CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Reporting scale scores at GCSE and A level

Research Report

Tom Bramley
Carmen Vidal Rodeiro
Frances Wilson
March 2024

Author contact details:

Tom Bramley, Carmen Vidal Rodeiro & Frances Wilson
Assessment Research and Development
Research Division
Shaftesbury Road
Cambridge
CB2 8EA
UK

tom.bramley@cambridge.org
carmen.vidalrodeiro@cambridge.org
https://www.cambridge.org/

As a department of the university, Cambridge University Press & Assessment is respected and trusted worldwide, managing three world-class examination boards, and maintaining the highest standards in educational assessment and learning. We are a not-for-profit organisation.

Cambridge University Press & Assessment is committed to making our documents accessible in accordance with the WCAG 2.1 Standard. We're always looking to improve the accessibility of our documents. If you find any problems or you think we're not meeting accessibility requirements, contact our team: Research Division
If you need this document in a different format contact us telling us your name, email address and requirements and we will respond within 15 working days.

How to cite this publication:

Bramley, T., Vidal Rodeiro, C.L., & Wilson, F. (2024). *Reporting scale scores at GCSE and A level.* Cambridge University Press & Assessment.

Note
This was originally an internal report from June 2014. It has been lightly edited before publication in March 2024.

Contents

1. Introduction

The recent plans to reform GCSE qualifications have raised the question of what the best format is for reporting examination results. Traditionally in England, exam results in GCSEs (and before them O levels) and A levels have been reported as letter grades, with A (or A*) as the top grade, then B, C etc. The reforms gave the opportunity to revisit the arguments for different formats of reporting, and Cambridge Assessment contributed to the early debates with a short paper recommending the use of longer numerical scales (Bramley, 2013).
The purpose of this study was to explore in more depth the arguments for and against different reporting scales. We took as a given that in all cases there is an underlying 'raw score scale' obtained by adding up (possibly after weighting) the raw scores on the components of the examination, as happens with 'linear' assessment. The process of scaling is essentially finding a mapping from this raw scale into the reporting scale. Such a mapping is sometimes referred to as a 'scaling function' taking raw scores as its input and producing the scale scores as its output. It is irrelevant whether the labels for the reporting scale categories are letters or numbers – both are taken as implying an ordinal relationship, with alphabetical or numerical order implying greater (or less, depending on the scale) achievement in the examination. 'Grading' of linear GCSEs and A levels, and 'level setting' on National Curriculum tests are thus both examples of scaling.

Grading of modular or 'unitised' assessments at A level and GCSE is more complex. The raw scores are first scaled at unit level using the UMS scaling procedure, and then aggregated. A second scaling function then maps the aggregate UMS scores into letter grades, but this scaling function is fixed and known in advance. Some features of UMS scaling are explored in section 5 of this report.

Figure 1.1 below is an illustration of the scaling function for a linear Religious Studies (RS) GCSE taken in 2009. The maximum possible raw mark was 168, meaning that there were 169 possible raw scores (0-168). These were mapped into grades A* to U as shown in the figure.

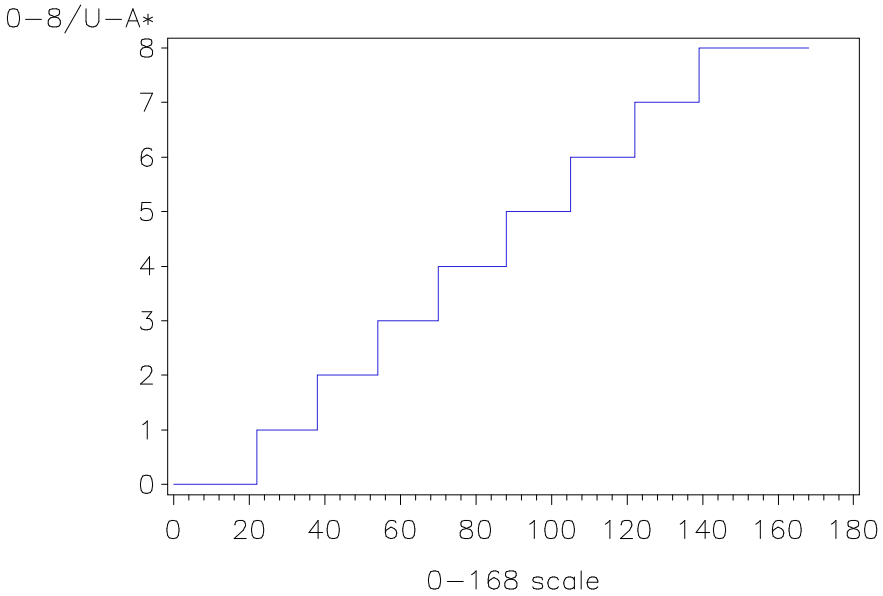

Figure 1.1: Scaling function for GCSE RS in 2009.

The same scaling function could be expressed as follows:

y=f(x), where y is the value on the reporting scale, x is the value on the raw scale, and:

$0 \leq x \leq 168$
$y=U$ for $x < 22$
$y=G$ for $22 \leq x < 38$
$y=F$ for $38 \leq x < 54$
$y=E$ for $54 \leq x < 70$
$y=D$ for $70 \leq x < 88$
$y=C$ for $88 \leq x < 105$
$y=B$ for $105 \leq x < 122$
$y=A$ for $122 \leq x < 139$
$y=A^*$ for $139 \leq x$.

Such an unwieldy formulation is not usually used in practice, of course – the lowest possible raw scores mapping to each scale score are known as 'grade boundaries' (or 'level thresholds', 'cut-scores' etc., depending on context) and the scaling function is most commonly represented as a simple tabulation of these boundaries, e.g.:

Grade	U	G	F	E	D	C	B	A	A*
Boundary	0	22	38	54	70	88	105	122	139

For many purposes (such as calculating average grades for summarising student or school performance) it is necessary to convert the letter grades to numbers: U=0…A*=8. Ofqual's proposed new reporting scale for the reformed GCSE[1] runs from 1 to 9 (with U=0), thus removing the need for this letter-to-number conversion.

It is axiomatic that the scaling function should be non-decreasing (i.e. higher raw scores map to the same or higher scale scores and never to lower ones), but apart from that what desirable features should it have? Many different desiderata have been proposed, but the most important one, stated by Petersen, Kolen & Hoover (1989), is that "the main purpose of scaling is to aid users in interpreting test results". Some other possibilities are listed below:

- It should minimise the loss of information inherent in the raw scale;
- It should be smooth, i.e. similar differences in raw score should lead to similar differences in scale score at most parts of the raw score range;
- It should discourage users from 'reading too much in' to insignificant differences;
- It should facilitate interpretations in subject-specific terms;
- It should support inferences about reliability (repeatability);
- It should be fair to all examinees;
- It should not create more scale score categories than raw score categories[2].

However, different users may want to make different inferences from examination (test) results, and therefore some of the above criteria proposed for desirable scaling functions are contradictory. Cresswell (1986) identified two general 'schools of thought' on the issue: one relating the number

[1] http://ofqual.gov.uk/news/design-details-of-new-gcses-in-england/ Accessed 02/06/14.
[2] This was one of six recommendations for scale scores in Dorans (2002) in the context of recentering the SAT. His other recommendations are arguably less relevant to GCSEs and A levels (for example that scale scores should be normally distributed and symmetric around the midpoint of the scale). But his cited recommendation directly conflicts with what we have advocated (e.g. Bramley, 2013) and therefore needs to be considered further.

of grades to the reliability of the underlying mark scale and tending to recommend fewer rather than more scale points; the other emphasising the loss of information when a larger number of distinctions are reduced to a smaller number and tending to recommend more rather than fewer scale points. Cresswell (ibid) showed that the different approaches can coexist depending on the purpose of the reporting scale. For example, he suggested that the public expect consistency (lack of variability) in reported outcomes and that lack of such consistency leads to lack of credibility and confidence in the exam system – implying that shorter scales are required to maintain credibility of public examinations. He also suggested that if the categories on the reporting scale need to be tied to descriptions of knowledge/skills of examinees with scale scores in those categories then again this is only feasible with fewer broad categories. On the other hand, for reporting smaller differences (such as a teacher reporting progress to an individual pupil) a finer grained reporting scale is more useful.

