Policy and New Products

Research Report 6

Marks into Grades

A Review of the Underlying Issues

Executive summary

Grades and grading

This report was produced by Dr Mike Kingdon, Principal Education Consultant, Entity Group Ltd, in March 2009.

Given the inevitable annual variations in the standards of individual examination components, most writers concur that marks alone are inadequate for reporting results. Converting marks to grades endows results with greater meanings — across diets, options, subjects, and in many cases across qualification streams — than marks alone can provide.

There is also a consensus that grading (many organisations use the term 'awarding') is the process of converting raw marks for components into component and subject grades. In the traditional examination cycle, grading is one step in the qualifications delivery process. In common with other UK regulators and awarding bodies, the principles underpinning SQA grading are founded in case law and informed by developments in psychometrics and ICT. Given its unique national status, SQA and its forerunners have been able to implement quality systems that have been impossible for other UK awarding bodies. However, despite this immense administrative and intellectual legacy, grading remains a judgemental process.

Efficient grading depends on the design of the assessment regime, the validity of the assessment components, plus the reliability and completeness of marking and marker standardisation. In their turn, the efficiency of grading decisions can be evaluated in terms of the:

- **Comparability** of the grades awarded — whether over time, across all of the routes that lead to the same subject grade, across recognised points of equivalence with other qualifications, or across other subjects at the same level
- **Validity** of the aggregation process —the greater the number and heterogeneity of the components to be combined into one grade, the more difficult it becomes to achieve adequate validity
- **Discrimination** of the final/subject marks — specifically the number of marks between adjacent grade boundaries
- **Robustness** of the grades

The grading of Scottish school leaving qualifications is subject to several challenges not experienced to the same extent in other UK systems and the majority of other national systems. First, the Scottish system is unique in having six overlapping streams of school leaving qualifications. As a result there are more grades to award. Second, there are recognised points of equivalence between the streams that need to be comparable. Third, UCAS recognises some of the Scottish grades as equivalent to those of other UK systems. Traditionally, SQA (and its forerunner) have participated only occasionally in the comparability studies of school leaving qualifications organised by other UK awarding bodies,

or in any of the comparisons organised routinely by QCA. A final UCAS benchmarking for the qualifications introduced in 2000 was planned for 2008. However, SQA does use systems of national and international benchmarks to monitor standards over time and between countries. All of these systems for monitoring the standards of SQA examinations continue to evolve. Nevertheless, users' assumptions about grade equivalence should be treated with caution.

A general concern is expressed about the ways many users of SQA and other UK qualifications are treating reported grades. UCAS tariffs and most value-added calculations assume that grades are at the interval or even ratio levels of measurement. SQA more modestly claims only to be assessing at the interval level. (These concepts are explained in Annex 2.)

Section 3 of the report investigates different approaches to the awarding of grades for components, subjects and groups of qualifications. The position of SQA's National Qualifications on the norm-referenced, criteria-referenced continuum is determined. The implications for grading following from the use of criteria rather than norms are discussed, with an example. Possible methods of extending the principles of mark-based grading to competence-based awards are proposed.

It is argued on resource and psychometric grounds that for grading to be efficient the numbers and variety of components used to assess particular qualifications and subjects should be kept to the minimum compatible with adequate coverage of the content and skills that characterise them. Similar cautions are made about the aggregation of grades from heterogeneous subjects to form group awards. Annex 2 discusses the psychometric background to these arguments.

The future

Ongoing societal changes discussed in Section 3 continue to create new needs and assessment opportunities. Section 5 discusses a number of new developments — in qualification delivery and responses to perceived needs — that will all create their own grading demands.

Background material

Annex 1 reports an analysis of the grading systems of school leaving qualifications from 62 countries outside the UK, which considers 72 different national, regional and other systems. Three approaches to grading are identified — percentiles, quality of work and marks. In common with the majority of countries that use grades, SQA bases its grades on the quality of candidates' work, moderated by percentiles. It is also concluded that, with a family of five overlapping streams of school leaving qualifications, the current Scottish system is the most complex of those investigated.

Annex 2 contains a detailed discussion and example of the technical concepts and criteria involved in grading.

Contents

1 Introduction

Grading has always been central to the work of all awarding bodies, whether it is the grading of external exams or that of internally produced evidence or coursework. However, some of the greatest grading challenges are in the immediate future. E-assessment, increasing modularisation (unitisation) of courses, greater use of selection tests, the assessment of functional skills, diplomas, on-demand assessment — whether it is, or is not, part of personalised learning and assessment, credit-based continuing professional development (CPD) schemes that build towards recognised qualifications and time-limited qualifications, will all require their own approaches to grading.

The purpose of this paper is to assist SQA to anticipate the grading challenges that the new developments in qualifications, assessments, needs and opportunities will create.

1.1 Origins of grades in Scotland and England

The decisions to report the results of school leaving examinations using grades rather than marks were made in 1960. The Scottish University Entrance Board (SUEB) was already using grades for its examinations but the implementation of grades for Scottish Highers was postponed until the introduction of the Scottish Certificate of Education two years later.

In England, the Secondary Schools Examination Council (SSEC) approved the use of grades in 1960 from that summer's A-level examinations. Three of the English GCE awarding bodies made the decision to use grades to report their O-level results as well, but the SSEC did not concur so the awarding bodies in question were not allowed to report the grades on their O-level certificates until much later.

2 Definitions of grades and grading

Grades are understood by most people as indicating the level of performance which has been achieved. A Grade A for instance is widely recognised as the highest level of achievement, while a Grade C is lower. In many cases a grade is understood to also convey whether the level of performance is acceptable. In Scotland, for example, Grade C has traditionally been the lowest grade regarded as a Pass.

Grading or awarding is defined by many authors as the process of deciding where to place grade boundaries. Examples include:

> *Grading is the process by which mark scales are divided into mark bands, such that marks within each band represent a particular grade. The minimum mark required for the award of each grade is known as the 'grade boundary' mark. Grading, therefore involves the identification of grade boundary marks.* Newton (2007) p26

> *The awarding process is one of weighting the evidence and coming to judgement on where to locate grade boundaries.* Robinson (2007)

Robinson (2007) also outlined the fundamental grading dilemma facing such experts:

> *Although the awarding process relies heavily on awarders' understanding of their subject, it has been shown that even experienced examiners are relatively poor at making judgements about the 'grade worthiness' of a piece of work on one mark compared with another without something other than their subject matter understanding to guide them. In the past examiners had somehow, simultaneously, to account for both the quality of the candidature and the difficulty of the paper. This was essentially a circular task, as the examiners had nothing to judge the quality of the candidature by except the performance of the candidates on the paper whose difficulty they did not know.* (p 121)

Bruce (1969) listed the functions of grading to be, *'placing the mark boundaries for a grade to take account of:*

1. *The year-on-year differences in the difficulty of an individual subject;*
2. *The differences between (any) different options* (or syllabuses) *for the same subject;*
3. *Different subjects examined in the same year.'*

Bruce (1969) also considered how different examination boards in England needed to alter the proportions of candidates awarded particular grades for individual subjects in individual years to reflect:

4. *Trends in the numbers of candidates and the types of centres from which they came;*
5. *How the candidature for their board might differ from those of others.*

Petch (1953) in a more discursive consideration of the same issues, to which Bruce referred, provided tables to illustrate the year-on-year and subject-by-subject variations in percentile marks and changes to candidatures.

Efficient grading depends on the design of the assessment regime, the validity of the assessment components, plus the reliability and completeness of marking and marker standardisation. In their turn, the efficiency of grading decisions can be evaluated in terms of the:

- **Comparability** of the grades awarded — whether over time, across all of the routes that lead to the same subject grade, across recognised points of equivalence with other qualifications, or across other subjects at the same level
- **Validity** of the aggregation process — the greater the number and heterogeneity of the components to be combined into one grade, the more difficult it becomes to achieve adequate validity
- **Discrimination** of the final/subject marks — specifically the number of marks between adjacent grade boundaries
- **Robustness** of the grades

Comparability will be discussed in this chapter, the other factors that determine the efficiency of grading decisions (aggregation, discrimination and robustness) will be discussed in Section 4 and in more technical detail in Annex 2.

2.1 Grading and comparability

After issues of validity and reliability have been addressed earlier in the examination cycle, comparability is the next most important issue underpinning fair and equitable grading.

Bruce and Petch were referring to (but not naming) aspects of comparability. Robinson (2007) makes a direct link between grading and comparability:

> *'It (grading) is an essential part of the comparability process: raw marks are too variable to be relied on.'* Robinson (2007)

Current authors distinguish several forms of comparability:

- **Comparability over time** — This is a concern for all qualifications, especially when the candidature is changing — not just increasing or decreasing but changing in its nature —eg its gender balance, the types and locations of centres submitting candidates and/or the other subjects and qualifications that the candidates might be studying.

- **Comparability between different syllabuses for same subject and level** — This is not confined to the English educational system, in which several examination boards provide syllabuses and exams being awarded the same qualification — even the same board may provide alternative syllabuses and

exams. It can be generalised to apply to all the varying routes there may be to the award of a grade for a particular subject and level, including tiered papers and unitised courses, group awards or frameworks. ExoD (2005C) have pointed out that the introduction of e-assessed courses will create new varieties of this general form of comparability.

- **Comparability between the syllabuses of different awarding bodies (instead of between examination boards)** — At first sight this form of comparability is not of concern to SQA, at least not for most qualifications intended for school leavers. It does apply however to a range of general and vocational qualifications for which equivalent alternatives are offered by other organisations. (See sections 'Implications for SQA' below.)

- **Comparability between subjects** — The UCAS university entrance system, and most measures of value added, assume that the same grade awarded for different subjects from the same qualification level implies equal attainment. The standard method used by all UK awarding bodies to evaluate this form of comparability is subject-pairs analyses. Each subject is paired in turn with all the other subjects, with which it has common candidates and a weighted average grade difference is calculated. The equivalent SQA method is known as 'National Ratings'. Differences that exceed a criterion level indicate potential leniency or difficulty. Any adjustments that might be necessary to achieve inter-subject comparability are made at the subsequent grading meeting. The technique is not without its critics so most awarding bodies limit the subject pairings used in the analyses to those that are based on adequate numbers of pairings and where the correlation between the two subjects is sufficiently high to indicate good predictive validity between them.[1]

The approach adopted by almost all awarding bodies is to use a variety of statistical and historical reports to question or confirm the judgements of quality made by the subject experts. Robinson (2007) summarises the current QCA position:

> *'Awarders are now in a position to test their decisions in a number of ways, retaining their professional input, but complementing it with the power of quantitative analyses to come closer to the goal of comparability that they are seeking.'* (p 121)

The issue with Robinson's statement is whether equivalent levels of comparability could have been achieved using marks. The consensus is that they could not.

