Qualification (GNVQ): 'a broad-based vocational qualification, assessed to national standards, which attests attainment, i.e. general skills (including Core Skills) knowledge and understanding which underpin a range of occupations, providing certification of achievement which may act as a springboard to enter employment or pursue further or higher education and training' (Burke, 1995, xii).

Grade descriptors: a description of the qualities likely to be found in the performance of a candidate who has achieved a particular grade.

Grade related criteria: a description of what a candidate needs to achieve to be awarded a particular grade.

Graduateness: the attributes shared by all graduates (Greatorex, 1998).

HEQC: Higher Education Quality Council.

InCCA: Inter Consortium, Credit Agreement

Key Skills: previously Core Skills (see above).

Knowledge: 'the 'know-how' or cognitive component that underpins competence or attainment, which may include facts, theories, principles, conceptual frameworks etc. It subsumes 'understanding' (Burke, 1995, xii).

Learning outcomes: a statement of the learning that is to be achieved (Greatorex, 1998).

Level descriptors (Generic level descriptors): descriptions of the attributes generally associated with learners at particular levels of their HE career, for example, a Masters' student may be considered to be autonomous.

Level:

- 'A measure of intellectual demand or difficulty....;
- A measure of progression through a curriculum or syllabus.....;
- A discriminator in the grading of academic performance.' (HEQC, 1997, 11).

National Curriculum: The UK government uses this national system of classifying attainment targets and statements of attainment by subjects and levels. The required education provision for school children aged 5 to 16 years in state schools' (Burke, 1995).

NCVQ: National Council for Vocational Qualifications

National Vocational Qualification (NVQ) or Scottish Vocational Qualification (SVQ): 'a qualification that is accredited by NCVQ and allocated a place within the NVQ framework. NVQs are required to meet specified criteria for accreditation' (Burke, 1995).

NICATS: Northern Ireland Credit Accumulation and Transfer Scheme

Normative: pertaining to norms or norm groups.

Norms: Statistics or tabular data that summarise the test performance of specified groups and are assumed to represent a larger population.

NVQ framework: 'the national system of classifying NVQs according to area of competence and level' (Burke, 1995).

Occupational Standards: consist of:-

Elements of competence which state the functions which are needed in particular occupational areas;

Performance criteria, which are attached to each element, describe the quality of the outcomes of successful performance;

The indicators of range describe the potential dimensions or parameters of the function - what is included in the coverage of the element and performance criteria and what is not' (Mitchell, 1995, 75).

Predictive criterion-related evidence of validity: Evidence of criterion-related validity in which criterion scores are observed at later at a later date.

Progression: 'the development or accumulation of competence or attainment by an individual through successive learning opportunities (programmes/courses/qualifications/ experiences) in a systematic manner. Also the related advancement in an individual's career through successive jobs' (Burke, 1995).

QAA: Quality Assurance Agency for Higher Education who aim to establish comparability between degree standards and to hold Higher Education accountable to stakeholders.

Qualification Frameworks: Map the relationship between different qualifications.

Qualifications: formal certification, issued by a relevant approved body, in recognition that a person has achieved learning outcomes or competencies relevant to identified individual, professional, industry or community needs.

Recognition of prior learning: the process by which credits are granted towards qualifications through assessment of an individual's knowledge and skills gained through education, training, work and life experience.

SEEC: South East England Consortium

Skills: 'the 'performance' component that underpins competence. Distinguished from competence by being more fundamental and frequently common to a variety of different competences.....Skills may be manual or cognitive, or combination of both' (Burke, 1995).

TDRU: The Test Development and Research Unit was established in 1968 by the three GCSE boards associated with the Universities of Oxford and Cambridge (OCSEB, UODLE and UCLES).

Threshold: a point of reference that is a minimum requirement.

THES: Times Higher Educational Supplement.

Transfer: previously acquired knowledge and skills can be 'transferred' i.e. applied to a new context, this means that in the new context the learner becomes competent more quickly or that they can perform the new competence without prior experience.

Vocational Assessment: 'an assessment may be defined as vocational to the extent that its validity has an external referent that derives from the world of employment' (Wolf 1996)

Bibliography

AQF (1998) Implementation Handbook. 2nd Edn. Carlton, Vic.: Australian Qualifications framework (AQF) Advisory Board.