This report explores some of these issues in more detail. Section 2 considers approaches based on the idea that the reporting scale should embed (or take account of) the reliability of the reported scores, applying some formulae given in the literature for determining the appropriate number of scale categories from GCSE and A level examples. Section 3 explores the 'information loss' perspective using a simulation based on data from a linear GCSE. Section 4 considers the implications for reporting of scale scores if statistical equating were used to link raw score scales from one year to the next. Section 5 describes the UMS scaling function used in modular A levels and GCSEs and illustrates some of its less publicised features. The discussion and summary section attempts to bring together the arguments and findings from the different sections of the report.

2. Reliability-based approach to determining the number of scale categories

Several authors have claimed that the rational or scientific way to decide on the appropriate number of scale score categories is to utilise information about the reliability of the raw scores. We consider two such approaches below.

Skurnik and Nuttal (1968) argued that since a candidate's true score is likely to lie (approximately) within ±2 standard errors of measurement (SEM) of their observed score, it is possible to calculate the number of non-overlapping bands which cover the full range of raw marks by dividing the range of marks available on the assessment by four times the SEM[3].

$$number\ of\ non-overlapping\ bands\ = \frac{RANGE}{4SD\sqrt{1-r}} \qquad (1)$$

where RANGE is the number of marks available (i.e. the total possible test score), *r* is the reliability as estimated by, for example, Cronbach's Alpha, and SD is the standard deviation of observed scores. If it is then assumed that examination scores typically have a range of six standard deviations, the formula for the number of non-overlapping bands can be reduced to (2):

$$number\ of\ non-overlapping\ bands\ = \frac{3}{2\sqrt{1-r}} \qquad (2)$$

Skurnik and Nuttal (ibid, henceforth S&N) proposed that it is reasonable to set grade boundaries such that that the true score for 95% of candidates lies within one grade of their observed score. If this is the case, then the number of grades is twice the number of non-overlapping bands, so the number of grades could be calculated by multiplying the right hand side of equation (2) by 2. However, that would overestimate the number of grade categories, because it assumes that a student's true score falls within 4 SEM, when this may not be the case. S&N proposed adding an arbitrary but useful correction, by subtracting one grade from the number of grades:

$$number\ of\ grade\ categories = \sqrt{\frac{9}{1-r}} - 1 \qquad (3)$$

This correction would have the advantages of giving a more sensible result for very low reliabilities and also of giving close to the result that would be obtained if five rather than six SDs was taken as the range of scores in (2) above.

For this research we used two indices based on S&N's scheme. The first ('SkNut1') was based on Equation 1, multiplying by two to get the number of grade categories and subtracting one grade as in Equation 3. The second ('SkNut2') used Equation (3) directly. The first index therefore explicitly uses information about the total number of marks available, whereas the second assumes this is roughly six SDs.

A similar approach was recommended by Mitchelmore (1981), who used the following as a criterion: "A grading scale is acceptable in a given assessment situation if the average probability of a student being regarded within one scale point of the original grade on a parallel assessment is at least 90%" (Mitchelmore, 1981, p219). This differs from S&N in two main ways: first the

[3] In the denominator of equation (1) the SEM is expressed as $SD\sqrt{1-r}$ which follows from the definition of reliability in classical test theory.

probability is lower (90% rather than 95%); and secondly the comparison is between two hypothetical observed scores and thus is conceptually more related to classification consistency than classification accuracy (the comparison between 'true' scores and observed scores as in S&N).

Defining grade boundaries … b_{-2}, b_{-1}, b_0, b_1, b_2 … an examinee's second grade (scale score) is within one category of their first score if the first score x lies between b_i and b_{i+1} and the second score y lies between b_{i-1} and b_{i+2}, for any value of i. The probability p of this is:

$$p = \sum_i P(b_i \leq x < b_{i+1}) \times P(b_{i-1} \leq y < b_{i+2}) \quad (4)$$

Assuming normally distributed errors, equal SEM across the score range and equally spaced boundaries it is possible to find the ratio of grade bandwidth to SEM giving an average value for p of 90%. Mitchelmore found the required bandwidth to be 1.635 SEMs wide, giving a value for p of 89.8% when the observed score is in the middle of a grade and 90.2% when it is on a boundary. The number of scale categories implied is therefore equal to the full mark range divided by 1.635 times the SEM.

Tables 2.1 to 2.3, Figure 2.1 and Figure 2.2 below apply these formulae to data from the June 2013 session of OCR exams at GCSE and A level. Reliabilities are routinely calculated at unit/component level for all units/components marked on screen. Analysis was restricted to the 414 units/components marked on screen and taken by 100 or more examinees.

Table 2.1: Distribution of reliability index[4].

Qual	Tier	N	Mean	Std Dev	Minimum	Maximum
A level	Untiered	200	0.837	0.087	0.460	0.950
GCSE	Foundation	59	0.798	0.084	0.460	0.931
	Higher	59	0.838	0.074	0.490	0.933
	Untiered	96	0.849	0.055	0.668	0.970

Table 2.2 on the next page shows that the Mitchelmore criterion is less stringent than the S&N ones (as might be expected from the fact that it uses a 90% probability value instead of 95%). A level units report 6 categories (A-E,U), GCSE foundation tier units report 6 categories (C-G,U), GCSE higher tier units report 7 categories (A*-E, U) and untiered GCSE units report 9 categories (A*-G, U).

[4] Cronbach's Alpha for units/components with no question choice (the majority), Backhouse's P for the rest.

Table 2.2: Percentage of units/components reporting ‘too many’ grades according to the criteria of Skurnik & Nuttal and Mitchelmore.

Qual	Tier	Variable	N	% not meeting criterion
A level	Untiered	SkNut1	200	24
		SkNut2	200	30
		Mitchelmore	200	2
GCSE	Foundation	SkNut1	59	10
		SkNut2	59	58
		Mitchelmore	59	3
	Higher	SkNut1	59	53
		SkNut2	59	47
		Mitchelmore	59	2
	Untiered	SkNut1	96	82
		SkNut2	96	83
		Mitchelmore	96	26

It can be seen from Table 2.2 that the majority of untiered GCSE units report too many grade categories by the Skurnik and Nuttal criteria. This is illustrated below in Figure 2.1.

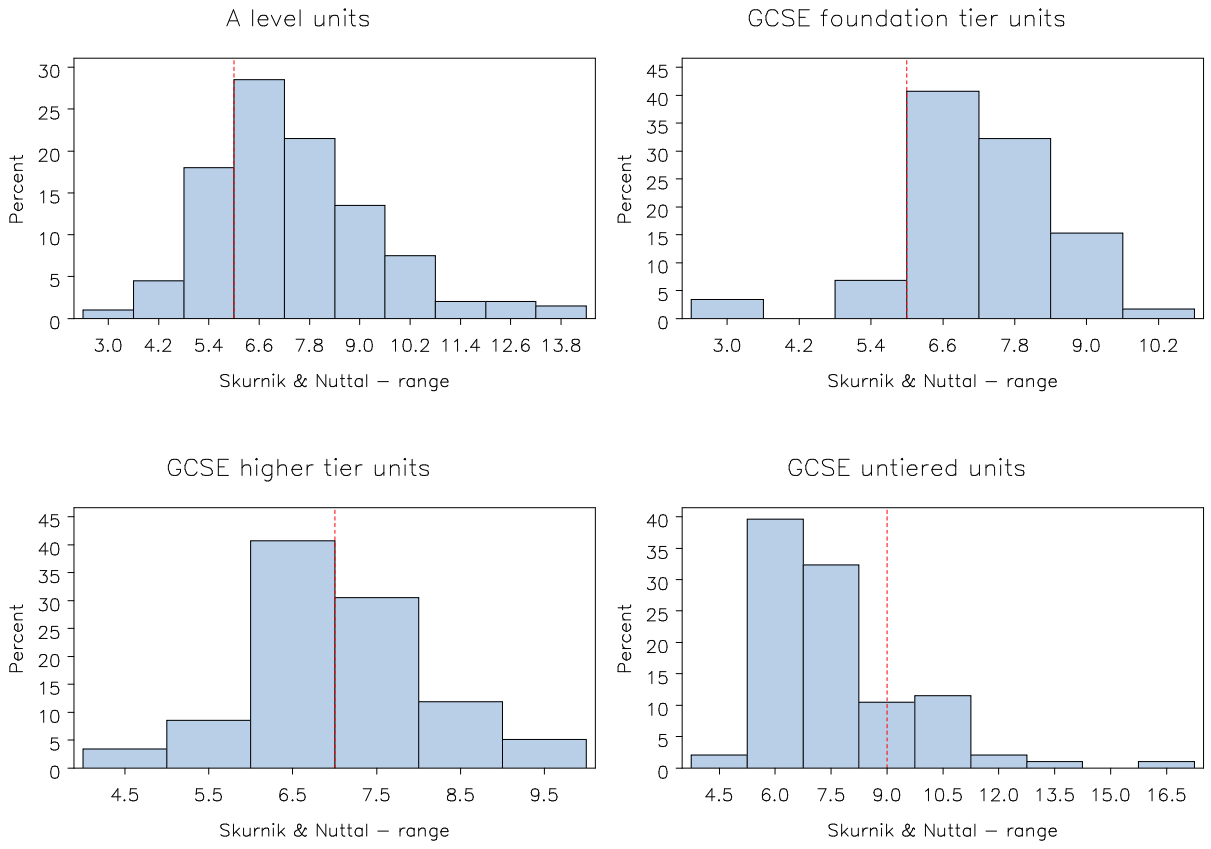

Figure 2.1: Distribution of the number of reporting categories implied by the first of the S&N indices (red dashed line indicates the number of grades actually reported).