2.2 Reporting using marks in the English school system

Prior to the introduction of the first school leaving grades in 1960, the results of UK school examinations were reported to candidates and their schools in terms of

[1] The discussion of different forms of comparability is based on Kingdon (1991) and updated with ideas from Newton (2007A & B).

marks. However, how the marks were interpreted and certified differed between national systems and awarding bodies. For example:

- At GCE Ordinary Level 45% was a pass and 75% generally recognised as a distinction, but only pass results were reported on certificates.
- At GCE Advanced Level 40% was a pass, 75% was generally recognised as a distinction, but again only pass results were reported on certificates. Marks between 30% and 40% led to the compensatory award of 'Allowed Ordinary Level'[2].

Inevitably, each year, each subject differed in the average number or percentage of marks achieved and in its mark distributions, so almost all sets of marks had to be scaled to fit the reporting parameters. University students were employed to scale the marks and they seem to have worked in one of two ways (depending on the awarding body). They were given either:

- a conversion table which mapped each raw mark to a scaled mark, eg 'A raw mark of 49% becomes a scaled mark of 51%'

  Or

- ranges of marks to which a constant was to be added, eg 'For raw marks between 41% and 50%, add 2 marks'

By reporting using grades it became possible to impose levels of inter-subject, and inter-year comparability, not only at the pass and distinction levels, but also at other points along the distributions of marks.

There were also very practical reasons for the move to grades. The attraction of a computer that could perform the mark conversions was obvious, but the University of London's systems analyst proposed a simpler solution. He suggested that grades should be mapped to hypothetical mark ranges. Rather than scaling the raw marks, eg changing 49% into 45%, the mark boundaries for the grades could be altered, say, by changing the boundary of the lowest Pass grade from 45% to 49%, to reflect annual and subject differences. When the SSEC accepted the use of grades for A level, a nine-grade system was agreed. There were six grades of pass, starting with 1 (distinction) to 6 (raw pass), and three grades of fail. The two higher fail Grades 7 and 8 were equated to the 'Allowed Ordinary Level'. The University of London introduced the system on a trial basis for its January 1960 A-level examinations and, with two other awarding bodies, used grades for their O-level examinations as well.

The SSEC quickly became concerned about the narrowness of the ranges of Grades 2 to 5 — just 5 percentage marks — and the unequal size of Grade 6 (40 to 49 marks). To address these concerns a five-grade system, with pass Grades A to E, based on percentages of candidates instead of marks, and two grades for fail (O and F) were introduced from the summer 1963 examinations. Subsequent

[2] It was assumed — wrongly — that schools would bypass O-levels in candidates' specialist subjects and prepare them directly for A levels.

evaluations suggested that the achievement of inter-year and inter-subject comparability was easier using five grades that the original six, unevenly spaced, grades.

It is of interest here that the first Scottish school leaving grades were also based on a nine-point system. Indeed, marks for National Qualifications are still converted into nine bands as well as into grades, and both grades and bands are reported to centres and to UCAS. Grades A, B, and C have two bands each, Grade D is just one band wide (band 7). Bands 8 and 9 are a 'No Award'. Band 9 ranges down from band 8 to zero.

2.3 The uses of grades

The claimed functions of examination grades are to:

- provide a standardised measure of a student's performance
- certify that a course of study has been completed
- certify that particular standards of attainment have been achieved
- *certify that prescribed content has been mastered and/or defined skills have been demonstrated*
- *give common meanings to standards achieved in different options and subjects*
- certify readiness to enter further education (FE) or higher education (HE) and particular forms of employment
- select from among qualified candidates for entry to high-status or popular courses/employment

Several of the above are contentious and the two most so are italicised.

The extent to which candidates, who have achieved particular grades in *the prescribed content* for a subject, should be expected to demonstrate particular skills in new contexts remains an issue for debate, that is whether exams should measure exam preparation or a more general level of ability within the subject, such as problem solving or communication.

Similarly, the expectation that grades have *common meanings* across subjects is popular with the users of qualifications and many selection systems are based on that assumption.

2.4 Marks and grades compared

Many contemporary educationalists in England were critical of the change from marks to grades, typically claiming that information was lost. However, the relationship between marks and grades is complex:

- Given the year-by-year variations in difficulty of exam papers of individual subjects, plus the differences in the spread of marks and shapes of the distributions of marks across different subjects, even scaled marks towards the middle of the range of pass marks may only have meaning within the

particular subject and/or diet. For instance, 55% may be above average in a difficult subject, especially in a difficult year, but below average in an easy subject, or in a subject only chosen by well-motivated and well-prepared candidates. It must also be remembered that the marks[3] used to report school results up to the early 1960s had already lost some of their subject meaning but gained inter-subject equivalence at the pass and distinction points. Subjects had started to use similar percentages for their pass and distinction points.

- The process of grading is an attempt to address inter-subject, inter-option and inter-diet differences and impose a common meaning to reported results.

- Both marks and grades indicate orders of ability, in that 75 marks is better than 70 marks, and Grade A is better than Grade C. However, apart from that, not all grades span the same range of marks; equal steps in marks or grades do not necessarily imply equal steps in ability. For instance, an increase from 10 to 15 out of 100 marks might be much easier to achieve than an improvement from 90 to 95. The step from a Grade D to a Grade C in SQA's National Qualifications (from just below a pass to a pass), is unlikely to be equal to the step from Grade B to Grade A. It is of concern that many of the uses to which grades are put, eg university entrance and school improvement analyses, assume that differences between grades correspond to equal differences in ability (See Annex 2 for more details on levels of measurement.)

In Scotland some of the above issues are mitigated through the bands used to report results to centres.

In conclusion, while detail may be lost within an individual subject examination, because grades show fewer gradations in achievement than marks, turning marks into grades increases their comparability across years and subjects. A total of 50 out of 100 marks in one subject does not have a clear meaning when other subjects score out of different totals, require at least 60 for a pass, or when this year's exam paper was exceptionally easy.

2.5 Use of grades in other countries

Annex 1 contains an analysis of the school leaving qualifications of 62 countries — excluding the UK — and 72 different national, regional and other variations. The conclusion is drawn that Scotland's current family of school leaving qualifications is the most complex of the systems analysed.

Annex 1 reveals other differences between the Scottish and other UK grading systems. While England, Wales and Northern Ireland use five grades of pass for

[3] To understand just how much scaling might have altered marks, the author tried to replicate the process used by the English awarding bodies prior to 1960, using results from a number of different subjects. The resultant distributions showed significant spikes and/or gaps as a result of instances when:

- more than one mark on the raw distribution mapped to a single mark point on the scaled distribution, producing a spike in the distribution
- a mark on the scaled distribution did not map to any mark points on the raw distribution, producing a gap

their GCE A- and AS-level examinations, Scotland uses only four grades for its Intermediates, Highers and Advanced Highers. As the Scottish qualifications overlap, and Grade D as equated to Grade A at the level below, 12 grades are used in total. There are further equivalences between the two Intermediate levels and four of the grades used in the Standard Grade qualification. See Table 1 below for details.

2.6 Implications for SQA

The definitions of grades and grading used in SQA and elsewhere in the UK are the same. Indeed, through the regular meetings of staff at all levels in the qualifications systems research, documentation and understandings have been shared. While QCA and the school leaving awarding bodies for the other three countries make repeated and explicit references to all forms of comparability, in SQA documents such as SQA (2007B) and SQA (2007C) references to comparability between different syllabuses, units, or exam options for same subject (and level), and comparability with other awarding bodies are mostly implicit.

With respect to the former, the existence, currently[4], of five overlapping school leaving qualifications, means that comparability issues exist whenever grades in the different qualifications are equated as being equal. UCAS (2008) provides tariffs for 10 different pass grades drawn from the school leaving qualification streams.

The underlined grades in Table 1 are equated by UCAS with the same numbers of points as GCE A level Grades A to D, and other GCE A/AS and GCSE grades map to intermediate positions (see http://www.ucas.com/students/ucas_tariff/tarifftables/). Two issues for SQA are:

1. The extent to which the common UCAS tariffs for some Higher/Advanced Higher and GCE A-level grades should become a factor in the future grading of the Scottish qualifications
2. Whether the assumptions of equivalence being made by UCAS should be tested formally in joint Scottish/English grading studies

In 2008, SQA's Advanced Higher and Higher exams were compared with exams from another UK examination board in order to establish comparability according to a new set of procedures and criteria[5]. The current tariff point had been established temporarily, on the basis of an analysis of qualifications which were to be introduced in 2000.

Finally, current developments in qualifications mean that new comparability and therefore grading issues are constantly being created.

[4] In 2008, the Scottish Government issued a consultation on the new generation of qualifications. It proposed to replace all Standard Grades and Intermediate exams with just two exams: General and Advanced General.
[5] No report available at the time of writing.

Table 1: equivalences between SQA grades and their UCAS tariff points

Advanced Higher	Higher	Intermediate 2	Intermediate 1	Standard Grade	Access	UCAS Tariff
A						120
B						100
C						80
D	A					72
	B					60
	C					48
	D	A				42
				Credit (1)		38
		B				35
		C		Credit (2)		28
		D	A			
				General (3)		
			B			
			C	General (4)		
			D			
				Foundation[6] (5, 6)	Access 3	
					Access 2	
					Access 1	

[6] Foundation is assumed to be equivalent to 'Access', but this is not externally assessed or graded.

3 Ongoing pressures for change

3.1 Societal issues in the UK

The first UK public examinations were university initiatives and date from as early as the 1830s. They were intended to regulate entry to university but quickly became used for other certification and selection purposes. English university examination boards introduced school examinations that anticipated the twentieth century School Certificate (SC) model, from 1858, while Scottish universities favoured the School Leaving Certificate (SLC) model (introduced from 1864). From 1911 the SC and Higher School Certificate (HSC) model became the standard for England, Wales and Ireland. In Scotland a centrally delivered SLC was preferred. A common feature of both systems until the 1950s was the use of group awards.

Across the British Empire and later the Commonwealth, the SC/HSC examination model has been used to recruit, select and train small elites to govern and administer individual countries, officer their armed services, manage their factories and businesses, staff their professions, and teach in their schools. As late as the 1960s, only the more able were entered for school examinations, and examiners had become skilled in discriminating between the candidates to be awarded different grades. As a result, candidates at the lower end of the prescribed ability ranges typically gained very few marks for subjects that they might have studied in secondary schools for up to five years. The negative mark schemes used for some question types and examination components[7] must have reinforced the general sense of failure felt by many.

The move to technological and information-based societies meant that significantly larger proportions of the populations had to be educated and certified to levels that had previously been reserved for national elites[8]. As a result, access to qualifications was expanded, examinations were required to assess what the candidates could do rather than what they could not, and grades had to be awarded for positive achievements.