Atkins, M. J., Beattie, J., & Dockrell, A. (1993). Assessment and student learning. In M. J. Atkins, J. Beattie, & A. Dockrell. Assessment issues in Higher Education. School of Education, University of Newcastle. Newcastle upon Tyne: Employment Department, Further and Higher Education.

Ball, C. (1996) Life-long learning for the 21st century, keynote address at the 21st Improving University Conference. Nottingham: Nottingham Trent University.

Bardecki, M. J. (1984). Participants' Response to the Delphi Method: An Attitudinal Perspective. Technological Forecasting and Social Change, (25), 281-292.

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

Barnett, R. (1994). The Limits of Competence, Knowledge, Higher Education and Society. Milton Keynes: The Society for Research into Higher Education and the Open University Press.

Bell, J. (1996). Graduateness: some early thoughts. A presentation at a seminar on 'Standards' as part of the Giving Credit Network at the University of Leeds, 29 Oct, Department of Law, University of Leeds.

Bell, J.F. (1998) Patterns of subject uptake and examination entry 1984-1997. Paper presented at the British Educational Research Association Annual Conference, Queen's University, Belfast, 26-30 August. (Available from Education-line at http://www.leeds.ac.uk/educol/BEID.html).

Bell, J. F. (2000) Review of the use of Thurstone pair methodology to monitor examination standards. A report for the QCA Project: Review of models for maintaining and/or monitoring examination/test standards over time (unpublished).

Bell, J. F., Bramley, T., & Raikes, N. (1997a). Standards in A-level Mathematics 1986-1996. Paper presented at the 1997 BERA conference at the University of York.

Bell, J. F., Bramley. T., & Raikes, N. (1997b). Investigating A level mathematics standards over time.

British Journal of Curriculum and Assessment, 8(2), 7-11.

Benner, P. (1984). From Novice to Expert. Menlow Park, California: Addison Wesley.

Bligh, D., Caves, R., & Settle, G. (1980). 'A' level scores and degree classifications as functions of university type and subject. D. Billing (Editor), Indicators of performance . Guildford: Society for research in Higher Education.

Bloom, B. (1956). Taxonomy of Educational Objectives, Book 1: Cognitive Domain. London: Longman.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). Handbook on Formative and Summative Evaluation of Student Learning. McGraw-Hill, Inc.

Bloomfield, B., Dobby, J. and Duckworth, D. (1977). Mode comparability in the CSE: a study of two subjects in two examining boards. Schools Council Examinations Bulletin, 36. Evans/Methuen Educational.

Boam, R. & Sparrow, P. (1992). Designing and achieving competency. London: McGraw-Hill.

Bobbit, F. (1918). The Curriculum. Boston M: Houghton Mifflin.

Bourner, T., & Hamed, M. (1987). Entry qualifications and degree performance. London: Council for National Academic Awards.

Boyatzis, R. (1982). The Competent Manager: A Model for Effective Managers, New York: Wiley.

Bramley, T., Bell, J. F., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone paired comparisons. Education Research and Perspectives, 25(2), 1-23.

Broadfoot, P. (1996). Education, Assessment, Society. Buckingham: Open University Press.

Brown, S. (1980). What do they know? A Review of Criterion-Referenced Assessment. Occasional Papers. Scottish Educational Department. Edinburgh: HMSO.

Burke, J. (1995). Outcomes, Learning and the Curriculum Implications for NVQ's, GNVQ's and other qualifications. London: The Falmer Press.

Burroughs, G. E. R. (1971). Design and Analysis in Educational Research, Educational Monograph No. 8, Oxford: Alden Mowbray Ltd.

Burstein, L. (1992) The IEA Study of Mathematics III: Students growth and classroom process. Oxford: Pergamon press.

Catterall, M., & Maclaran, P. (1997). Focus Group Data and Qualitative Analysis Programs: Coding the

Moving Picture as Well as the Snapshots. Sociological Research Online, 2 (1).

Charters, W. W. (1924). Curriculum Construction, New York: Macmillan.

Christie, T., & Forrest, G. M. (1981). Defining Public Examination Standards. London: Schools Council.