Both the S&N and Mitchelmore formulae assume that the mark range is divided into equal sized grade bands. In practice this is not what happens at GCSE and A level – usually it is the case that the extreme grade categories (e.g. A and U for an A level unit) cover a wider range of marks than the intermediate grades. A more appropriate comparison therefore is arguably that of the actual grade bandwidth (defined here as the distance in marks between the A and B boundaries for A level units and GCSE higher or untiered units, and between the C and D boundaries for GCSE foundation tier units) and the minimum grade bandwidth implied by the S&N and Mitchelmore formulae.

Table 2.3: Percentage of units/components with too low a grade bandwidth according to the criteria of Skurnik & Nuttal, and Mitchelmore.

Qual	Tier	Variable	N	% not meeting criterion
A level	Untiered	SkNut1	200	95
		SkNut2	200	96
		Mitchelmore	200	72
GCSE	Foundation	SkNut1	59	88
		SkNut2	59	90
		Mitchelmore	59	42
	Higher	SkNut1	59	88
		SkNut2	59	85
		Mitchelmore	59	32
	Untiered	SkNut1	96	97
		SkNut2	96	95
		Mitchelmore	96	77

Table 2.3 gives a more discouraging picture than Table 2.2. It is only for the tiered GCSE units using the Mitchelmore formula where the majority meet the criterion.

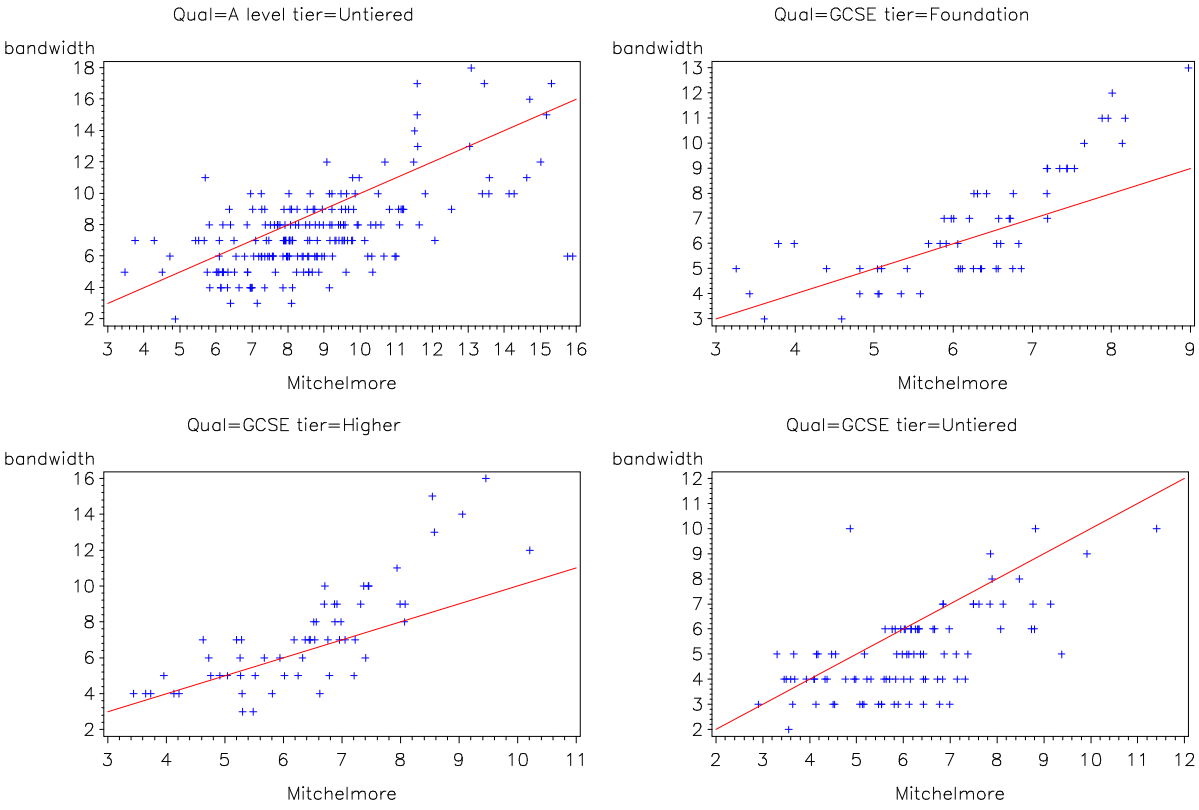

Figure 2.2: Plot of actual bandwidth against minimum recommended by Mitchelmore criterion (units/components above the red x=y reference line meet the criterion).

However, it is more important that the reporting scale for the full qualification meets the criteria, rather than the individual units/components comprising it. As noted by Bramley & Dhawan (2013) and Benton (2014) it is not straightforward to derive an estimate of composite reliability for unitised (modular) qualifications because of the variety of routes through these qualifications, the fact that unit raw scores are scaled before aggregation, and the fact that an estimate of unit-level reliability is often not available for some units (for example controlled assessment, portfolio, practical units and other units not marked on-screen).

Nonetheless, we felt it was important at least to attempt to apply the S&N and Mitchelmore formulae at qualification level to get a feel for the results, particularly since the number of reporting categories will be increased in the reformed GCSE (from 9 to 10). It was possible to do this for a selection of A levels and GCSEs by making the assumption that the SEM is relatively invariant to changes in the cohort and that therefore a composite SEM can be estimated as the square root of the sum of the squares of the unit level SEMs, even though these units were taken by different (albeit overlapping) groups of examinees. The results below can be taken as applying to a quasi-linear examination obtained by simply adding up the unweighted raw scores on the units, and using grade boundaries corresponding to the sum of the unit boundaries. It was not possible to use the second of the S&N formulae because there was no way of estimating the reliability of the scores at aggregate level.

Table 2.4: Number of scale categories and grade bandwidth at qualification level (quasi-aggregate of units) for selected A levels and GCSEs.

Qual.	Name	Tier	Spec	Total Mark	Max. number of grades			Min. grade bandwidth		
					Actual	SkNut1	Mitch	Actual	SkNut1	Mitch
AL	Maths		7890	432	7[5]	16.6	21.5	38	26.0	20.1
AL	Biology		H421*	320	7	16.4	21.3	21	19.5	15.1
AL	Bus. Stud.		H430	300	7	11.9	15.7	22	25.3	19.1
AL	Economics		H461	240	7	10.0	13.4	20	24.1	17.9
AL	French		H475*	280	7	14.6	19.1	24	19.1	14.7
AL	Geography		H483	300	7	16.7	21.6	22	18.0	13.9
AL	Psychology		H568	360	7	17.9	23.1	26	20.2	15.6
GCSE	Chemistry	F	J244*	180	6	11.8	15.6	20	15.3	11.5
GCSE	Chemistry	H	J244*	180	7	11.7	15.6	19	15.4	11.6
GCSE	Science B	F	J261*	160	6	11.8	15.6	19	13.6	10.2
GCSE	Science B	H	J261*	160	7	11.4	15.2	20	14.0	10.5
GCSE	Greek	U	J291*	220	9	20.3	26.1	16	10.8	8.4
GCSE	Law	U	J485	240	9	13.0	17.1	16	18.5	14.1
GCSE	Rel. Stud.	U	J620	204	9	14.7	19.2	12	13.9	10.6
GCSE	Sociology	U	J696	240	9	13.3	17.4	15	18.1	13.8

*Not all units used in aggregation.

Table 2.4 shows a much more encouraging picture at qualification level. In every case the number of scale categories that could be reported according to the S&N and Mitchelmore criteria exceeded the number actually reported, and in the majority of cases the actual grade bandwidth was wider than the minimum implied by the criteria.