Developments within qualifications have paralleled the societal changes:

- In England, Wales, Northern Ireland and Scotland, there was a move from group to single subject awards.
- On both sides of the England/Scotland border:
  - there were strong movements from:
    - informal to formal — and certified — vocational training

[7] Unless stated otherwise, the term component includes units of qualifications.

[8] In the 1980s Mauritius was still predominantly an agrarian society. 8% of boys and 8% of girls in Mauritius were educated to school certificate/GCE Ordinary level standard. About half of each cohort went on to study for HSC/A Levels and less than half of these went to universities in the UK, USA, Australia and India. Even with a significant proportion of the graduates deciding not to return, this left enough qualified personnel to staff the island. Today, Mauritius seeks to be the ICT leader for Africa and one of its links to the rest of the world. It is also a popular tourist destination. As a result, Mauritius now requires that 40%, preferably 50%, of each cohort are trained to the level of the previous elite in order for the island to staff its developing service industries and infrastructure. Inevitably, opportunities for emigration have become tightly constrained.

  - emphasising content to emphasising skills
  - restricted access to qualifications to more open access
- the numbers of professional and vocational qualifications increased many-fold in number and complexity
- vocational qualifications enjoyed greater esteem and became accepted routes for entry to HE and high-status employment
- the numbers of subjects available for study within individual professional, vocational and school qualification streams increased
- within individual qualification streams — including school qualifications — the numbers and variety of assessed components increased[9]
- syllabuses and arrangements expanded from a page or so in the 1960s to many pages in the 2000s

The arguments for a greater number and variety of assessment components were that assessing new skills and assessing established skills in new ways — both possibly over periods of time — increased validity. However, the claimed gains in validity may have had implications for the same qualifications' reliability, authenticity, and discrimination (see Section 4.8). In Section 4.9.5 it is argued that grading is dependent upon all components being 'addable'. (A theoretical justification for these statements is provided in Annex 2.)

3.2 The development of vocational and professional qualifications

Both England and Scotland have strong traditions of professional qualifications. Only in the UK, the Commonwealth and USA have groups of co-workers gathered together to share craft and professional issues and grow into nationally, even internationally, recognised institutions independently, without being supported by, or absorbed into, state structures. Alongside the qualifications provided and/or regulated by states, England, Scotland and many other English-speaking countries have a wealth of regulated and unregulated craft and professional provision.

The methods of assessment adopted by the developing craft and professional qualifications owe their origins to earlier — often informal — craft or university approaches. As a consequence they have different assessment priorities, can be idiosyncratic, and may emphasise workplace or professional validity over reliability. Nevertheless, they have often pioneered assessment and grading methods. Some of the UK industry-specific awarding bodies have been among the first organisations to adopt e-learning and on-demand, on-screen assessment. For some, on-screen assessment has become their main means of qualification delivery.

[9] Kingdon (1991) and (2000) reported that the numbers of components in English school examinations are doubling about every 16 years.

3.3 Developments in vocational and professional qualifications

Current developments in the vocational and professional domains include:

- credit-based continuous professional development (CPD) schemes offering both horizontal and vertical learning paths
- time-limited certificates to practise, and
- employers and their clients accessing national databases to confirm that employees' qualifications are valid and current

3.4 Service issues

Awarding bodies have also had to become more efficient. Not only have they had to develop ways of delivering more, and more complex, qualifications to ever more candidates without extending the timescales, quality changes have had to be achieved without proportionate increases in cost. The driving forces in England have been fierce competition between awarding bodies and the opportunities offered by new technology. The effects of the former have been mixed. While competition has kept the English awarding bodies lean and hungry for entries, they have not always been able to adopt all the more expensive quality procedures and have had to wait for ICT-driven alternatives.

UK awarding bodies are now exposed to new risks and threats. Most of these follow from the growth in qualifications activity and that increases in resources to deliver the qualifications — human, technical, time, financial and physical — will fail to keep pace with expectations.

3.5 New challenges

Developments elsewhere in the UK and Europe could be threats to SQA's status if it failed to respond. Examples include:

- Demands for quicker turn round of results
- Demands for feedback on results from:
  - candidates — to inform their learning
  - centres — to evaluate and develop their teaching
  - society — to monitor national standards and improve the curriculum
- E-leaning and e-assessment
- On-demand assessment
- Personalised learning and assessment
- Greater input by employers into the learning and assessment processes
- Credit-based CPD leading to formal qualifications
- Time-limited certification of competence to practise crafts and professions

Grading — or at least the formal recognition of achievement — is at the heart of all of the above.

3.6 Implications for SQA

Ongoing changes within the Scottish qualifications system are inevitable as SQA responds to accelerating societal needs, educational developments and curriculum reviews, service issues and the opportunities offered by information technology.

SQA does not have to compete with other UK awarding bodies for candidate entries, in non vocational qualifications, so is able to institute improvements in its procedures which would not be possible for awarding bodies in England. Similarly, SQA does not have to compete with the UK awarding bodies in how it uses the resources of technology, time and cost. As a result, SQA is in a position to improve on what other UK awarding bodies provide. However, Scotland's different ratio between resources and candidate numbers imposes limitations.

UK-wide trends to increase the numbers and variety of components used to assess qualifications are a case in point. While there may be valid educational reasons for increasing and diversifying components, the extra resources required have to be justified. Further, it is reasoned in Section 4.9 below that there can be tensions between use of more complex component structures and maintenance of a qualification's psychometric integrity.

4 Approaches to converting marks to grades

Countries across the world show considerable variation in how grades are awarded. For example, three distinct approaches are used to generate initial component or subject grades (see Annex 1):

1. Use of percentages of candidates, eg equating the most able 10% with the highest grade
2. Use of traditional or written criteria against which to assess the quality of candidates' work
3. Use of fixed conversions of marks to grades, eg by always equating 45% with a pass

Further for multi-component qualifications, some awarding bodies grade components and then aggregate the resultant profile to form the subject grade, while others combine the component marks to form a subject total and award a single subject grade. For the convenience of this paper, the former will be referred to as component/subject method of grading and the latter as the direct subject method.

Across the world, the first or second of the above approaches to grading predominate as the basis of grading (see Annex 1). However, whichever of the three methods is preferred, most awarding bodies use one of the other two to question or confirm the mark points generated by their preferred method. The issue, unfortunately, is that even using pairs of methods, none of the possible combinations guarantee the accuracy of the grades reported.

It is, of course, possible to use all three approaches at different stages in a single grading process. In England, the SSEC method of awarding A-level grades remained in place until the summer of 1988. The SEC — the SSEC's second replacement — imposed a new system of grades for A level, following concerns about the narrowness of many mark boundaries, and thus the low discrimination of the examinations. The unusual feature of this system was the use of differing methods to set individual grades:

- Grades A, B and E were all set by quality of work
- Grades C, D and a new Grade N — a fail grade which replaced the Allowed Ordinary grade — were set using ranges of marks
- Grade U (for unclassified) was given to those candidates who failed to achieve any of the higher grades
- All Grades A to N were moderated using percentages of candidates

4.1 Awards based predominately on norms

All UK awarding bodies are accused of norm referencing or not norm referencing, depending on whose prejudices have been offended. The concerns of UK teachers are often that particular grade boundaries are set using proportions of candidates

while the inherent quality of the candidates' answers is ignored, which would be norm referencing. The concern of media pundits is usually that the proportions awarded higher grades are varying, which would mean that there is no norm referencing, and/or that any increases in the percentage of candidates achieving top grades must mean that standards are declining.

Perhaps the most extreme example of the norm-referenced approach to grading is grading by ranks most often found in the USA. Candidates are ranked according to their scores. Grades are awarded using fixed proportions, eg the best 10% of the candidates are awarded Grade A, the next 20% Grade B, the next 40% Grade C, the next 20% Grade D and the remaining 10% Grade E. This method does not take the quality of their work into account. There are usually clear instructions about how remaining candidates are to be graded when their numbers are not exactly divisible by 10. The method is used in some US school systems to prevent grade distortion and has been used in Sweden to impose social equality on university selection procedures. It has also been used in Africa and Asia where access to higher education and/or employment has to be restricted because of lack of opportunities and/or resources.

The main criticisms of grading by ranks are that:

1. Small differences in candidates' performances can lead to large differences in the grades awarded and therefore the opportunities offered to the affected candidates. While some discrepancies between performance and grades are characteristic of all grading systems, they are very unlikely to approach the magnitude of those generated by this technique. Further, grading by ranks lacks the moderating mechanisms usually associated with other approaches to grading
2. Except at the extremes of the ability range, the grades are likely to lack comparability:
    a. between different groups of candidates, and
    b. over time

4.2 Awards based predominantly on criteria

The perceived alternative to norm referencing is criterion (more accurately 'criteria') referencing. The most commonly quoted example of criteria referencing — the UK driving test — is based on a concept of normal driver behaviour. Wood (1991) suggested that if grades were ever to become fully criteria referenced their meaning would become so embedded within the contexts of individual subjects that the first casualty would be inter-subject comparability.

The more practical matter is that criteria have proved to be notoriously difficult to write and use. Attempts to express in writing the characteristics of responses awarded particular A-level grades, in other than very general terms, have failed. More worryingly, very small changes in the wording of individual criteria can produce significant, and unpredictable, shifts in the number of candidates who meet them. Full criteria referencing of assessments of the complexity of school examinations probably remains a theoretical model.

However, as Wood also pointed out, norm and criteria referencing represent opposing ends of a continuum; norms are often defined in criteria terms and vice versa. Current SQA grading procedures are located in the middle range of the continuum and more towards the criteria referenced end. Its grade-related criteria are general definitions, which are complemented by marking guidelines for specific exam papers. In addition, the percentage of candidates to be awarded a grade on the basis of their scores and the quality of their work is compared with the percentage of candidates usually being awarded the grade.

4.3 When norm referencing and criteria referencing conflict

As the following example illustrates, the reporting of the first level 3 results for the New Zealand Certificate of Educational Attainment (NCEA) in January 2005 provided a case study of how, as Wood predicted, criteria referencing can generate wide inter-subject differences in the results. University entrance in New Zealand is based on candidates' performance at NCEA level 3, but an extension of the examination — formerly called the scholarship examination — is used to award competitive bursaries to part-fund university fees. When the results of the NCEA equivalent of the scholarship examinations (variously referred to in the media as level 3+ or level 4) were published, the proportions of language candidates receiving awards were in the 60 to 75% range, while the proportions for science candidates were in single figures. Only 2% of Biology candidates passed — and a significant proportion of them only on appeal. The social science and humanities candidates and mathematicians were distributed between these two extremes.

The media storm can be imagined and administrative failings were soon revealed[10]. The immediate response of the government was to create further scholarships for the subjects most affected. Their longer-term response was to restore a strong measure of norm referencing and review the criteria[11] [12].