Clarke, M. J., & Youngman, M. B. (1987). Dispositional associates of GCE performance in sixth form and further education college students. British Journal of Educational Psychology, 57, 191-204.

Cohen, L., & Manion, L. (1994). Research Methods in Education. London, New York: Routledge.

Coles, M., & Matthews, A. (1995). Fitness for Purpose: A means of comparing qualifications. A report to Sir Ron Dearing to be considered as part of his review of 16-19 education.

Council for National Academic Awards (CNAA). (1989). CNAA Handbook. London: CNAA.

Cresswell, M. (1996). Defining, setting and maintaining standards in curriculum-embedded examinations: Judgemental and statistical approaches. In H. Goldstein, & T. Lewis (Eds) Assessment problems, developments and statistical issues. Chichester: John Wiley.

Cresswell, M. J. (1987). Describing Examination Performance: grade criteria in public examinations. Educational Studies, 13(3), 247-263.

Cresswell, M. J., & Houston, J. G. (1991). Assessment of the National Curriculum - some fundamental considerations. Educational Review, 43(1), 63-78.

Davis, B. D., & Burnard, P. (1992). Academic Levels in Nursing. Journal of Advanced Nursing, 17, 1395-1400.

Dearing, R. (1996). Review of Qualifications for 16-19 year olds. London: SCAA.

Dearing, R. (1996). Review of Qualifications for 16-19 Year Olds. Mathematics and the Sciences. London: SCAA.

Debling, G. (1995). Panacea and Parity: Addressing the Myths about NVQs and SVQs. Competence & Assessment. Special Issue: Higher Level Vocational Qualifications (Chap. 29, pp. 6-9). Sheffield: Employment Department.

Debold, R. (1996). The nominal group technique. http://www.radix.net/~ash2jam/TQM/nominal.htm

Delbecq, A., Van de Ven, A. and Gustafson, D. (1975). Group techniques for program planning: A Guide to Nominal Group and Delphi Processes. Middleton, WI: green Briar Press.

Dewey, J. (1938). Experience and Education. New York: Collier MacMillan Publishers.

Dexter, T., and Massey, A. (2000) Conceptual issues arising from a comparability study relating IGCSE grading standards with those of GCSE via a reference test using multilevel models. Paper to be presented at the 22nd Biennial Conference of the Society for Multivariate Analysis in Behavioural Science. LSE. London.

DFE (1995). Key Stages 1 and 2 of the National Curriculum. Department for Education, London: HMSO.

DfEE (1999). Key Skills Explained, Nottingham: DfEE.

DfEE. (1999). Key Skills Explained. Nottingham: DfEE.

Employment Department. (1995). Degrees of Influence. Sheffield: Employment Department.

Eraut, M. (1989). Specifying and Using Objectives. M. Eraut (ed), The International Encyclopedia of Educational Technology. Oxford: Pergamon Press.

Erffmeyer, R. C., Erffmeyer, E. S., & Lane, I. M. (1986). The Delphi Technique: An Empirical Evaluation of the Optimum Number of Rounds. Group and Organizational Studies, (11), 120-128.

FEU. (1992). A Basis for Credit? London: FEU.

Fielding, A. (1999). Why use arbitrary points scores? Ordered categories in models of educational progress. Journal of the Royal Statistical Society, Series A - Satistics in Society, 162(3), 303-328.

Fitz-Gibbon, C. T. (1992a). The Design of Indicator Systems. Research Papers in Education, 7(3), 271-300.

Fitz-Gibbon, C. T. (1992b). Multilevel modelling in an indicator system. S. W. a. W. J. D. Raudenbush Pupils and classrooms: International Studies of Schooling from a Multilevel Perspective . London and New York: Acadamic Press.

Fitz-Gibbon, C. T., & Vincent, L. (1994). Candidates' Performance in Public Examinations in mathematics and Science. London: SCAA.

Fitz-Gibbon, C.T., & Vincent, L. (1997). Difficulties regarding subject difficulties: developing reasonable explanations for observable data. Oxford Review of Education, 23(3), 291-298.

Flynn, J. R. (1986). Massive IQ gains in 14 nations: what IQ tests really measure. Psychological Bulletin, 101(2), 171-191.

Forrest, G. M., & Vickerman, C. (1982). Standards in GCSE: subject pairs comparisons, 1972-1980. Manchester: Joint Matriculation Board.