This suggests that the linear reformed GCSEs with 10 reporting categories should achieve the desirable property (if it is desirable!) that an examinee's 'true score' should be within ± 1 scale point of that observed. This assumes that the majority of error in exam scores comes from sampling of questions (rather than markers), because this is what is captured by Cronbach's Alpha. Bramley & Dhawan (2012) showed that this assumption is reasonable for the kind of subjects with units marked on-screen at the time of their report, which did not include most long-answer / essay-based units. However, if the majority of marker-related error is unsystematic (i.e. markers are as likely to be severe as lenient across questions) then Cronbach's Alpha will include the effects of marker unreliability to a certain extent (Hutchison & Benton, 2009, p40-41). There is a growing body of evidence to support the contention that this is in fact the case – i.e. most marking error is not attributable to systematic severity or leniency (see for example Benton 2006; Baird, Hayes, Johnson, Johnson & Lamprianou, 2013).

[5] A* is reported at A level at qualification level, but not at unit level.

3. Information-loss approach to determining the number of scale categories

The basis of this approach is the recognition that information is lost when applying a scaling function leading to a reporting scale with fewer categories than the raw score scale. The resulting scores are less precise, in the everyday sense of the word. Just as a measurement in millimetres is less precise when rounded to the nearest centimetre, so a score on a raw score scale with 100 categories is less precise when reported on a scale with 10 categories. (The function creating the reporting scale need not literally involve dividing the raw score by a constant and rounding the result to the nearest integer, although this is one way to create a reporting scale with equally spaced intervals). Information is lost in the sense that discriminations that could be detected on the original scale (e.g. between objects with length 968mm and 973mm) become lost on the new scale (both reported as length 97cm). The argument of those advocating longer scales is essentially along the lines of 'having gone to the trouble of collecting the information at a given level of precision, why throw some of it away by reporting on a more coarse-grained scale'? This argument was forcefully given by Ebel (1969):
"Regardless of the inaccuracy of the basis for grading, the finer the scale used for reporting grades, that is the more different grade levels it provides, the more accurate the grade reports will be." (p217).
This is shown by considering the formula for the variance of the sum of two components where E represents measurement error and R represents rounding/grouping error:

$$\sigma^2_{E+R} = \sigma^2_E + \sigma^2_R + 2cov(E, R)$$

The variance of the sum can never be less than the variance of either the measurement error or the rounding/grouping error unless they somehow have a negative covariance, which could only arise in specially contrived circumstances, and not in general.

One way of quantifying the information retained when a finer-grained scale is grouped or rounded for reporting is using the squared correlation ρ^2 between the two scales. Shaw, Huffman and Haviland (1987) noted three advantages of this index: i) maximising ρ^2 is associated with choosing the linear function of the integers 1... k that best approximates the original values; ii) ρ^2 is unaffected by any linear transformation of the reporting scale (e.g. converting a 0-9 scale to a 100-190 scale with possible values 100, 110 … 180, 190); iii) ρ^2 is the proportion of variance in the original distribution retained in the distribution of scale scores (in the sense of linearly predictable from them). By varying the number of intervals and the interval widths when grouping continuous data from four reference distributions (normal, uniform and two chi-square) and calculating ρ^2 they concluded that:

- Increasing the number of intervals reduces information loss, provided these intervals break up the central part of the distribution;
- Interval width is not as influential a factor as number of intervals because for a given number of intervals the width can deviate substantially from the optimum without significantly affecting loss;
- Information retention is similar across all four distributions, increasing with increased dispersion (as is intuitively obvious – if all observed scores map to the same grade, all the information in the scores will be lost);
- Eight categories retain 95% of the information and beyond this the information retained only increases slowly.

This again would suggest that both the existing 9-category and new 10-category reporting scales for GCSE retain an acceptable amount of information. However, ρ^2 is not the only possible index of information retention. We thought it would also be of interest to know what proportion *p_agree* of the ordinal comparisons between pairs of examinees A and B (i.e. A<B, A=B, A>B) on the raw score scale are preserved on the reporting scale, across all possible pairs[6].

To investigate information retention in the context of measurement error and rounding/grouping error we used a larger and more realistic simulation than that of Ebel (1969). The basis of the simulation was OCR's linear GCSE in Religious Studies from 2009 (see earlier for the scaling function for this examination). First, 10,000 'true scores' were simulated using the beta distribution with parameters chosen (by trial and error) to give a score distribution in the range 0-168 with a similar shape to that of the actual distribution, and a variance of 90% of the observed variance. Second, an error score drawn from a normal distribution with mean 0 and SD (x) was added to each true score, where (x) was the value needed to contribute 10% of the observed variance. Thus observed scores with a reliability of 0.9 (ratio of true variance to observed variance) were simulated. This is a reasonable, perhaps even slightly cautious value given the results in Table 2.4. The simulated observed scores were thus continuous and could fall outside the range 0-168. The next step was to round both the true scores and the observed scores to the nearest integer, converting observed scores below 0 to 0 and scores above 168 to 168. These integer scores were taken as the basis for all future rounding/grouping operations.

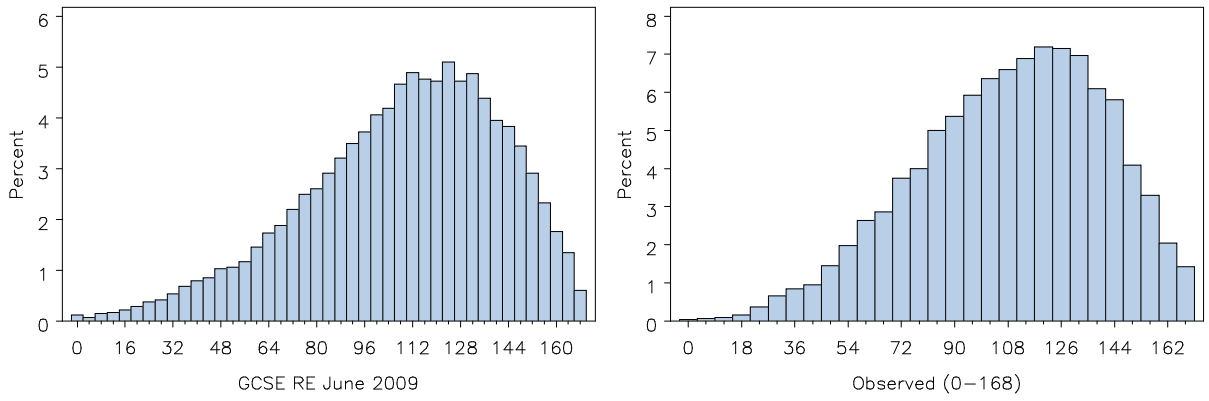

Figure 3.1: Distribution of actual (left) and simulated (right) observed scores.

Further scores were produced on a 101-category scale (i.e. 0-100), a 51-category scale (0-50), an 11-category scale (0-10), a 4-category scale (0-3) and a 2-category scale (0-1). In all cases the scale scores were produced by multiplying the 169-category score by a factor of m/168 where m was the maximum possible score on the coarser scale, and then rounding to the nearest integer. This process gave equally spaced intervals across the majority of the score range with half-width intervals at the extremes (see Figure 3.2 below). A scaling function corresponding to application of the original grade boundaries was also used to give a 9-category scale, and alternative versions of the 4- and 2-category scales were created by applying boundaries at the 75th, 50th and 25th percentile points for the 4-category scale and at the 50th percentile (median) for the 2-category scale. This was to achieve the best possible discrimination for these coarse-grained scales.

[6] *P_agree* is related to the indicator of ordinal agreement gamma (Goodman & Kruskal, 1979) by the formula gamma=(2**p_agree*)-1, providing tied pairs are excluded. The inclusion of ties was deemed important here because we we treat as 'loss' a situation where two examinees found to be different on a finer scale are regarded as the same on a coarser scale. But it should be noted that the inclusion of ties can have a significant effect on the size of this and similar statistics.

These scaling functions are illustrated on the next page.