4.4 Awards based predominately on marks

There are still UK awarding bodies and universities that use fixed mark points for their awards and do not acknowledge any diet-by-diet differences in the standards of their examinations. Proponents claim that they can grade directly fail, borderline, pass and higher-level answers to questions irrespective of any variations in their difficulty.

When this approach is supported by an adequate understanding of the cohort's characteristics — such as a university lecturer might have of the final year students in a university faculty — there may be some justifications for it. However, when applied to external candidatures, and without other evidence to

[10] When I visited the NZQA in January 2007 the 're-calibration of the criteria' was progressing. The initial problems were attributed to different groups having produced the curriculum, the assessment instruments and the criteria, plus the language criteria group's preference for the word 'or' while the scientists used 'and'!
[11] See Benson-Pope (2005) for the official report on the NCEA problem.
[12] For a fuller discussion of this issue see ExoD (2005B).

support the decisions being made, the approach becomes potentially dangerous. The risk is that the examiners will prejudge candidates using ephemeral criteria and award marks that reflect their prejudices. If the views expressed by Robinson (2007) can be taken as authoritative, this approach to grading no longer has QCA approval.

A further issue here is that preset mark boundaries — if only for pass/fail decisions — are re-appearing as part of e-assessment, specifically on-screen testing. Arguments for this form of delivery include the advantages of on-demand assessment, immediate results, rapid feedback on candidate performance, and ease of re-sits. It assumes that a test is available or can be assembled automatically which fits the preset mark boundaries (as well as a range of other criteria ensuring validity and reliability for instance). The hope is that the mark boundaries that underpin instant marking and the test assembly have been based on adequate psychometric evidence and that their use is kept under review[13].

4.5 The practicalities of grading

There are no theoretical models or mathematical formulae for converting assessment marks into subject or component grades. There are only principles and guidelines that can inform grading decisions[14]. The current principles and guidelines have not been developed from assessment theory. Instead, they are the product of the aggregated case law of past failures and issues, informed by improved understanding of assessment issues, developing concepts of good practice, and the opportunities offered by advancements in ICT.

When discussing grading, a much used concept is that of possible misqualifications, or errors. This concept is based on the two types of errors statisticians realised one can make when determining whether a particular observed pattern in data is significant or not (for instance, whether one marker is lenient compared to other markers):

- Type 1 errors in which a statistical test identifies a data pattern as significant although it is not, and
- Type 2 errors in which a statistical test does not identify a data pattern as significant although it is[15].

When applied to the setting of individual grade boundaries, Type 1 errors can be equated with leniency in grading (a Pass is awarded for work that does not meet the criteria) and Type 2 with severity (a Pass is not awarded, although the work does meet the criteria). Using this thinking, the responsibility of examiners becomes the setting of each boundary so that potential Type 1 and Type 2 errors in grading are minimised. In practice, of course, there is usually a zone of uncertainty a mark or so either side of a mark boundary. If the risks of Type 1/Type 2 grading errors cannot be reconciled, the response of most awarding

[13] For a report of QCA practice in re-calibrating on-demand unit assessments see ExoD (2005C).
[14] See for example QCA (2000), SQA (2003A & B), and SQA (2007A).
[15] See for example Guildford and Fruchter (1978).

panels is to err towards Type 1 errors. The finalisation procedure used by SQA is designed to identify candidates in this zone and correct both types of errors.

4.6 Grading academic examinations using decision theory

Attempts have been made to minimise the possibilities of Type 1 and Type 2 errors using decision theory. Decision theory in mathematics and statistics is concerned with identifying the values, uncertainties and other issues relevant in a given decision, and in the resulting optimal decision. The growing availability of computer workstations in the early 1990s meant that the production of examinations reports and statistics, and thus aspects of examination processing, could be devolved to groups of examiners. In particular, it became possible for examiners to interact with the IT systems to investigate the implications of potential grading decisions in real-time. For example, in its Decision Analytic Aids to Examining (DAATE) project, QCA and its immediate forerunners explored the use of decision theory to set grade boundaries by applying examiners' decisions about sample scripts to all scripts (see French et al (1990).

4.7 Grading competence-based qualifications

While the DAATE grade-setting process proved to be too protracted to become a standard method, it did generate new insights into how grading decisions are made. Building on this, Entity has generalised the DAATE approach to a wider set of situations, such as the evaluation of sets of competences. The particular example concerned decisions about which graduate members of a professional body, with widely differing training and professional experience, were 'ready' for promotion to chartered status. Using the technique awarding bodies will be able to identify which decisions are 'secure' and which are outliers that require further investigation (Entity (2008).

4.8 Evaluating grading decisions

Whatever the principles underpinning a qualification, all grading decisions depend on:

- the availability of adequate resources — human, technical, time, financial, and physical — to inform and implement them
- assessments with adequate psychometric efficiency — specifically validity, reliability, completeness, discrimination, and robustness

SQA's guide to assessment (SQA, 2008D) sets out the Authority's principles of assessment. These require external assessments to be:

1. valid
2. reliable
3. practicable
4. equitable and fair

Grading decisions should also exhibit:

5. completeness
6. discrimination, and
7. robustness

Production of grades that are valid, reliable, discriminating and robust is in turn dependent upon the marks on which the grades are based being as complete (as well as valid, reliable, discriminating and robust) as possible when grading takes place. There also need to be means to support and/or question grading decisions. All awarding bodies have their own lists of reports, statistics and samples that they have found useful for this task.

For immediate purposes a question, component or subject can be said to have discriminated if candidates acknowledged to have differing abilities are adequately separated by the marks and grades they got for the question, component or subject.

Similarly, grading decisions are robust when the number of changes of grade that follow single mark changes to any grade boundaries remain small.

Further definitions and explanations are provided in Annex 2.

The rules for the addition of the components (or units, modules, subjects) also need to be valid. Examples of these rules are, for instance, how to compute an overall Standard Grade, how to combine the (internal and) external components of an exam for a National Qualification, or whether or not to compute an overall grade for a Group Award consisting of a range of subjects.

Section 4.9.5 refers to 'addability'. This quality is not always easy to achieve. There is a tension between, on the one hand, using a large variety of very different components to cover a subject domain, and on the other hand, the use of one overall grade that discriminates adequately between candidates. Somewhere along the route that begins with the combination of two relatively homogeneous components and ends with a large number of heterogeneous ones, a qualification can cease to discriminate adequately, not using the full range of grades anymore but awarding fewer grades to more and more candidates.

The critical factor is the correlation between the components or subjects to be aggregated. Correlations tend to be low, for instance when many candidates perform well on the first component and poorly on the second, but many other candidates' profiles are the other way round. Each component can have a wide

range of candidate marks, but when the marks for both components are added up, good results will compensate for the poor results, and most candidates will in the end have a more similar, closer to average, total mark or overall grade. This is called 'regression to the mean'. The result is that the overall marks or grades are too close together to reliably discriminate or distinguish between candidates. The following discussion indicates where regression effects are likely to be an issue and alternatives to simply adding marks, bands or subject grades need to be found. Regression plus all of the terms used above are discussed further Annex 2.

4.9 Aggregating component grades into subject grades and subject grades into group awards

The very limited literature on grading contains few suggestions about how awards for individual components should be summed to create subject grades and how complete subject grades should be combined to create grades for group awards. Those that are to be found have often been written more for public relations than academic or regulatory purposes.

4.9.1 The advantages and disadvantages of component grading

There can be little doubt that grading of components and the resultant profiles are popular with candidates and centres. While component grades provide more information than a single summative grade, and give some clues to candidates' strengths and weaknesses, the difference is one of degree only.

The main disadvantages of component grades are that they are too coarse a measure of attainment to lead to practical strategies for improvement and that they fall a long way short of the detailed feedback that can be generated by e-assessment systems.

4.9.2 Aggregation of component grades to form overall subject grades

Reviews of awarding body literature revealed the following methods of aggregating component grades into subject grades:

- **Inspection** — worrying given the potential inaccuracies and biases but nevertheless still in use.
- **Conversion of component grades to points followed by averaging or direct aggregation** — the former is often referred to as the GPA (grade point average). Information loss can result and regression effects can be significant if the number of components to be aggregated is large and/or the inter-component correlations are low.
- **Using the component grades for information only and aggregating the component marks to form the subject grade** — this avoids the information loss associated with aggregating grades but is still dependent for its validity on the components being sufficiently homogeneous.

- **Using conversion tables in which the axes are grades for particular types of components** — eg coursework versus examination grades. The cells in the table show the combined result. Using this method it is possible to combine more heterogeneous components, and impose desired weightings and control for factors such as grade inflation. The main difficulty with the method is the problem of representing more than two components/dimensions on a single table.
- **Using profiles** — the components are separately awarded and reported. This was the approach adopted when the first Scottish school leaving examinations incorporated practical and oral examinations at the beginning of the twentieth century. More complex profile reporting systems can specify minimum criteria for the award of particular grades plus thresholds for individual components — defined in terms of marks, grades, study hours etc. Again, the validity of the approach is dependent upon any components that are to be directly combined being sufficiently homogeneous.

4.9.3 Aggregation of module/unit grades to form overall subject grades

Aggregation of units on modern modularised (often referred to as 'unitised') examination structures constitutes a particular case of component aggregation for several reasons. First, the units may have been completed over a period of time. Second, if the units are graded, as for A levels and GCSEs, the candidates already know some of their results and may have re-sat one or more of them to improve their ultimate grades. Third, one of the units may be synoptic and seek to assess candidates' abilities to draw on content and skills from across the other units, as in the graded unit within SQA's Higher National courses.

A current example of a modular-based grading system is the award of unitised A levels. Candidates can take three prescribed modules and achieve an AS (Advanced Supplementary) level award or all six modules for the full A-level award. Both levels are graded A to E (all passes) or N (no award).

Candidates' marks are converted into marks on the Uniform Mark Scale (UMS). The maximum UMS mark for each module depends on the proportion of the final marks that it contributes for the A and AS awards. So, if the three AS modules have the same weight and the total for the AS is 300, each module will have 100 UMS marks. The unusual aspect of this process is that the UMS marks have a fixed relation to grades. For a unit with a UMS mark out of 100, the ranges of UMS marks for each grade are:

E: 40–49
D: 50–59
C: 60–69
B: 70–79
A: 80–100

So, if the examiners decided that a raw mark of 90 out of 120 on a unit was the lowest mark for an A grade, then that 90 raw mark would become 80 when

translated into the UMS mark out of 100. Therefore, all issues of comparability between units are dealt with while setting grade boundaries for the units. (Robinson (2007) provides a worked example based on data from the AQA awarding body.)

Credit-based CDP schemes that build towards full qualifications may create further issues — the numbers and types of units to be aggregated may be considerably greater than anything considered so far, and the units may have to be time-limited so that they are still current when the final award is made. SQA's HND qualifications for instance consist of many Units, but only one of these is graded.