Forrest, G.M., & Shoesmith, D.J. (1985). A second review of GCE comparability studies. Manchester:

Joint Matriculation Board.

Forster, M., & Gray, E. (2000). Impact of Independent judges in comparability studies conducted by Awarding Bodies. Paper to be present at BERA 2000. Cardiff: University of Wales.

Fraser, S. (1964). Jullien's plan for comparative education. 1816-17. New York: Columbia Unversity, Teacher College.

Freudenthal, H. (1975) Pupils' achievement internationally compared – the IEA. Educational Studies in Mathematics, 6, 127-186.

Fulcher, G. N. (1993). Unpublished PhD dissertation, Lancaster University.

Gagné, R. M. (1965). The analysis of instructional objectives for the design of instructions. In R. Glaser (ed), Teaching Machines and Programmed Learning, Volume 2 Data and Directions. Washington DC: Department of Audio-visual Instruction, National Education Association (NEA).

Gibbs, A. (1998). Focus Groups. Social Research Update, (19).

Gipps, C. (1990). Assessment: A Teachers' Guide to the Issues. London: Hodder & Stoughton.

Glaser, R. and Klaus, D. J. (1962). Proficiency measurement: assessing human performance. In R. Gagné, (ed.). Psychological Principles in System Development, 419-72 New York: Holt, Rinehart and Winston Inc.

Goldstein, H., & Cresswell, M. (1996). The comparability of different subjects in public examinations: a theoretical and practical critique. Oxford Review of Education, 22, 435-442.

Goldstein, H., & Thomas, S. (1995). Using examination results as indicators of school and college performance. Journal of the Royal Statiistical Society, Series A, 159, 149-163.

Gray, E. (1999). A Comparability Study in GCSE Science 1998: A study based on the Summer 1998 examination . OCR (MEG) on behalf of the Joint Forum for the GCSE and GCE.

Greatorex, J. (1998). Educational Levels in a Higher Education and Healthcare Context. Unpublished doctoral dissertation, University of Derby.

Greatorex, J. (1999a). Generic Descriptors: a health check. Quality in Higher Education, 5(2), 155-166.

Greatorex, J. (1999b). Making the Grade - A Novel Method for Developing Grade Descriptors. British Educational Research Association Annual Conference 2-5 September, University of Sussex at Brighton.

Greatorex, J., & Dexter, T. (1998). An accessible analytical approach for investigating what happens between Delphi rounds. British Psychological Society London Conference .

Greatorex, J., & Nyatanga, L. (1994). Academic Levels in the Accreditation of Prior Learning. Perspectives on Experiential Learning: The 1994 International Experiential Learning Conference.

Hadfield, G. (1980). Sources of variation in what constitutes 'good' art examinations at sixteen plus. Unpublished doctoral dissertation, University of Manchester.

Haque, Z., and Bell, J.F. (2000). Evaluating the Performances of Minority Ethnic Pupils in Secondary Schools. Paper to be present at BERA 2000. Cardiff: University of Wales.

Hartog, P., & Rhodes, E. C. (1936). The Marks of Examiners. London: Macmillan.

Harvard, N. (1996). Student attitudes to studying A-level sciences. Public Understanding of Science, 5(4), 321-330.

HEQC. (1997). Graduate Standards Programme. London: HEQC.

Hirst, P. H., & Peters, R. S. (1970). The Logic of Education. London: Routledge and Kegan Paul.

Hodgson, A., & Spours, K. (2000). Qualifying for Success: Towards a Framework of Understanding. Institute of Education, University of London.

Holsti, O. (1969). Content Analysis for the Social Sciences and Humanities. Reading, MA: Addison Wesley.

Hoskins, S., Newstead, S. E., & Dennis, I. (1997). Degree performance as a function of age, sex, prior qualifications and discipline studied. Assessment and Evaluation in Higher Education, 22, 317-328.

Hughes, S., Pollitt, A., & Ahmed, A. (1998). The development of a tool for gauging the demands of GCSE and A level exam questions. British Educational Research Association Annual Conference.

Husén, T. (1967) International Study of Achievement in Mathematics, Vols. I and II. Stokholm: Almquist and Wiksell.

InCCA (1998). A Common Framework for Learning. DfEE.