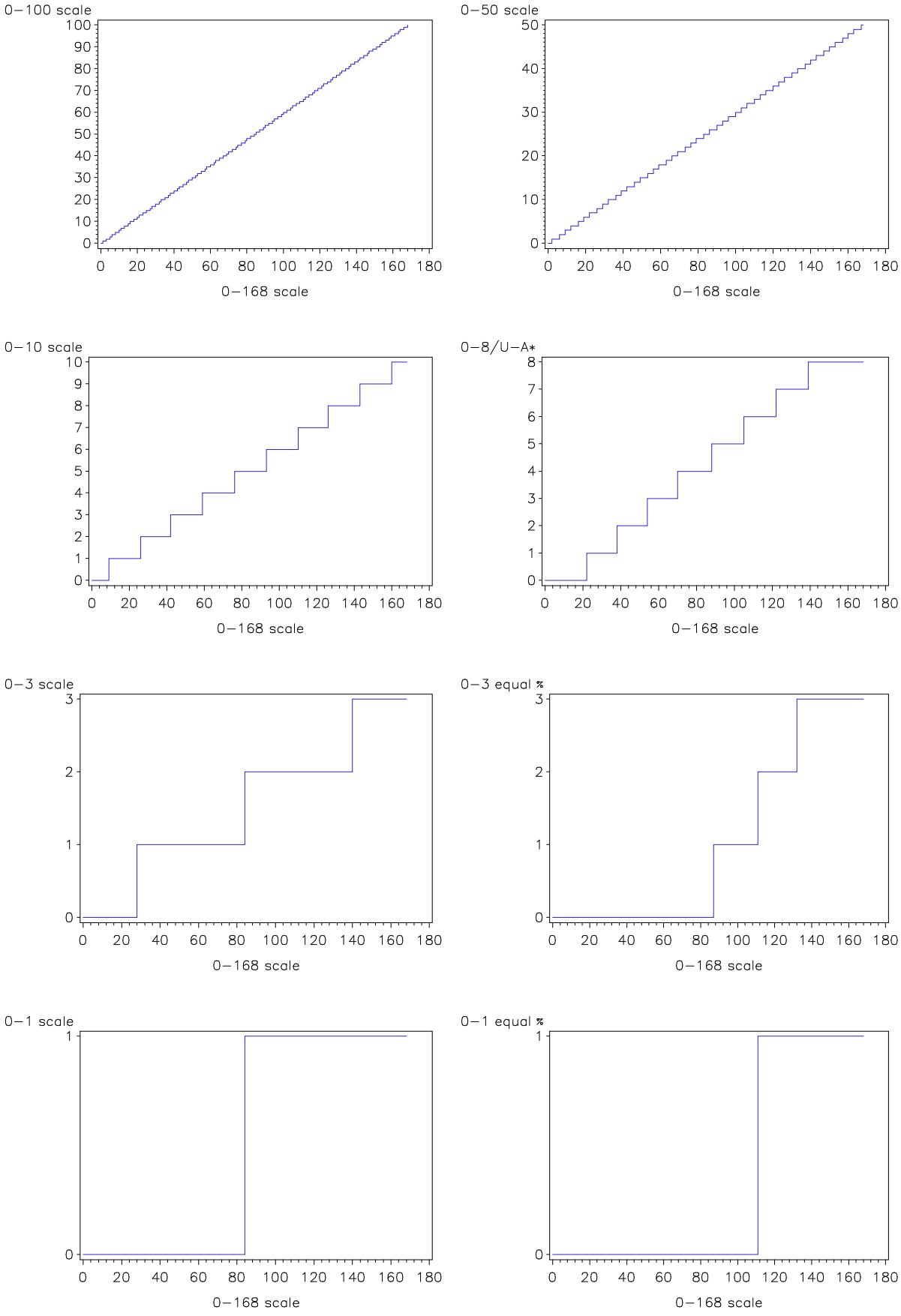

Figure 3.2: Illustration of scaling functions applied to the simulated RS data.

For each scale, the effect of measurement error on accuracy and precision was assessed by comparing the true scores with the observed scores. The effect of rounding/grouping into *n* categories was assessed by comparing the true scores on the 169-point scale with true scores on the *n*-point scale; and observed scores on the 169-point scale with observed scores on the *n*-point scale. The combined effect of measurement error and rounding error was assessed by comparing the true scores on the 169-point scale with the observed scores on the *n*-point scale. The results are in Table 3.1 below.

Table 3.1: Comparison of simulated true and observed scores at different scale lengths.

Comparing…	…with	ρ^2 (%)	*p_agree* (%)	Classification accuracy (%)
true_score	obs_score	90.14	89.91	0
true_score	true_169	99.99	99.06	
obs_score	obs_169	99.95	99.11	
true_score	obs_169	90.22	89.46	
true_169	obs_169	90.22	89.03	4.24
true_169	true_101	99.97	99.24	
obs_169	obs_101	99.98	99.28	
true_169	obs_101	90.20	88.68	
true_101	obs_101	90.18	88.33	7.52
true_169	true_51	99.90	97.72	
obs_169	obs_51	99.91	97.86	
true_169	obs_51	90.16	87.98	
true_51	obs_51	90.07	87.07	13.71
true_169	true_11	97.43	85.39	
obs_169	obs_11	97.71	86.31	
true_169	obs_11	88.14	80.71	
true_11	obs_11	85.93	80.13	54.79
true_169	true_9	97.07	84.42	
obs_169	obs_9	96.66	84.66	
true_169	obs_9	88.09	79.49	
true_9	obs_9	86.34	80.40	60.35
true_169	true_4	77.10	55.43	
obs_169	obs_4	79.29	57.05	
true_169	obs_4	70.64	55.02	
true_4	obs_4	65.90	78.51	84.70

Comparing…	…with	ρ^2 (%)	*p_agree* (%)	Classification accuracy (%)
true_169	true_4e[7]	88.32	75.93	
obs_169	obs_4e	87.68	75.88	
true_169	obs_4e	80.08	72.22	
true_4e	obs_4e	80.81	79.63	75.22
true_169	true_2	59.42	34.87	
obs_169	obs_2	58.90	35.70	
true_169	obs_2	55.58	34.99	
true_2	obs_2	67.07	88.12	93.77
true_169	true_2e	67.39	50.93	
obs_169	obs_2e	66.75	50.87	
true_169	obs_2e	61.90	49.51	
true_2e	obs_2e	67.85	83.54	91.19

The rows in Table 3.1 containing a value for classification accuracy illustrate the effect of measurement error alone for a given scale length. The values for both indices of information retention (ρ^2 and *p_agree*) generally decreased as the number of scale categories decreased. This re-iterates Ebel's point that it is not possible to compensate for measurement error by reporting in broader categories (i.e. by rounding or grouping). The only exception to the decreasing trend was that the categories actually used (9-category scale) had better information retention than the 11-category scale created by the rounding formula. This is because they 'broke up the distribution' better in the sense of Shaw et al (1987). Similarly the divisions of the 4 and 2-category scale that were based on equal percentages of the distribution (_4e and _2e) had higher values for ρ^2. A further advantage of the ρ^2 indicator in this context is that when true scores are compared with observed scores it is equal to the reliability (by definition). The 169-, 101- and 51-category scales had the same reliability (90%) as the original continuous scores: it was not until the scale was reduced to 11 categories that a reduction to 85% was observed.

However, the intuition that grouping ought to compensate for measurement error is given some support by the values for classification accuracy (the proportion of examinees with a true score the same as the observed score), which increased from 0 (for continuous scores) to 93.77% (for a 2-category scale with a boundary at half-marks). For both examples of scales of the same length with different boundaries, the 'worse' boundary in terms of retaining information (ρ^2) had a higher value for classification accuracy. This shows that using classification accuracy alone as an indicator of scale quality is potentially misleading, a point that has been made before (e.g. Bramley, 2010).

The first two rows in each block of four in Table 3.1 show the effect of rounding / grouping from a 169-category scale to an n-category scale, whether from true to true or from observed to observed. For both indicators of information retention, rounding to a 101-category scale (e.g. a percentage scale) retained over 99% of the information, and for ρ^2 even rounding to an 11-category scale

[7] The _4e and _2e are the cases where the boundaries were set to give equal percentages in each scale category.

retained over 97%. The *p_agree* index tailed off more quickly as the number of scale categories reduced, showing that from the perspective of correctly ranking pairs of examinees, more information is lost by rounding / grouping.

This last point is illustrated in Table 3.2 below, which shows the indices of agreement for the 169-category scale compared with the actual (9-category) grade scale in a hypothetical selection scenario where the pool of applicants consists of those with an observed score of 5 or 6 (C or B).

Table 3.2: Indices of agreement for the sub-set of simulated examinees with an observed grade of B or C.

Comparing…	…with	ρ^2 (%)	*p_agree* (%)	Classification accuracy (%)
true_169	true_9	87.95	68.67	.
obs_169	obs_9	74.68	52.72	.
true_169	obs_9	36.04	43.58	.
true_9	obs_9	31.46	54.62	56.91
true_169	obs_169	48.95	73.00	.