4.9.4 The advantages and disadvantages of group awards

The underlying issues are how breadth of study should be valued and whether skills can be transferred from subject to subject or component to component. Group awards are favoured when breadth is emphasised. Academic traditions that emphasise and reward (or have rewarded) breadth of study include most European systems, especially those that follow the Napoleonic tradition, the International Baccalaureate, and the Scottish School Leaving Certificates. Today, however, the strongest pressures for breadth come from the professional and vocational domains.

In many vocational and professional contexts, of course, both breadth and depth may be applicable. Practitioners may require breadth of knowledge and understanding so that they can respond to issues outside their particular expertise, if only to refer them to more qualified colleagues. Therefore, some measures of professional or vocational competence require candidates to demonstrate specialist knowledge in core areas and breadth of understanding —or rather minimum competence — of others.

Building on the comments of Bruce (1969), the issues with group awards are, given that some candidates will fail particular components of the award: what the centrality of these components is to the overall award; which value is attached to them by the users of the qualification; and what the effects are of their failure on the teaching of the overall subject.

4.9.5 Preferred methods of aggregating subject results to form group awards

Methods that are not typically used to aggregate subject grades to form a group award include inspection, aggregation of subject marks and use of conversion tables. The heterogeneity of the subjects to be aggregated for many group awards probably precludes all of these approaches.

The preferred methods of producing group awards are therefore:

- **Converting subject grades to points and averaging or summing them** — The UCAS system is based on the latter, and a similar method is used for evaluating the level and hours for which individuals' may have studied[16]. However, there are dangers in the approach. For example if too many different subjects are included and/or they are too heterogeneous, regression effects may mean that it becomes difficult to discriminate between candidates in the middle of the ability range.
- **Using profiles** — These can be very flexible tools for producing group awards and aggregating candidates' performances in different qualifications systems. When direct combination of some subject grades is not advisable due to their heterogeneity, hurdles can be set, based on completion, teaching hours or minimum performance. Again, the above caveat applies.

The issues with both of these approaches to group awards are:

- The 'addability' — specifically the validity of the aggregation procedures — of the components. While any sets of marks can be added it does not follow that the results will be meaningful. For instance, if a candidate presented an overall grade C for Maths, English, Art, and PE together, what exactly could that mean? The generally accepted view on what is 'addable' is that to be combined, the subjects (or components) need to be positively and significantly correlated. Expressed another way, the subject marks, bands or grades to be added need to have sufficient mutual predictive validity.
- Whether minimum performances in one or more components are compulsory. If such compulsory components represent a real barrier for significant numbers of candidates, they can significantly distort how they are taught. Bruce (1969) — somewhat cynically — suggested that making a particular component result compulsory can also lower its standards because examiners become reluctant to fail otherwise good candidates and invent ways to pass them. He quoted SC English as his prime example.

4.10 Implications for SQA

Current trends in the UK are emphasising the importance of breadth of study plus greater number and/or heterogeneity in the components to be aggregated into single awards such as Diplomas and Baccalaureates. The problem for SQA — and indeed all awarding bodies — is that the extent to which a final grade can be valid, reliable, complete, discriminating and robust depends on the psychometric properties and the Scottish Credit and Qualifications Framework (SCQF)[17] levels of the components to be combined. While the possible methods of combining components can be large, not all will produce a grade that is compliant with the aims of the qualification in question and current principles of good assessment practice. The solution to both the psychometric concerns expressed above, and the

[16] See SQA (2006) and QCA (2004A)

[17] See SCQF (2007) and Holmes (2007).

resource issues discussed in Section 3, is for the number of component grades or marks summarised in one overall grade to be kept to the minimum compatible with adequate coverage of the content and skills that characterise particular subjects and qualifications.

The issues are increasing. The introduction of e-assessment, unitisation and the other developments reported elsewhere in this paper are all likely to increase the number and/or the heterogeneity of the components that will be aggregated to form single awards.

5 New developments in assessment

Newton (2007B) identifies the following as some of the immediate challenges to current UK grading systems:

- E-assessment
- Unitisation
- Diplomas

To these might be added:

- On-demand assessment
- Personalised learning and assessment
- Credit-based CPD schemes
- Time-limited qualifications

On-demand assessment and personalised learning and assessment are considered together in Section 5.3. Credit-based CPD schemes and time-limited qualifications are considered in Section 5.2 on unitisation.

5.1 E-assessment

E-assessment is a collective term for a number of different applications of ICT to assessment:

- E-marking
- E-portfolios
- E-testing — also called ‘on-screen assessment’

5.1.1 E-marking

Many UK awarding bodies now electronically scan candidates’ scripts and make the images of individual question answers — instead of the scripts — available to markers. As markers input their marks online, their rates of working and quality of marking can be monitored. E-marking can generate significant gains in the reliability of the results produced. Specifically, marker standardisation can become an ongoing process that is integrated with the marking process. Feedback to markers is no longer so dependent on the physical movement of scripts. Supervising markers are able to have a genuine dialogue with their markers without being at the same location and cascade any updates to the mark scheme to all markers within a very short time.

E-marking has reached such sophistication in some awarding bodies that script images are split up, with answers to some questions being separated for automatic marking, some answers sent to clerical markers and others to expert markers, as required. Further, the awarding bodies have greater control of the marking process. They can identify marking or resource issues earlier and have the

flexibility to redirect images and resources at short notice. If necessary, images can be sent to markers across the world. The technique generates detailed information that can be used in grading, and to inform training of examiners and markers. Valuable feedback can also be provided to candidates and their centres.

While e-marking techniques can be expected to improve the quality of marking, empirical feedback from a small number of awarding bodies and examiners suggests that excessive separation (or 'decollation') of candidates' answers can make grading more difficult. Examiners responsible for grading appear to need access to complete sets of scripts, close to critical mark points, if they are to be able to evaluate the overall quality of candidates' work and feel confident about grading decisions.

UK regulators are already trying to anticipate how the current principles and guidelines for grading will have to be developed to encompass e-assessment. It is not clear at this stage how successful *a priori* principles and guidelines will be, given the speed with which technology is advancing and the relatively slow take-up of e-assessment. It will be interesting to see just how much the *a priori* statements come to be revised as experience with e-marking begins to build up.

5.1.2 E-portfolios

The term e-portfolio is being used in two ways. First, many UK awarding bodies use the term to refer to electronic methods of compiling project reports, coursework folders, samples of work and personal resource material that candidates may submit for assessment or use to support personal study. A number of proprietary systems exist to assist electronic compilation of those materials, plus their submission to an awarding body for assessment. To avoid issues of compatibility, some awarding bodies accredit particular systems. There are no specific grading issues associated with e-portfolios as long as the material submitted remains a component of another assessment. Therefore, they are not discussed further in this paper.

The second use of the term e-portfolio is specific to SQA and under this (and other terms) can provide a general solution to a problem created by e-assessment and unitisation. In SQA (2007C) the idea was mooted of providing an e-portfolio for candidates to accumulate the results (marks or grades) obtained through e-assessment and other processes. It is assumed from the context that the e-portfolio would be created and maintained by SQA and that candidates would access their details via a web link. Under different names, similar systems are already provided by a number of awarding bodies and ExoD (2005C) have suggested how they could be adapted to make provisions for candidates to investigate what grades they might already be entitled to, and what levels of performance they might require in outstanding tests/units, if they wish to be awarded particular grades.

5.1.3 On-screen assessment

The potential of on-screen testing has yet to be fully realised. One issue is the limited range of question types that are currently being used. Multiple choice and

other objective question types predominate. While notable advances have been made in extending the range of e-question types in some subject areas, on-screen assessment has yet to find its niche in the overall scheme of UK school examinations. Take-up has been greatest in lower level vocational qualifications, higher level UK professional qualifications, and in universities. In many vocational and professional qualifications, on-screen testing is the dominant technique and only gives way to more traditional forms of assessment for the final professional components.

Kingdon (2004) and ExoD (2005C) have argued that the full potential of on-screen assessment is unlikely to be realised until there are increases in the numbers and quality of e-learning courses and large numbers of assessments are available on-demand. In the meantime, its use is likely to be limited to:

- test-when-ready/on-demand assessment of individuals or small groups of school candidates
- resits by individuals or small groups of school candidates who have failed to achieve expected results in paper-based assessments
- monitoring standards or moderating teachers' assessments by assessing sample students

The significance of on-screen assessment is its ability to be delivered on-demand and provide immediate results, detailed feedback, and early resits. Underpinning these developments there will need to be calibrated question, section and/or paper banks, with associated pass marks, from which tests can be supplied to centres on-demand. (Calibration means that questions or sections which have been answered by different groups of test-takers are made comparable by using one scale. This is necessary to select items from a bank and to report comparable results.)

The challenge for future grading systems is the need to provide supporting:

- mechanisms for producing pre-set pass marks
- websites or e-portfolios for candidates to bank e-credits gained and make informed decisions about when to cash them in to be converted to grades

5.2 Unitisation of courses and assessment

Some of the issues associated with the aggregation of results for modules (or units) that belong to a single qualification have been considered in the section on component grading (Section 4.9.1). This section considers some of the additional issues associated with awards which are based on units selected from a large number of options, possibly drawn from different qualification streams.

Unitisation creates grading issues beyond the numbers of results to be combined and the need for candidates to be able to make informed choices about when to apply for grading:

- The heterogeneity of the units and whether they can be validly combined together. Ideally, this is an issue that should be addressed at the time candidates enter for unitised schemes, however, problems may only appear once units are completed. The profile approach to grading appears to offer a more valid approach to aggregating heterogeneous units than simple addition of marks.
- Contemporary initiatives such as Assessment for Learning place greater emphasis on internally (teacher) assessed work, which tend to have very different distributions of marks or grades from externally assessed components. Attempts to make internally-assessed components fit the characteristics of externally assessed ones have not always been successful. Indeed, in extreme cases that may have distorted the characteristics — and potentially the value — of one or both components. Therefore, the grading of qualifications, based wholly or partly on internal assessments, may necessitate use of one of the approaches suggested in Section 4.9.2.
- For qualifications that allow credit accumulation — perhaps as part of vocational continuous professional development (CPD) schemes — candidates' opportunities to study particular units may vary, increasing the heterogeneity issue reported above.
- If the CPD scheme were also to have units that were time-limited then there would need to be mechanisms to indicate to candidates that their credits for some units might expire shortly.

Increasingly, complete qualifications to practise a craft or profession may become time-limited. Existing examples include first-aid certificates and registration qualifications for electricians, plumbers, gas fitters, street workers, etc. At the higher end of the qualifications scale, professional licences to practise are becoming conditional on the collection of specified CPD units in stated timescales. So far, time-limited qualifications tend to be limited to those crafts and professions where failure to remain up to date could result in injury or death to the practitioner or his/her clients. Pressures from the EU professional indemnity insurers plus health and safety advisors all mean that time-limited qualifications can be expected to increase.