Jaeger, R. M. (1978). A proposal for setting a standard on the North Carolina High School Competency Test. Annual Meeting of the North Carolina Association for Research in Education North Carolina: Chapel-Hill.

James, C., & Redfern, L. (1995). The description of levels in nursing degrees: an illustration and analysis of the variations. Journal of Clinical Nursing, 4, 311-317.

Jessup, G. (1991). Outcomes: NVQs and the Emerging Model of Education and Training. London: The Falmer Press.

Jessup, G. (1995). Outcome Based Qualifications and the Implications for Learning. J. Burke (ed), Oucomes, Learning and the Curriculum: Implications for NVQs, GNVQs and other qualifications . London: The Falmer Press.

Johnson, S. (1989). National Assessment: The APU Science Approach. London: HMSO.

Johnson, S., & Bell, J. F. (1985). Evaluating and predicting survey efficiency using generalizability theory. Journal of Educational Measurement, 22, 107-119.

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

Jones, B. E. (1997). Comparing Examination Standards: is a purely statistical approach adequate? Assessment in Education, 4(2), 249-262.

Jones, J., & Hunter, D. (1995). Consensus methods for medical and health services research. British Medical Journal, (311), 376-380.

Kahney, H. (1993). Problem Solving: current issues. Milton Keynes: Open University Press.

Kandola, R., & Pearn, M. (1992). Identifying Competencies. In R. Boam, & P. Sparrow Designing and Achieving Competency: A Competency based approach to developing people and organizations. London: McGraw-Hill Book Company.

Kastein, M. R., Jacobs, M., Van der Hell, R. H., Luttik, K., & Touw-Otten, F. W. M. M. (1993). Delphi, the Issue of Reliability: A qualitative Delphi study in Primary Health Care in the Netherlands. Technological Forecasting and Social Change, (44), 315-323.

Kastein, M. R., Jacobs, M., Van der Hell, R. H., Luttik, K., & Touw-Otten, F. W. M. M. (1993). Delphi, the Issue of Reliability: A qualitative Delphi study in Primary Health Care in the Netherlands. Technological Forecasting and Social Change, (44), 315-323.

Kelly, A. (1976). A study of the comparability of external examinations. Research in Education, 16, 38-63.

Kelly, G. A. (1955). The Psychology of Personal Constructs, vols I and II, Norton, New York.

Kim, K.-S. (1990) Multilevel Data Analysis: a Comparison of Analytical Alternatives. Ph.D. thesis, University of California, Los Angeles.

Kingdon, M., & Stobart, G. (1988). GCSE Examined. Lewes: The Falmer Press.

Kolb, D. (1985). Experiential Learning: Experience as the Source of Learning and Development. Prentice Hall: New Jersey.

Korb, L. J. (1982). Profile of American youth. Washington, DC: Office of the Assistant Secretary of Defense.

Kreft, I. G. G. (1996) Are multilevel techniques necessary? An overview, including simulation studies. Los Angeles: California State University.

Kreitzer, A. E., & Madaus, G. F. (1994). Empirical investigations of the hierarchical structure of taxonomy, a forty-year perspective, Ninety-third yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press.

Leon, P. (2000). Trying to box clever. Times Higher Educational Supplement. No. 1,429, March 31, 39.

Linstone, H., & Turoff, M. (1975). The Delphi Method: Techniques and Applications. London: Addison Wesley.

Linstone, H., & Turoff, M. (1975). The Delphi Method: Techniques and Applications. London: Addison Wesley.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hilsdale, NJ: Lawrence Erlbaum.

Lord, K., Wake, G. D. and Williams, J. S. (1995). Mathematics for progression from Advanced GNVQs to Higher Education; The Mechanics in Action project, Manchester: University of Manchester.

Mager, C. (1991). Assessment in Open College Networks. Leicester: National Institute for Adult Continuing Education.

Mansfield, B. and Horton, P. (1986). Work Based Assessment and its Certification, Leeds and Yorkshire Humberside Association for Further and Higher Education.

Mansfield, B. and Mitchell, L. (1996). Towards a Competent Workforce. England: Gower Publishing Ltd

Martino, J. P. (1993). Technological Forecasting for Decision Making. Dayton: McGraw-Hill Inc.