Comparing the 'obs_169' with the 'obs_9' shows that a quarter of the information about this sub-set has been lost by reporting on the grade scale (using ρ^2), and nearly half of the information (using *p_agree*). The comparisons of true_169 with obs_9 and with obs_169 show that rounding/grouping error has had a substantial extra effect (on top of measurement error). It is not unreasonable to assume that different courses / institutions / jobs attract applicants from different parts of the ability range, so this illustrates that the information lost by grouping/rounding could be a more serious problem than the figures in Table 3.1 might have suggested.

4. Reporting on a percentage scale

Bearing in mind the overarching aim of scaling to aid users in interpreting test results, in this section we consider whether a much longer scale than the traditional grade scale could achieve this. We have seen earlier that Cresswell (1986) noted that a short scale is necessary if the categories are to be related in some way to descriptions of what examinees labelled in that category know and can do. There seems to be a general consensus, however, that the grade labels at GCSE and A level do not really provide this kind of information. This can be seen both from the general 'narrative' around the lack of skills of examinees entering higher education or the workforce, whether in literacy, numeracy, ICT, critical thinking etc; and also in the literature relating to the usefulness or otherwise of 'grade descriptors' summarising the putative knowledge and skills of examinees on the borderline (or in the middle) of a given grade category (see for example Robinson, 2007).

Taking the pessimistic view that a particular grade label such as 'B' in practice can often mean little more than 'similar to other examinees with a B, not as good as those with an A, but better than those with a C', suggests that in fact the categories are often used (for example in a selection context) to make the kind of ordinal comparisons captured by the *p_agree* statistic. The longer the

scale, the more accurate this kind of comparison is, as shown in the previous section. Clearly though, to persuade a wider audience to adopt a longer scale it would help to have some way to link the scale scores to knowledge and skills.

A percentage scale would have the obvious attraction that the meaning of an individual scale score would be readily apparent – the percentage of the maximum available raw marks achieved by the examinee. On a test containing only dichotomous items it would be the percentage of questions that the candidate answered correctly. Reporting percentages also accounts for differences in the lengths (in maximum marks or number of questions) of different examinations. The percent-correct score is easy to calculate and understand and it is already used, for example, in classrooms for score reporting.

However, it is hard to use the percentage score for fair comparisons of candidates' performances on different examinations of the same qualification (*e.g.* January or June series; different years). For example, getting 50% of the marks on a difficult test may mean the candidate has more knowledge and skill than another candidate getting 60% of the marks on a relatively easier test. In other words, two candidates with the same percentage score on two different examinations of the same qualification do not necessarily have the same level of achievement.

In order to be able to compare results from different examinations the scores need to be comparable, that is, scores on different forms of an examination should indicate the same level of performance no matter which form of the test the candidate took. This is the main reason why percentage scores are not usually used as a scale for reporting assessment results because such scores are not comparable across forms. The complex grade awarding procedures for GCSE and A level (Ofqual, 2011) that are followed to locate the boundaries on the grade scale aim to find a scaling function that supports claims of equivalence of scale scores both over time within a board, and across boards at any particular time.

Focusing first on the more tractable problem of comparability within a board over time, Bramley & Vidal Rodeiro (2014) argued that with linear examinations it is possible to link (equate) complete raw score scales using established statistical equating methods, showing that the 'comparable outcomes' approach (Ofqual, 2012) can be viewed as a special case of one of these methods (the frequency estimation equipercentile equating method). If we have an exam taken in years Y and Y+1, it is possible to find an equating function that maps raw scores on year Y+1 onto the raw score scale of year Y. Then whatever scaling function was applied for the reporting of scores in year Y can be applied to the equated scores on year Y+1. This could be a case of simply applying the same grade boundaries – but it need not be. If scores in year Y were reported as percentages, the scores in year Y+1 could also be reported as 'equated percentages'. The interpretation of a scale score of 65% on year Y+1's test would then be 'a performance of equivalent quality to one which gained 65% of the available marks on the year Y test'. If the year Y test and its mark scheme were published and available for the perusal of teachers / admissions tutors / other stakeholders this would in principle allow at least some interpretation of scale scores in terms of knowledge and skills.

To illustrate this, the 2010 version of the same GCSE RS examination described earlier was equated to the 2009 version, using the open source statistical software R (R core team, 2013) with the package *kequate* (Andersson, Branberg & Wiberg, 2013). The equating method was a kernel approach to smooth the frequency estimation equipercentile equating technique (see Bramley & Vidal Rodeiro (2014) for a comparison of multiple equating methods in this context and a full description of the method used here).

Figure 4.1 below shows the results of the equating, with the 2010 test score on the *x*-axis, and the difference between the equated scores and the 2010 scores on the *y*-axis. Negative values mean that the 2010 test was easier than the 2009 test (*i.e.* you would need to reduce scores to equate them to 2009) and vice versa. So, in this case, the 2010 examination seemed easier at the top of the mark range (marks in the A and A* ranges) than the 2009 test. On the other hand, it was more difficult than the 2009 in the F to B ranges.

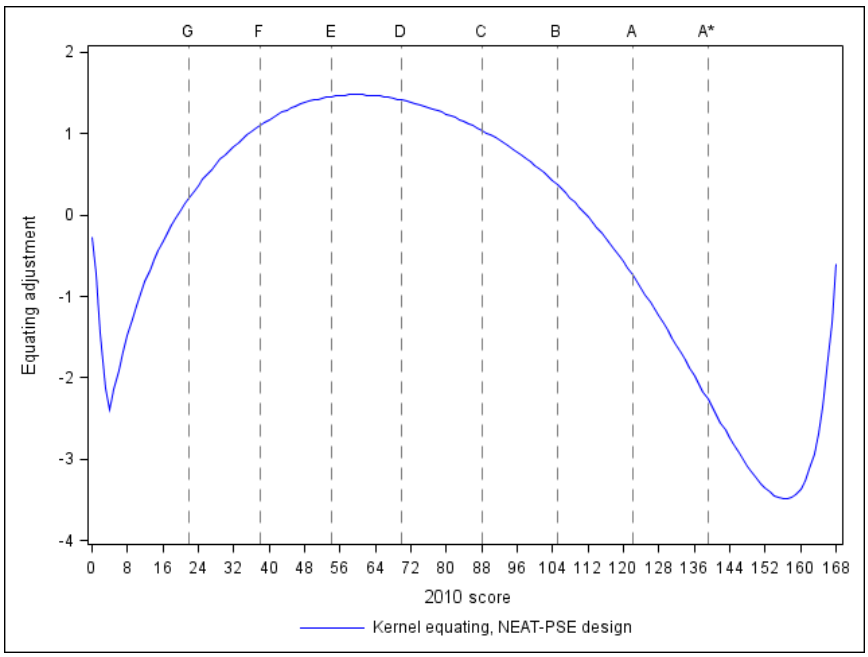

Figure 4.1: Equating adjustment (2010 raw scores equated to 2009 raw scores).[8]

[8] The grade boundaries in Figure 4.1 are those from 2009: A*=139; A=122; B=105; C=88; D=70; E=54; F=38; G=22.

Table 4.1: Selection of raw scores, equated scores and rounded percentage scores from the 2009 and 2010 exams.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
2009 raw	2009 %	2009 % rounded	2010 raw equated	2010 equated %	2010 equated % rounded
0	0.000	0	-0.263	-0.156	0
1	0.595	1	0.288	0.172	0
2	1.190	1	0.550	0.328	0
3	1.786	2	0.882	0.525	1
4	2.381	2	1.606	0.956	1
5	2.976	3	2.850	1.696	2
…	…	…	…	…	…
40	23.810	24	41.177	24.510	25
41	24.405	24	42.210	25.125	25
42	25.000	25	43.240	25.738	26
43	25.595	26	44.268	26.350	26
44	26.190	26	45.294	26.961	27
…	…	…	…	…	…
80	47.619	48	81.245	48.360	48
81	48.214	48	82.223	48.942	49
82	48.810	49	83.200	49.524	50
83	49.405	49	84.176	50.105	50
84	50.000	50	85.152	50.686	51
…	…	…	…	…	…
120	71.429	71	119.425	71.086	71
121	72.024	72	120.348	71.636	72
122	72.619	73	121.270	72.184	72
123	73.214	73	122.190	72.732	73
124	73.810	74	123.108	73.279	73
…	…	…	…	…	…
164	97.619	98	161.313	96.019	96
165	98.214	98	162.644	96.812	97
166	98.810	99	164.083	97.668	98
167	99.405	99	165.671	98.614	99
168	100.000	100	167.396	99.640	100

Column 1 of Table 4.1 shows the possible raw scores on each test. Column 2 shows these scores as a percentage (to 3 decimal places) for the 2009 test. No information has been lost by this transformation, but rounding to integers in Column 3 means that some percentage scores are mapped to by more than one raw score. Column 4 shows scores from 2010 expressed on the raw score scale of 2009 (again rounded to 3 d.p.). For example, a score of 81 in 2010 corresponds to a score of 82.223 in 2009. These scores are expressed as a percentage in Column 5 and then finally as a rounded (integer) percentage in Column 6. Thus a raw score of 81 in 2009 would receive a scale score of 48, and a raw score of 81 in 2010 would receive a scale score of 49. In the former case this scale score can be interpreted directly as the percentage of marks gained, and in the latter case it can be interpreted as the percentage of marks that the examinee would have gained if they had taken the 2009 exam. In this way the scale scores are tied to the 2009 question

papers and mark schemes, and provided both exams were constructed to the same specification, inferences about knowledge and skills should be possible.