A factor associated with many unitised qualifications and credit-based CPD schemes in the vocational and professional areas is the use of a hierarchy of test situations in which candidates may complete assessments. To offer the greatest opportunities for candidates to complete them, the advantages of on-screen and on-demand assessment are obvious. ExoD (2005C) offered a list of test situations that started with a candidate's home workstation, through workplace situations, to secure assessments taken in commercial test centres.

5.3 On-demand assessment

On-demand assessment is not restricted to on-screen delivered assessments, nor are the types of questions restricted to those that require choice of options or one word/short answers. Using modern ICT technology there are several ways that tests and examinations can be compiled and delivered to centres and candidates' responses captured for return to the awarding body.

The challenges for awarding bodies and regulators arise when immediate results are required. These imply a paradigm shift from post-marking and post-marker standardisation grading to *a priori* grading. ExoD (2005C) have suggested how such grades might be set and standards maintained.

The greatest significance of on-demand assessment, using e-assessment procedures and e-learning systems is not for awarding bodies and regulators but institutions. Kingdon (2004) and ExoD (2005C) have speculated on how fully personalised learning might be organised and some of the issues that it will create.

5.4 Implications for SQA

All of the developments considered in this section have implications for grading. In addition to the issues of the number and heterogeneity of components to be aggregated into single grades, considered in Section 4 this section raises issues about the timing of awards. While some of the developments require *a priori* grading to fulfil their potential roles in future qualifications, others will involve the aggregation of component results gained over longer periods of time — possibly some years. Future grading systems may have to accommodate components that have been re-sat one or more times, and components that are time-limited.

Some UK candidates already make decisions about when to have banked component results graded, for instance whether to accept the result for a component of the first half of an A level or to decline it and resit the assessment in the final half. SQA is developing electronic reporting services (e-mailing results to candidates in 2008). Candidates in all SQA qualifications have always been kept informed about the Units and credits achieved so far on paper, and at set times.

6 Conclusions and implications

Grading (some organisations and writers use the term ‘awarding’) is the process of converting marks for components into component and subject grades. In the traditional examination process it is one step in the qualifications delivery process.

In common with other UK awarding bodies, the principles underpinning SQA awards are founded in case law and informed by developments in psychometrics and IT. SQA is able to draw on over 140 years of examining experience in Scotland, plus the many ideas and documents shared with the other UK qualifications regulators and awarding bodies. Given its unique national status, SQA and its forerunners have been able to implement quality systems that have been impossible for other UK awarding bodies. Despite this immense administrative and intellectual legacy, however, grading remains a judgemental process.

Efficient grading is predicated on the design of the assessment regime, the validity of the assessment components, plus the reliability and completeness of marking and marker standardisation. In their turn, the efficiency of grading decisions can be evaluated in terms of the:

- comparability of the grades awarded — over time, across all of the routes that lead to the same grade, across recognised points of equivalence with other qualifications, and across subjects at the same level
- validity of the aggregation process — the greater the number and heterogeneity of the components to be combined, the more difficult it becomes to achieve adequate validity
- discrimination of the final/subject marks — specifically the number of marks between adjacent grade boundaries
- robustness of the grades — their overall stability when individual grade boundaries are changed by single marks

The grading of Scottish school leaving qualifications is subject to several challenges not experienced to the same extent by other UK systems and the majority of the qualifications systems analysed for this report (see Annex 1). The current Scottish system is unique in having six streams of qualifications with recognised points of overlap. First, there are more grades to award. The resultant number of grades is larger than almost all of the other grading systems explored. Second, one of the aims of grading should be achievement of adequate comparability across recognised points of equivalence between the different qualification streams. Third, UCAS recognises some of the Scottish grades as equivalent to those of other UK systems. Traditionally, SQA (and its forerunner) have not participated in the comparability studies of school leaving qualifications organised by other UK awarding bodies as other than observers. The validity of the UCAS assumptions has been tested in a comparison study in 2008.

Some users of Scottish qualifications, in common with those of other UK qualifications systems, continue to assume that reported grades can be manipulated in the same ways as other quantities. The UCAS tariffs and value

added systems assume that the grades are at the interval level of measurement, and equal steps in grades represent equal steps in attainment. Some users even assume that a Grade 4 indicates attainment which is twice as good as that at Grade 2, and that lower grades can be added together to equal a higher grade. UK awarding bodies — including SQA — are more modest, and claim only to be working at the ordinal level of measurement.

Grading in Scotland and the UK as a whole will continue to be an important issue. Ongoing societal changes continue to create new needs and assessment opportunities. Section 5 discussed a number of new developments — in qualification delivery and responses to perceived needs — that will all create their own grading demands.

Annex 1: School leaving qualifications in other countries

Introduction

In order to understand how Scottish school leaving qualifications contrast with those of other countries, a literature search was conducted using education libraries and the internet. As insufficient reports were available on vocational systems, the analysis concentrated on school leaving qualifications and access to higher education. Similarly, examinations and tests taken earlier in students' school careers, such as the French 'Cadet' examination and England's National Curriculum Tests, were excluded.

The comparators included:

- whether the system reported applied to all schools — state, regional and other variations were considered separately. In total, the qualifications systems of 62 countries and 72 national plus regional award systems were identified
- the stated bases of awards — quality of work, percentiles or marks
- where reported, the method used to moderate the awards
- whether grades or marks are used and their range
- the number of pass grades or the usual minimum pass mark
- where reported, the proportions of candidates that obtain the highest and cumulative pass grades/marks
- the way that subject grades are combined for reporting and selection purposes
- How the main academic and professional selection processes take place — using the qualification alone, in combination with university entrance examination and/or the early years of HE

A number of other factors were explored but discounted when it was found that they did not differentiate (or differentiated so much that qualifying examples were limited).

A number of recurrent patterns in school leaving examinations were identified:

1. The most common age for students to take school leaving examinations is 18. The UK is alone in organising school leaving examinations for 16 year olds[18].
2. Fifty-four of the systems considered use quality as the prime factor for placing marks and grades. Unfortunately, there is insufficient information to be able to report how initial decisions on marks or grades are confirmed.
3. Fourteen of the countries analysed — including France and Francophone countries across the world — use a 20-point mark-based system, with a pass mark of 10[19]. The definitions of individual mark points seem to be implicit

[18] Kingdon & Stobart (1988) and Kingdon (1991) have traced the reasons why the UK 16+ examinations remained so important in the UK.
[19] Portugal uses 9.5 marks.

and in the case of France and former French colonies, probably traditional[20]. Full awards require passes in prescribed groups of subjects and (usually) a minimum grade point average (GPA). Passing the qualification granted access to university but selections tended to take place in the early years of HE. In the Francophone countries, holding the qualification also grants social status and access to key professions.

4. A further 12 countries — mostly European but others with strong links to Europe — use a 10-point mark-based system. Pass marks vary from 4 marks (Latvia, Lithuania and Finland) to 7 marks (Argentina and Brazil). The modal range being 5 or 5.5 marks. These differences apart, this approach to awarding is very similar to the Francophone/Napoleonic tradition outlined above, so is assumed to be derivative.
5. The Italian system bridges the gaps between these two traditions by using a 10-point system with half mark divisions.
6. Further variations on the use of marks included:
    a. Mexico and Costa Rica use percentage marks with pass levels being set at 60% and 70% respectively
    b. Luxembourg and others use a 60-mark system, again based on quality of work, with 30 marks set as the pass level
    c. Russia is moving from a 5 grade system to a 12 mark system with a pass level of 2 marks
    d. Denmark also uses a 12 mark system, with a pass level of 6 but with 3 negative marks to differentiate further levels of failure.
7. A small majority of the countries investigated use grades rather than marks, although the distinction seems to be a fine one in some cases. Both alpha characters — usually starting with A (or the equivalent in the local character set) as the highest — and ascending or descending numeric systems are used. The typical number of pass grades is 4 or 5, with one or more fail grades.

Two traditions are identifiable among the countries that use grades:

8. Most schools in the USA, Canada, China, Australia and many of their near neighbours use the ABCDF system where F is a fail grade. Some US and Canadian states have expanded the discrimination of the system by introducing additional grades eg A*, A- etc, while Japan and Peru use just ABCF. The basis of awards in all of these countries is norm referencing therefore, the first selective device is the grade point average (GPA). However, entry to the more prestigious US and Canadian universities is dependent upon SATs scores and further rigorous selection takes place in the early years of HE. Australia aside, the other countries also had additional systems of selection for HE.
9. Twelve countries, most in Europe but some elsewhere, use 5- or 6-point numeric grade systems based on quality of work. Germany and Austria have additional selection systems for some universities and faculties but the strongest selection takes place within the early years of HE.
10. There are a few national variations on the use of alpha and numeric grades based on quality of work:

[20] Colleagues who work in the French educational system have reported that the 20-mark system, with a pass mark of 10 marks, was instituted by Napoleon. Therefore, it is appropriate to talk about a ‘Napoleonic tradition’ in school leaving examinations.

   a. New Zealand, Romania and Sweden all use abbreviations for 4 levels of performance (3 of pass and fail)
   b. Nepal uses Division 1* to Division 111 (4 grades of pass) and F.

11. The countries in the UK are differentiated from the others analysed in several ways:
   a. In very few other countries did the main selective function for HE and the professions rest with the school leaving qualifications. By far the most common patterns are for HE candidates to take university entrance examinations and/or for the most rigorous selection to take place in the early years of HE. Other examples, of the system of 'sponsored mobility'[21] used in the UK, were not found except among a small number of Commonwealth countries.
   b. The system of aggregating results for different subjects and qualifications used in the UCAS system again is unique to the UK. Almost all other countries that sought to aggregate results use grade point averages or weighted grade point averages if different types of qualifications are to be combined.
   c. All the non-UK countries analysed have very much simpler school leaving qualification systems.
12. The systems of international qualifications that were considered — principally the International and European Baccalaureates (IB and EB) — showed characteristics of both UK and other systems but are generally distinct from all other school leaving qualifications analysed.

Trends

In so far as it was possible to identify trends, they are confined to the countries that use grades. While a small number of countries have increased the number of grades in recent years, an almost equal number have decreased them.