Massey, A. J. (1982). Assessing 16+ Chemistry: The exposure-mastery gap. Education in Chemistry, (September), 143-145.

Max Planck Institute of Human Development. (1983). Between elite and mass education in the Federal Republic of Germany. Albany: State University of New York Press.

McKelvey, R.D., and Zavoina, W. (1975) A statistical model for analysis of ordinal level dependent variables. Journal of Mathematical Sociology, 4, 103-120.

Mitchell, L. (1993). NVQs/SVQs at Higher Levels: A Discussion Paper to the 'Higher Levels' Seminar, October 1992. Competence and Assessment Briefing Series, Number 8. Sheffield: Employment Department.

Mitchell, L. (1995). Outcomes and National (Scottish). Vocational Qualifications. In J. Burke (ed), Outcomes, Learning and the Curriculum: Implications for NVQs, GNVQs and other qualifications. London: The Falmer Press.

Moon, J. (1999). Describing Higher Education - Some Conflicts and Conclusions. In H. Smith, M. Armstrong and S. Brown, Benchmarking and Threshold Standards in High Education, Staff and Educational Development Series. London: Kogan Page.

Morrison, H. G., Busch, J. C. and D'Arcy, J. (1994). Setting Reliable National Curriculum Standards, Assessment in Education Principles, Policy and Practice, 1, 2, 181-199.

Morrison, H., Wylie, C., McFaul, P. & Thompson, W. (1996). The passing score in the Objective Structured Clinical Examination. Medical Education, (30), 345-348.

Mulgrave, N. W., & Ducanis, A. J. (1975). Propensity to Change Responses in a Delphi Round as a Function of Dogmatism. H. Linstone, & M. Turoff (eds), The Delphi Method: Techniques and Applications (pp. 288-290). London: Addison Wesley.

NCIHE (1997). Education in the Learning Society, Report of the National Committee (the 'Daring report'). London: HMSO.

New Zealand Vice-Chancellor's Committee (1994). The National Qualifications framework and the Universities. Wellington: New Zealand Vice Chancellor's Committee.

Newbould, C., & Massey, A. (1979). Comparability using a common element. Cambridge: UCLES.

Newbould, C.A. (1974). Technical drawing reference test 1973. Appendix A to the report on the 16+ feasibility study in technical drawing by AEB and UCLES. Cambridge: UCLES.

Newton, P. E. (1997). Measuring comparability of standards between subjects: why our statstical techniques do not make the grade. British Educational Research Journal, 23(4), 433-449.

NICATS. (1998). Report of the Northern Ireland Credit Accumulation & Transfer System (NICATS) Project (draft). NICATS.

NICATS. (1999). Report of the Northern Ireland Credit Accumulation & Transfer System (NICATS) Project. NICATS.

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

Nuttall, D.L. (1971). The 1968 CSE monitoring experiment. Schools Council Working Paper, 34. Evans/Methuen Educational.

OFSTED, & FEFC. (1999). GNVQs: Evaluation of the Pilot of the New Assessment Model, 1997-1999. Coventry: Further Education Funding Council.

O'Neil, M., & Jackson, L. (1983). Nominal group technique: A process for initiating curriculum development in higher education. Studies in Higher Education, 8(2), 129-138.

O'Reilly, D. (1990). Hierarchies in mathematics: a critique of the CSMS study. P. Dowling, & R. Noss (eds), Mathematics versus the National Curriculum . Basingstoke: The Falmer Press.

Otter, S. (1992). Learning Outcomes and Higher Education. Leicester: Unit for Development of Adult Continuing Education.

Otter, S. (1995). Learning Outcomes in Higher Education. J. Burke (ed), Outcomes, Learning and the Curriculum: Implications for NVQs, GNVQs and other qualifications . London, Washington: The Falmer Press.

Patel, V. L., & Groen, G. J. (1992). The Representation of Medical Information in Novices, Intermediates and Experts. J. Lun et al. (eds), MEDINFO . North Holland: Elsevier Science Publishers B. V.

Peddie, R. (1997) Difficulty excellence and level: implications for qualifications frameworks. Australian and New Zealand Journal of Vocational Educational Research, 5(2),56-76.