Of course, equating is not error-free and usually there is more equating error[9] in the parts of the mark range where the data is sparse. Furthermore, there is an increase in the size of the standard error of equating in small samples (this could have an implication for syllabuses with small entries). Figure 4.2 below shows the error equating of the Kernel frequency estimation equipercentile equating carried out for the Religious Studies dataset. As expected the standard error of equating was higher at the bottom of the mark distribution, which corresponds with the part of the distribution with fewer candidates. This might lead to problems in the comparability of the raw scores, and therefore the percent-correct scores, in those mark ranges. On the other hand, fewer examinees would (by definition) be affected by this.

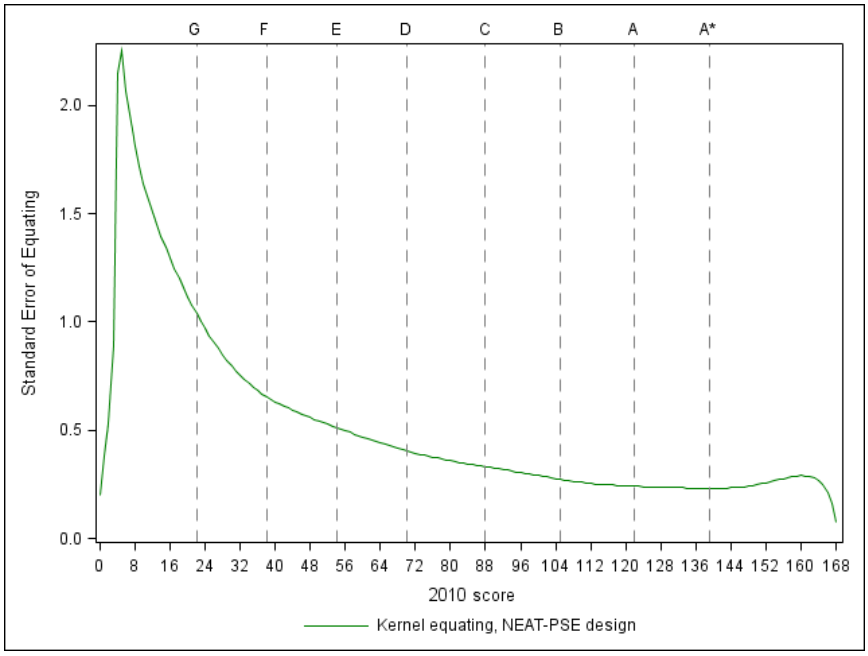

Figure 4.2: Standard error of equating, Religious Studies

The possibility of 'gaps' in the reporting scale

One feature of linear examinations in the current system, so obvious that it is hardly worth mentioning, is that each grade is mapped to by some range of scores – that is, the procedures guarantee that each grade always has some range of marks associated with it, even if no examinees score in that range. However, with a non-linear equating function and a reporting scale with a large number of categories (like the percentage scale) it is not necessarily the case that every scale score from 0 to 100 is associated with a raw score, even if there are more raw score points than scale score points, as in the RS example. For example, a raw score of 143 might map to 84 and a raw score of 144 might map to 86, creating a 'gap' at 85 which would be unobtainable on this particular test. When using relatively long scales gaps are a problem that would probably affect a relatively small number of examinees (assuming they are distributed reasonably over the whole scale), but it is possible to imagine a scenario where two people are applying for a course (or job), one with a scale score of 84 and one with a scale score of 85, and it so happens that the score of 85 was unobtainable on the version of the exam taken by the applicant with a score of 84.

[9] Here 'error' relates to the possibility of different equating outcomes with different samples from the same population. Since the entire dataset is used, arguably this is not relevant – but there is still a strong and reasonable intuition that equating outcomes based on little or no data should be treated with caution.

They might argue that this is unfair. Two legitimate responses would be that the applicant with a score of 85 would have a higher equated score on the raw scale and is thus (however slightly) likely to be of higher achievement; and second that the likely measurement error means that it is reasonable to consider both applicants to be of similar achievement and look for another basis on which to discriminate between them. (See Cresswell (1986) for further discussion of this issue).

It was of interest to explore how likely this situation would be to arise in practice. We used the same examination (GCSE RS) but this time used all the years from 2006 to 2010 with 2006 as the reference year. The reason for this is that the grade boundary locations were very different in 2006 from subsequent years implying a test of different difficulty, and hence greater non-linearities in the equating functions that would in turn be likely to produce the 'gaps' described above. On the 'flight of stairs' type plots like Figure 1.1 and Figure 3.2 such gaps would correspond to a part of the graph where a vertical segment covers more than one scale score unit. Conversely the more common and indeed inevitable instance when going from a 169-point scale to a 101-point scale is where more than one raw score maps to the same scale score, indicated by a horizontal segment covering more than one raw score unit.

We found that using a percentage scale and a reference year of 2006, there was no instance in the years 2007-2010 where any scale score was unavailable. We therefore experimented with 'stretching' the percentage scale to 111 categories (i.e. 0-110), 126 categories (0-125) and 151 categories (0-150) to see if this would lead to any gaps. The following results were obtained:

- All scores in the 0 – 110 scale were available to at least one candidate in 2007, 2009 and 2010. However, there were some scores not available in 2008.
- All scores in the 0 – 125 scale were available to at least one candidate in 2007 but not in subsequent years.
- In the 0 – 150 scale, there were some scale scores in all years from 2007-2010 that were not available.

An example of some 'gaps' in the scaling function at the bottom end of the scale in 2008 is shown in Table 4.2 below.

Table 4.2: Equated scores from 2008 on the 2006 scale, reported in scales of different lengths.

		2008 scaled score Number of categories			
2008 raw score (169 categories)	2008 on 2006 raw scale	101	111	126	151
0	-0.080	0	0	0	0
1	0.633	0	0	0	1
2	1.754	1	1	1	2
3	3.398	2	2	3	3
4	5.036	3	3	4	4
5	6.661	4	4	5	6
6	8.265	5	5	6	7
7	9.862	6	6	7	9
8	11.436	7	7	9	10
9	12.999	8	9	10	12
…	…	…	…	…	…

It can be seen that scale score 8 was unavailable for scales with 111, 126 and 151 categories, scale score 2 unavailable for the scale with 126 categories, and that scale scores 5 and 11 were unavailable for the scale with 151 categories. We consider further in the discussion section how undesirable it is for such 'gaps' to exist, but before that we explore the scaling function in the current system for reporting grades in modular A levels and GCSEs.

5. The UMS scaling function

The Uniform Mark Scale (UMS) was devised not as a final reporting scale, but as a necessary intermediate step to solve (or provide the least unsatisfactory solution to) the problem of how to aggregate results from units (modules) of modular assessments. Because such modules are taken by different groups of examinees at different times and can be added up in a multitude of different combinations to achieve a valid final result, there needed to be a way to allow fairly for differences in difficulty (lack of comparability of the raw mark scales) at unit level before aggregation. See Gray & Shaw (2009) and AQA (2009) for detailed descriptions and explanations of the use of UMS, which will not be repeated here. Briefly, the UMS is applied at unit level and is a scaling function that maps raw scores into 'uniform marks' on a scale where the raw grade boundaries correspond to fixed percentages of the maximum uniform mark. At A level these are 80% for A, 70% for B, …, 40% for E[10]. The maximum uniform mark itself depends on the weighting of the unit in the overall aggregation. A raw mark between two grade boundaries maps to the equivalent proportional point between the corresponding UMS boundaries. UMS scores are rounded to integers before aggregation.