Implications for Scotland

As far as could be determined from the information analysed:

- the Scottish system of school leaving qualifications is the most complex of those analysed
- the total number of grades that can be derived from Scottish school leaving qualifications is among the highest of the countries analysed
- Scottish school leaving qualification, in common with other UK countries, a small number of Commonwealth countries and qualifications such as the IB and EB, have a selection function that the other school leaving systems analysed do not
- the Scottish system of school leaving qualifications (in common with other UK countries) appeared to have been the subject of far more changes in recent decades that most other systems
- Scottish and other UK school leaving qualifications are graded on the basis of quality (in common with many other countries). However, from the sources used, it was not possible to identify whether written or 'cultural' definitions of

[21] See Turner (1960).

the quality associated with individual grades (or mark points) are used in particular countries. Empirical evidence supports the latter. If the latter definition is assumed, a loose association can be drawn between countries that use quality as the basis of their awards of grades and those that have enjoyed stable qualification systems

Annex 2: Some technical aspects of grading

Introduction

In Section 4.8 it is stated that grades for qualifications should be valid, reliable, practicable, equitable and fair; and also complete, discriminating and robust. Later, in Section 4.9.5, it is suggested that components and subjects that are to be aggregated to form parts of larger awards should be 'addable'. This annex discusses the psychometric bases for these statements.

Unless otherwise stated, the remarks in this annex apply to both:

- the grading of components and their aggregation of form subject grades, and
- the aggregation of subject results to form group awards

Terminology

Levels of measurement

It is important to keep in mind what it is that is being measured or described, and what type of measurement scale is being used, because this determines how the measurements can be used. Statisticians distinguish the following main types of measurement scales and their restrictions:

Nominal scale

At this level descriptive categories are used, such as white/mixed/Asian/ black/other ethnic background. It is impossible to order these categories in more or less of the same, or better/worse. It is also impossible to summarize them into an average ethnic background. They can only be counted.

Ordinal scale

An ordinal scale is used to describe and measure attributes using categories which can be ordered, for instance blond/brown/black, or sufficient/good/excellent. Now it is possible to not just compare individuals, but to measure them roughly by stating that one has more or less of the underlying attribute (pigment, or skill). It is still not possible to compute an average, because it is not known how much darker or better one is than the other. Note that renaming or coding the categories into numbers (Grade 3 to 1 for example) doesn't change this, because the grades are still based on categories with a flexible width.

Interval scale

An interval scale is used when the difference between the ordered categories is known, and is the same, as is the case when bands with an even width are used. Now a band 4, 5, and 6 can be averaged to a mean of band 5, because band 6 is one bandwidth worse than band 5, and band 4 is one bandwidth better. This

assumes that the same increase in skill is required to move from band 5 to 4 as from band 6 to 5. When statisticians think this assumption is not correct, they will try to convert marks into an ability scale which has a true interval level.

Ratio scale

Ratio scales are used when the scale has a zero point, as for example household scales. Different readings on the same scale can be related to each other by stating that, for example, one kilo is two pounds. Although marking scales go from zero to a maximum, their zero score does not measure zero skill or knowledge, but the level of skill or knowledge where the assessment starts registering achievement. A zero score on an Advanced Higher does not necessarily mean that the candidate doesn't know anything at all in the subject. The lack of a zero point means that we cannot say that a candidate with 80 marks is twice as knowledgeable as one with 40 marks.

Measurement error

Every assessment instrument has a certain amount of error in its measurement. Partly this is inherent. Just as human weight scales are not sensitive enough to measure letters, let alone atoms accurately, assessments can't measure someone's true ability completely accurately. Partly this has something to do with the assessment itself. It may be too short to measure small differences in knowledge or skill and it may not have a lot of discrimination. Or some questions assumed background knowledge which was more common among boys than girls. Partly, the inaccuracy results from external conditions such as noise, and marking. And partly it results from the fact that we, the object of the measurement, cannot demonstrate (or even have) exactly the same degree of ability at all times, in the same way as our weight fluctuates during the day.

The amount of measurement error can be estimated using the standard deviation and the reliability of the assessment. This estimate is called the standard error of measurement. With the standard error of measurement we can estimate how close an actual score is likely to be to true scores. So, for an assessment with a standard error of measurement of 5 out of 100, we can state that the mark perfectly indicating their true ability would be within a range from 10 marks (twice the error) above to 10 marks below their actual mark for 68% of all candidates. This means that most of the time a mark of 70 on this assessment points to a true mark somewhere between 60 and 80. Measurement errors of this magnitude are usual, and this is one of the main reasons not to report 'raw' marks, but bands, or grades[22].

Reliability

An examination can be said to be reliable if candidates of equal ability achieve the same marks and grades. Typically, school-leaving exams focus on reliability of marking. The extent to which the number and selection of exam questions

[22] The standard errors of measurement in the assessments used in SQA science exams in 2008 (Standard Grade elements, or multiple choice sections of other qualifications) varied from 4% to 10% of the maximum score. The errors were larger as the assessments were shorter.

produce an accurate picture of a candidate's ability is also part of reliability. The ways in which marking has been specified, managed and standardised all impact on the confidence users have in final marks and grades. When the reliability of an assessment as a whole is low, this means that the mark a candidate has achieved might as well be several marks higher or lower, for instance if the assessment had been marked by another marker, consisted of other questions, or had been taken at another time of the day.

Completeness

Decisions based on complete and standardised marks are more likely to be valid than decisions based on partial and/or unstandardised marks. Sometimes, however, provisional grading has to take place using incomplete results. In such cases, the provisional grades need to be confirmed when the results are complete. In other situations, for instance in surveys, total results may be computed by estimating results for some components from the results for the other components. Obviously, this affects their accuracy. For this reason, these estimated individual results may not be reported, but may only be used to report group results.

Discrimination

The most useful measure of the discrimination of a component (or subject) is its standard deviation. Discrimination is highest when candidates' scores are spread out evenly over the whole marking scale. When many candidates have very similar marks, for instance when half of the candidates have a mark between 65 and 70 out of 100, the assessment does not distinguish very clearly between these candidates, especially not when the reliability (accurateness) of the marks is low as well. The standard deviation is used to measure this spread. For reasons that are discussed below, when component marks are summed to form subject totals, the standard deviation of the subject totals is usually less than the typical standard deviation of the component marks. The same phenomenon — regression — is observed when subject results are summed to form group totals (see Regression).

Robustness

The robustness of a set of component (or subject) grades is the degree of stability they exhibit when individual grade boundaries are changed by a single mark. The discrimination and robustness of sets of component (or subject) grades are related. If the discrimination is high — ie candidates are well spaced around the boundary — single mark changes have less impact on the grades as a whole and, therefore, have greater robustness than when candidates are bunched round a boundary.

Validity

This is an umbrella term for a number of related concepts. First, any qualification should have **content validity**, which is a measure of the degree to which the examination papers and mark schemes can be said to have sampled the subject domain, as represented in the specifications or syllabus.

An important part of the content validity of a qualification is the balance between the means and standard deviations (SDs) of the component marks. If a component mean is too high (or too low) it can have a distorting effect on the subject content and skills used to discriminate between candidates at the pass/fail boundary. For instance, candidates achieve much better on average on an internally assessed practical component of an exam than on the externally assessed theory, and both marks are simply added up. Passes are then likely to reflect achievement of the practical aspects of the course more than the theoretical aspects (unless the pass boundary takes this difference into account). If a component SD is relatively high (or low) the content and skills as measured by this component contribute more to the discrimination of the total mark.

Second, for components to be part of the same subject they need to have mutual **predictive validity**. It should be possible to estimate candidates' performance in one component from their performance in the other components that make up the subject. (A similar argument applies if subjects are to be combined into a group award.) Discussion of whether components have adequate predictive validity is usually conducted in terms of their correlations with each other. Correlations typically indicate the extent to which candidates achieve similar results on both components. They use the distance of each component score to the mean score for the component, or they look at the rank order of each candidate's grade on each component[23]. For the purposes of this paper it is helpful to consider the correlation between two components as being a measure of the extent to which they measure the same subject content and skills. The following points indicate how particular correlation values are interpreted:

- **A correlation of +1.0** between two components would indicate that they are measuring exactly the same content and skills. The implication of this is that only one of the two is required to assess the candidates.
- **Correlations approaching +1.0**. Correlations above +0.85 are common between components of mathematical subjects, are often observed for physical science and engineering subjects, but are rare in language, arts and humanities subjects. A correlation level this high indicates that the there is considerable overlap in what the two components are assessing. Unless they can both be justified on the grounds of syllabus coverage, correlations this high may indicate over-assessment of the subject content. The implication is that one of the two components might be reduced or dispensed with. (If this was practised, and teachers or candidate could predict which component would be assessed, it would obviously run the risk of reducing the taught curriculum to one component, which in turn might considerably lower the correlation.)
  If such correlations were to be observed between pairs of subjects, it would question whether they are indeed separate subjects.
- **Correlations in the range +0.5 to +0.7** are typical for many types of subjects but are low for mathematical ones.
- **Positive correlations below +0.35** between two components may indicate that they are linked by no more than a general ability factor. Correlations between randomly selected pairs of subjects typically fall in this range. Therefore,

[23] There are many formulas in use. The Pearson product-moment correlation is most often used for scores, the Spearman correlation for rank order.

correlations of this order between pairs of components question the integrity of a subject and/or the examination being used to assess it

- **Correlations close to zero** between two components indicate that they are effectively independent of each other and there is no overlap at all in the content and skills being measured. If near zero correlations were to be observed repeatedly between the same components they would question whether the components belong to the same subject.
- **Negative correlations** between components indicate that they are providing contradictory information about the candidates. As correlations approach –1.0 they indicate that more and more candidates who have performed well in one component have performed poorly in the other.

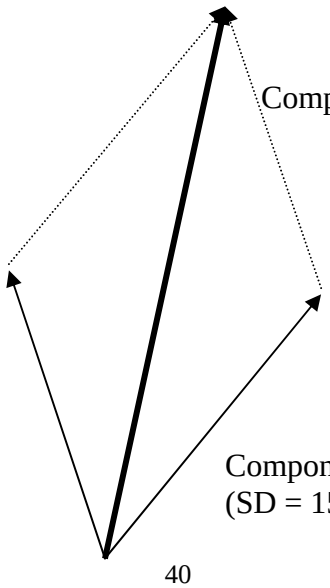
Similar arguments apply to subjects that might be combined to form a group award.

Regression

The correlations between components provide a convenient measure of the integrity of a subject. Very high correlations can reveal instances of over-assessment, even redundant components. Correlations approaching zero imply that the components have insufficient in common to justify inclusion in the same subject and that the addition of the marks may be invalid.

For the marks for different components to be 'addable', the correlations between pairs of components need to be positive and significant. However, even here there are problems. Whenever less than perfectly correlated sets of marks are combined the results bunch towards the middle of the distribution — a phenomenon called 'regression'. The cause is that, while the results for individual candidates can be added in the normal way, combining sets of marks for groups of candidates involves a process called a vector addition. The following example illustrates what happens:

Figure 1: Diagram to illustrate how the standard deviation (SD) and discrimination of an examination are reduced as two sets (columns) of marks with a correlation between the components of +0.6 are combined.