Peers, I. S., & Johnston, M. (1994). Influence of learning context on the relationship between A-level attainment and final degree performance: a meta-analytic review. British Journal of Educational Psychology, 64, 1-18.

Pollitt, A., & Murray, N. L. (1996). What raters really pay attention to. In M. Milanovic, & N. Saville (eds), Studies in Language Testing: 3 Performance Testing, Cognition and Assessment. Cambridge: Cambridge University Press.

Popham, W. K. and Husek, T. R. (1969). Implications of criterion-referenced measurement, Journal of Educational Measurement, 6, 1-9.

Postlethwaite, T. N. (1987). Comparative Education Review, Special Issue., 31(1).

Qualifications and Curriculum Authority. (1999). Curriculum Guidance for 2000 - implementing the change to 16-19 qualifications. London: QCA.

Quality Assurance Agency for Higher Education. (1999a). A consultative paper on Higher Education Qualifications Frameworks for England, Wales and Northern Ireland (EWNI), and for Scotland. Gloucester: QAA.

Quality Assurance Agency for Higher Education. (1999b) A Consultation Paper on Qualifications Frameworks: Postgraduate Qualifications [Web Page]. URL www.niss.ac.uk/eduation/qaa/pub98/pgqual/consult#Levels.

Quality Assurance Agency for Higher Education. (1999c). Consultation on Higher Education Qualifications Frameworks 1 Dec, University of Westminster.

Race, P. (1998 November). Give students the big picture. The Times Higher Education Supplement, p. 30.

Raggatt, P., & Williams, S. (1999). Government, Markets and Vocational Qualifications: An Anatomy of Policy. London: The Falmer Press.

Randall, J. (1998 October). Flexible friends. The Times Higher Education Supplement, p. i.

Redfern, L., & James, C. (1994). Credits, Levels and Professional Learning: Some considerations. Fifth Annual Nurse Education Tomorrow Conference .

Reeves, M. (1988). The Crisis in Higher Education: Competence, Delight and the Common Good. Milton Keynes: Society for Research into Higher Education and Open University Press.

Robbins Report. (1963). Report of the Committee on Higher Education. London: HMSO.

Robertson, D. (1994). Choosing to Change, extending access, choice and mobility in higher education: The report of the HEQC, CAT Development Project. London: Higher Education Quality Council.

Robertson, D. (1996). Credit transfer and the mobility of credits in UK higher education: the evolution of policies, meanings and purposes. Journal of Education Policy, 11(1), 53-73.

Robinson, W. P., & Tayler, C. A. (1992). Changes in pupils' self-perceptions and self-evaluations: from CSE/GCE to GCSE. Educational Psychology, 12(2), 107-112.

Roth, P. L. (1994). Group Approaches to the Schmidt-Hunter Global Estimation Procedure. Organizational Behaviour and Human Decision Making Processes, 59(3), 428-451.

Sackman, H. (1975). Delphi Assessment. G. Wright, & P. Ayton Judgmental Forecasting . Chichester, New York: John Wiley and Sons Ltd.

Scheibe, M., Skutsch, M., & Schofer, J. Experiments in Delphi Methodology. H. Linstone, & M. Turoff The Delphi Method: Techniques and Applications . London: Addison Wesley.

Schmidst, H. G., Norman, G. R., & Boshuizen, H. D. A. (1990). A Cognitive Perspective on Medical Expertise. Academic Medicine, 65(10), 611-621.

Schools Council (1966). The 1965 Monitoring Project. Evans/Methuen Educational.

Sear, K. (1983). The correlation between A-level grades and degree results in England and Wales. Higher Education, 12, 606-619.

SEEC (1996). Credit Guidelines, Models and Protocols. DfEE.

Shackleton, V. (1992). Using a competency approach in a business change setting. In R. Boam, & P. Sparrow Designing and Achieving Competency: A Competency based approach to developing people and organizations. London: McGraw-Hill Book Company.

Shoesmith, D.F., Newbould, C.A., & Harrison, A.W. (1977). A common element in GCE French Examinations. Oxford: Oxford and Cambridge Schools Examination Board.

Spencer, L. (1983). Soft Skill Competencies. Edinburgh: Scottish Council for Research in Education.

SRAC (1990). A Study of the Demands Made by the Two Approaches to 'Double Mathematics': An investigation conducted by the University of Cambridge Local Examinations Syndicate on behalf of the Standing Research Advisory Committee of the GCE Examining Boards. SRAC.

Steinaker, N. W., & Bell, M. R. (1979). The Experiential Taxonomy: A new approach to teaching and learning. Academic Press Inc.

Stenhouse, L. (1975). An Introduction to Curriculum Research and Development. London: Heinemann.

Stewart, V. (1998). Business Applications of Repertory Grid http://203.96.24.202/busiappof

Swann, W. (1992). Hardening the hierarchies: the National Curriculum as a system of classification. T. Booth, & W. Swann (eds), Curriculum for Diversity in Education . London: Routledge/Open University Press.

Tate, R., and Wongbundhit, Y. (1983) Random versus nonrandom coefficient models for multilevel analysis. Journal of Educational Statistics, 8, 103-120.

Taverner, S. and Wright, M. (1997) Why go modular? A review of modular A-level Mathematics. Educational Research. 39(1), 104-112.

Taylor, F. W. (1912). Scientific Management. New York: Harper.

Taylor, R. G., Pease, J., & Reid, W. M. (1990). A Study of the Survivability and Abandonment of Contributions in a Chain of Delphi Rounds. Psychology, (27), 1-6.

Taylor Fitz-Gibbon, C., & Vincent, L. (1997). Difficulties regarding subject difficulties: developing reasonable explanations for observable data. Oxford Review of Education, 23(3), 291-298.

Thorndike, E. L. (1913). Original tendencies as ends: emulation in the case of school "marks". In Thorndike, E. L. Educational Psychology, 286-9 New York: Teacher College, Columbia University.

Tyler, R. W. (1931). A generalized technique for constructing achievement tests, Educational Research Bulletin, 10 (8).

Tyler, R. W. (1949). Basic Principles of Curriculum and Instruction. Chicago: University of Chicago Press.

Tymms, P., & Vincent, L. (1995a). Comparing examinations boards and syllabuses at A-level: Students' Grades, Attitudes and Perceptions of Classroom Processes. Belfast: Northern Ireland Council for the Curriculum, Examinations and Assessment.

University of London Entrance and School Examinations Council. (1972). Advanced level physics comparability study. London: University of London.

Waterhouse, M., & Crook, G. (1998) Mapping Vocational Qualifications Against Academic Programmes. 24 April, University of Central England in Birmingham.

Watson, J., McEwen, A., & Dawson, S. (1994). Sixth form A level students' perceptions of the difficulty, intellectual freedom, social benefit and interest of science and arts subjects. Research in Science and Technological Education, 12(1), 43-53.

Weber, R. (1985). Basic Content Analysis. Beverly Hills: Sage Publications.

Willmott, A. S. (1977). CSE and GCE grading standards: the 1973 Comparability Study. Macmillam Education.

Willmott, A.S. (1980). Twelve years of Examination Research. ETRU, 1965-1977. London: Schools Council.

Winship, C., & Mare, R. D. (1984). Regression models with ordinal variables. American Sociological Review, 49, 512-525.

Winter, R. (1993). The Problem of Education Levels (Part 1): Conceptualising a Framework for Credit Accumulation and Transfer. Journal of Further and Higher Education, 17(3), 90-103.

Winter, R. (1994). The Problem of Educational Levels, Part II: A New Framework for Credit Accumulation in Higher Education. Journal of Further and Higher Education, 18(1), 92-106.

Winter, R., & Maisch, M. (1996). Professional Competence and Higher Education: The ASSET Programme. London: The Falmer Press.

Wolf, A. (1995). Competence-Based Assessment. Buckingham, Philadelphia: Open University Press.

Wolf, A. (1996). Vocational Assessment. In. L. T. Goldstein H. & Lewis, T. (Editors), Assessment: Problems, developments and statistical issues: a volume of expert contributions. Chichester: John Wiley & Sons.

Wolf, A., Sylva, K., Jones, P., Wakeford, R., Harrison, A., & Dockrell, J. (1997). Assessment in higher education and the role of 'graduateness'. Higher Education Quality Council.

Wrigley, J., Sparrow, F. H., & Inglis, F. C. (1967). Standards in CSE and GCE. English and Mathematics. Working Paper 9. London: HMSO.