Bold line shows the result of adding Component 1 marks to Component 2 marks
SD of the total = 13.4%

Component 1: (SD = 15%)

Component 2:
(SD = 15%)

Comments

1 A correlation between two components of +0.6 is good for most types of subjects. Even so, the SD and thus the discrimination of the examination, have been compressed by combining the two sets (columns) of marks

2 While the degree of compression is not large in this case, it would become larger (and more significant) if:
   a the correlation between the components were to decrease further
   b the marks of additional components were to be added to the two shown above

The following table illustrates how regression effects increase as the number of components increase and the correlations between pairs of components decline.

Tables 1A: Regression effects associated with increasing numbers of components and different levels of correlation between pairs of components.

Note: The figures in Table 1A are derived from the combination of two observed distributions of marks, both of which are normally distributed, have means of 50% and standard deviations (SDs) of 15%. If the two distributions were perfectly correlated and weighted 50:50 the resultant mean and SD would also be 50% and 15%. The regression effects and subsequent loss of discrimination **observed** when pairs of mark distributions, with progressively lower correlations, are expressed as percentage reductions in the SD.

Approximate reductions in the discrimination of an examination associated with the aggregation of two sets of components marks				
	Correlation			
	+0.9	+0.6	+0.3	0.0
Approx. reduction associated with aggregation of 2 components	2.5%	9%	15%	30%

It is very unusual to find three or more sets of component marks that all correlate to exactly the same extent, so Table 1B has been produced using mean correlation values. Again, equal weightings have been applied.

Table 1B

Approximate reductions in the discrimination of an examination associated with the aggregation of more than two sets of components marks		
No. of components	Average correlation between components	Projected reduction in discrimination of examination plus comments/implications
3-6	0.9	<5%. However, very few subjects have such high levels on inter-component correlations
4	0.6	Approx 15%
6	0.6	Approx 20%

6	0.3	Approx 30%. The reduction in discrimination is becoming significant
10	Near zero	60+%. Alternatives to direct aggregation of components should be investigated

A contemporary example of six components being aggregated into a single subject grade is the English A-level examination. Even with six components, the discrimination of the examination may be prejudiced if the inter-component correlations were to decline as far as 0.3 — the level usually associated with a general ability factor only.

If the trends in the regression effects illustrated in Tables 1A and 1B continue, it is easy to imagine how, if the number of weakly correlated components were to increase too far, two situations would become possible:

- The regression effects would increase to a level when the standard deviation of the overall examination declined so far that the discrimination of the examination was seriously impaired, and/or
- The introduction of further components would cease to have any material impact on the examination as a whole

While these scenarios are unlikely to arise in practice at the component level — the cost of designing and assessing additional components alone should preclude them — they represent a real danger where subject results are to be combined into group awards.

Given that regression effects are a statistical inevitability, how can they be mitigated and the integrity of published grades preserved?

At the subject level the only way to counter the implications of regression is to ensure that individual components are as psychometrically efficient as possible in assessing their subject, so that the number needed to assess the subject properly is kept to a practicable level. Steps to achieving this include:

- agreement of the aspects of a subject that are important and should be examined
- selection of components so that, collectively, they assess the important aspects of the subject in a balanced way
- design of the examinations and mark schemes for individual components so that they discriminate well and lead to marking with high standard deviations. When such component results are aggregated, the subject as a whole will also achieve good discrimination and this will be evidenced by a high standard deviation
- good levels of correlation between pairs of components — definitions of ‘good’ will, of course, vary with the type of subject being examined

- avoidance of peaks and gaps in the mark distributions for individual components which might inhibit the placement of grade boundaries and lower the robustness of the examination. Because in scaling some original marks can be clustered together into one new mark which may produce significant spikes and gaps, scaling (and standardising) of component and subject marks should be kept to a minimum.

The number of grades to be awarded is also an issue. Smaller numbers of grades lead to greater robustness at the component grade level and, if direct subject grading approach is applied, at the subject level too. However, if subjects are graded via the aggregation of component grades, using a greater number of grades at the component level leads to subject grades with higher discrimination and robustness.

References

Benson-Pope D (Chair) **(2005),** *Report of the scholarship reference group*, Office of the Associate Minister of Education, Wellington, NZ, March

Bruce G (1969), *Secondary school examinations: facts and commentary*, Pergamon Press, Oxford & London

Dunning J (Chair) **(1977),** *Assessment for all: report of the committee to review assessment in third and fourth years of secondary education in Scotland*, Scottish Education Department, HMSO, Edinburgh, ISBN 0 11 491505 9

Entity (2008), *personal communication*

ExoD (2005A), *A response to the Tomlinson report*, Exam on Demand Assessment Advisory Committee, Occasional Paper No. 1, Exam on Demand Ltd, UK

ExoD (2005B), *Tomlinson Revisited*, Exam on Demand Assessment Advisory Committee, Supplement to Occasional Paper No. 1, Exam on Demand Ltd, UK

ExoD (2005C), *The development of e-assessment: 2004 to 2014*, Exam on Demand Assessment Advisory Committee, Occasional Paper No. 2, Exam on Demand Ltd, UK

Hunter SL (1963), *Scottish education: changes in the examination structure in secondary schools*, International Review of Education, Vol 9, No 3, September, pp 310-324

French S, Willmott AS & Slater JB (1990), *Decision Analytic Aids to Examining (DAATA)*, Research and Monitoring Unit, School Examinations and Assessment Council (SEAC), London

Guildford JP & Fruchter B (1978), *Fundamental statistics in psychology and education*, 6th edition, McGraw-Hill, ISBN 0-07-Y85248-0

Holmes E (Ed), **(2007)**, *British qualifications*, 37th edition, Kogan Page, London & Philadelphia, ISBN-10 0 7494 4803 2

Johnson A (2007), *Update and assessment of the grading of the specialised Diploma*, letter from the Secretary of State for Education and Skills to English awarding bodies

Kingdon M (1991), *The Reform of Advanced Level*, Hodder and Stoughton, London

Kingdon M (2001), *Finite human resources: the approaching assessment crisis in England*, paper presented to the 27th annual conference of the International Association for Educational Assessment, Rio de Janeiro, Brazil, May

Kingdon M (2004), *The e-assessment agenda*, paper presented to the 30th annual conference of the International Association for Educational Assessment, Philadelphia, June

Kingdon M & Stobart G (1988), *GCSE Examined*, Falmer Press

McAlpine M & Ware M (2003), *Introducing computer assisted assessment in Scotland: laying the foundations for an integrated approach*, paper presented to the 29th annual conference of the International Association for Educational Assessment, Manchester, UK, October

Newton P (2007A), *Contextualising the comparability of examination standards*, Chapter 1 in Newton P, Baird J-A, Goldstein H, Patrick H & Tymms P (2007), *Techniques for monitoring the comparability of examination standards*, QCA, London ISBN 1-85838-97-1 (pp 9-42)

Newton P (2007B), *Comparability monitoring: progress report*, part of Chapter 10 in Newton P, Baird J-A, Goldstein H, Patrick H & Tymms P (2007), *Techniques for monitoring the comparability of examination standards*, QCA, London, (pp 452-486), ISBN 1-85838-97-1

O'Connor JJ & Robertson EF (2000), *The setting up of the Scottish School Leaving Certificate*, article based on Chapter 3 in Yousuf M (1990), St Andrew's University doctoral theses submitted January 1990
URL: **http://www-history.mcs.st-and.ac.yk/Education/scottishleaving.html**

Petch JA (1953), *Fifty years of examining: The Joint Matriculation Board 1903-1953*, George Harrap, London

Philip HL (1992), *The Higher tradition*, Scottish Examinations Board, D&J Croal, Haddington, ISBN 0 901256 84 9 pp 127

QCA (2000), *Arrangements for the statutory regulation of qualifications: in England, Wales and Northern Ireland*, Qualifications and Curriculum Authority on behalf of QCA, ACCAC and CEA

QCA (2004A), *The statutory regulation of qualifications: in England, Wales and Northern Ireland*, Qualifications and Curriculum Authority on behalf of QCA, ACCAC and CEA

QCA (2004B), *On-screen delivery of qualifications: the basic and key skills experience, Qualifications and Curriculum Authority*, March

QCA (2007A), *Codes of practice: GCSE, GCE, GNVQ and AEA*, Qualifications and Curriculum Authority on behalf of QCA, ACCAC and CEA

QCA (2007B), *Assessment and Grading of the Diploma*, QCA position paper dated 14 March 2007

QCA/ LSC (2004), *Principles for as credit framework for England*, Qualifications and Curriculum Authority

Robinson C (2007), Chapter 3 in Newton P, Baird J-A, Goldstein H, Patrick H & Tymms P (2007), *Techniques for monitoring the comparability of examination standards*, QCA, London, (pp 97-123) ISBN 1-85838-97-1

Scotland, J (1969), *The history of Scottish education, Volume 2, from 1872 to the present day*, University of London Press, ISBN 340 09548 2

The Scottish Office (1996*)*, *Scottish Certificate of Education: Standard Grade*, Factsheet 5, originated April 1996 and updated thereafter

SCQF (2007), *SCQF Handbook*, website www.scqf.org.uk

SQA (undated), *Higher Nationals — a short history*, Note ref needed in paper. URL**: http://www.sqa.org.uk/saq/**

SQA (2007A), *Conditions and arrangements for national qualifications 2007/2008*, SQA publication code BA0828

SQA (2007B), *Review of qualifications at SCQF levels 4 and 5: Assessment of qualifications at SCQF levels 4 and 5*, Assessment and qualifications task group, working group 1, SQA, May

SQA (2007C), *SQA's vision and strategy for e-assessment*, SQA, June

SQA (2008A), *Review of qualifications at SCQF levels 4 and 5*, Assessment and qualifications task group, version 3.1, SQA, January

SQA (2008B), *Induction training: national course grade boundary setting*, SQA PowerPoint examiner training presentation

SQA (2008C), *Induction training: standard grade: grade boundary setting*, SQA PowerPoint examiner training presentation

SQA (2008D), *Guide to assessment* SQA publication code AA4147, **http://www.sqa.org.uk/sqa/files_ccc/GuideToAssessment.pdf**

Turner RH (1960), *Sponsored and contest mobility in the school system*, American Sociological Review, vol 58, pp 247-56

Tomlinson M (Chair) ***(2004)***, *Curriculum and qualifications reform: final report of the working party on 14-19 reform*, DfES Publications (ref DfE-0976-2004), Nottingham, and **www.14-19reform.gov.uk**

UCAS (2008), *Tariff tables: find out how many points your qualification are awarded*, **http://ucas.com/students/ucas-tariff/tariftables/**

Wood R (1991), *Assessment and testing: a survey of research*, Cambridge University Press

University of London (1836-1846), *Minutes of the University Senate*, Volumes 1 & 2