If there are more UMS scores than raw marks available, gaps (marks unachievable on the UMS scale) are inevitable – but they are possible even when the number of UMS score categories is equal to (or even less than) the number of raw marks available, since they can arise whenever the distance between two raw grade boundaries is less than the distance between the corresponding UMS boundaries. Previous research (Bramley, 2012) showed that a large number of A level units had raw grade bandwidths that were less than the 'target' values (which are for there to be the same difference between boundaries in percentage points on the raw scale as on the UMS scale[11]). For units with the same number of raw and UMS score categories available, this would imply the existence of gaps.

Table 5.1: Relation between number of raw and UMS score categories (OCR GCSE and A level units, June 2013)[12].

Qualification	Tier	Number of units	Number of UMS categories…					
			…equal to raw		…higher than raw		…lower than raw	
			N	%	N	%	N	%
A level	N/A	426	142	33.3	272	63.8	12	2.8
GCSE	Foundation	63	0	0.0	63	100.0	0	0.0
GCSE	Higher	62	0	0.0	60	96.8	2	3.3

[10] See http://www.ocr.org.uk/Images/126279-unit-level-ums-grade-boundaries-november-2012-january-and-june-2013.pdf for full information.
[11] Note that this does not apply for tiered GCSEs where the concept of 'target boundaries' appears not to be so clearly defined.
[12] In contrast to Table 2.1, here all units/components including coursework, practicals etc. are shown. We faced the practical problem that sometimes essentially the same unit has more than one component code (for example for 'coursework carried forward', 'postal moderation' and 'OCR repository'). We made every effort not to double-count such units based on component labels in the ISP Warehouse, but it is possible that some duplicates slipped through the net.

GCSE	Untiered	242	67	27.7	132	54.5	43	17.8

Table 5.1 shows that for the majority of units, the number of UMS categories available exceeded the number of raw score categories available. This was almost always the case for tiered GCSE, but less so for untiered GCSEs. In other words, gaps are inevitable in the scaling function at unit level for majority of units.

But for a substantial minority of units, the number of scale categories available was equal to or less than the number of raw score categories. As described above, gaps are still possible on such units if the raw boundaries are closer together than the corresponding UMS boundaries.

Table 5.2: Frequency distribution of the number of gaps for units with the same or fewer UMS categories than raw categories.

Number of gaps	GCSE units (untiered)		A level units	
	N	%	N	%
0	10	9.1	11	7.1
1	6	5.5	6	3.9
2	7	6.4	20	13
3	3	2.7	7	4.6
4	12	10.9	8	5.2
5	10	9.1	5	3.3
6	4	3.6	1	0.7
7	3	2.7	7	4.6
8	5	4.6	2	1.3
9	6	5.5	8	5.2
10	5	4.6	11	7.1
11	1	0.9	7	4.6
12	2	1.8	1	0.7
13	9	8.2	2	1.3
14	6	5.5	8	5.2
15	3	2.7		
16	5	4.6	9	5.8
17	3	2.7	8	5.2
18	5	4.6	11	7.1
19	1	0.9		
20	2	1.8		
21			6	3.9
23			4	2.6
24			5	3.3
25			4	2.6
26	1	0.9		
27	1	0.9		
30			1	0.7
31			1	0.7
42			1	0.7

Table 5.2 shows that there were plenty of gaps, even on units where in principle it was possible to have a 1-to-1 mapping of raw scores to UMS scores. The modal number of gaps was 4 for GCSE units and 2 for A level units. Some examples of the scaling functions for units with and without gaps are in Figure 5.1.

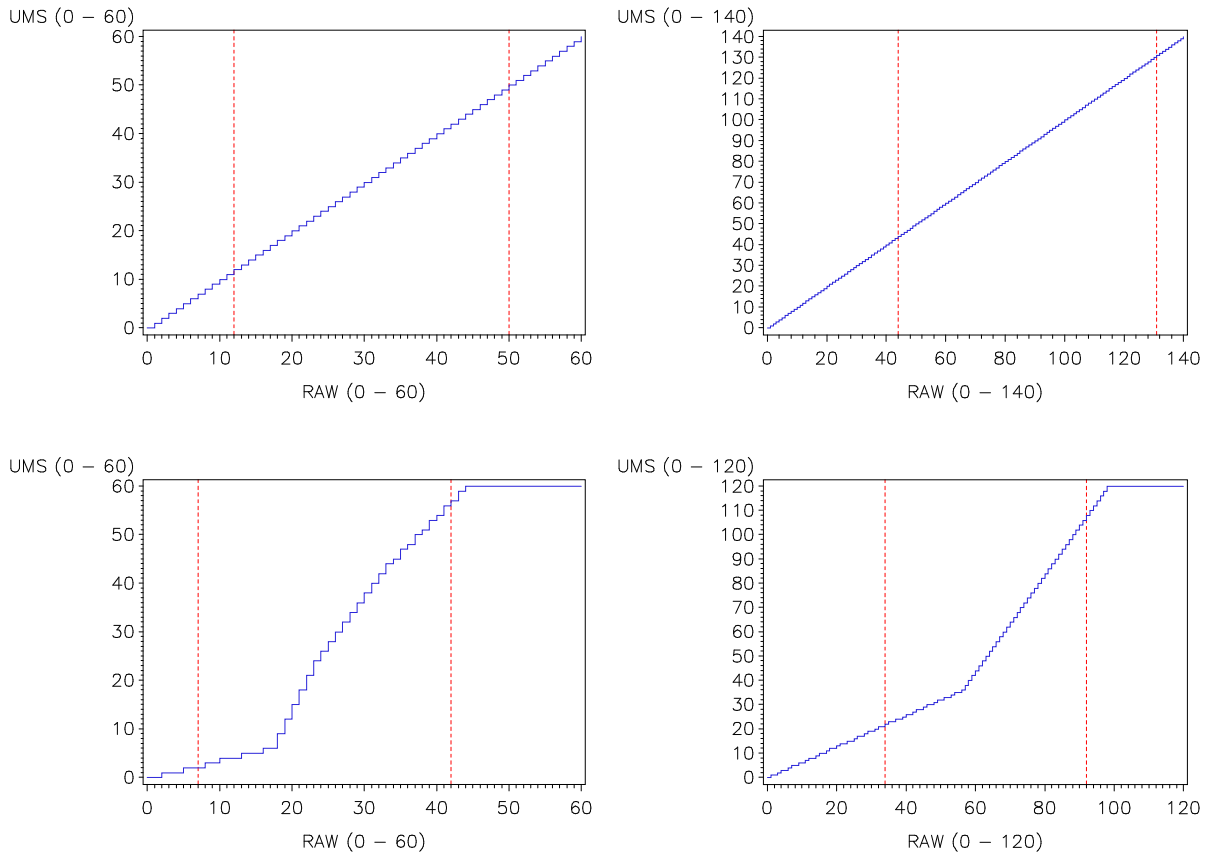

Figure 5.1. Top: a GCSE (left) and A level (right) unit with no gaps in the raw-UMS scaling. Bottom: the GCSE (left) and A level (right) units with the largest numbers of gaps[13].

The dashed red lines show the range of raw scores containing the middle 95% of examinees – that is, 2.5% scored less than the leftmost red line, and 2.5% scored more than the rightmost red line. The gaps in the UMS occur in the range of raw scores where the slope of the line is greater than one, and thus affect the majority of examinees.

As well as gaps, it is also possible to have UMS scores that are mapped to by more than one raw score (the converse can never happen). This will happen whenever there are more marks between raw boundaries than the corresponding UMS boundaries, but is perhaps most likely to happen when capping occurs. According to the rules defining the UMS conversion, when the raw score twice the A-B distance above the A boundary (the 'cap') is less than the maximum raw score, all raw scores including and above this point up to the maximum raw score achieve the maximum UMS score. The point of the cap is to ensure that there is not a sudden discontinuity in the conversion rate of raw to UMS marks just above the highest (A or A*) boundary, where a significant proportion of examinees might be affected – but the cost is the creation of a range of raw scores at the very top end where one more raw mark has no effect on scale score, potentially

[13] The actual units in Figure 5.1 were: top left – B234/02 Impact of modern technologies on manufacturing (from GCSE Manufacturing Double Award J510). Top right – F714 German: Listening, reading and writing (from A level German H476). Bottom left – B144 Consumer rights and responsibilities (from GCSE Law J485). Bottom right: G063 ICT systems, applications and implications (from A level ICT H517).

disadvantaging examinees when aggregating UMS scores across units. However, given that such examinees are first very rare and second almost certain to have achieved the top grade overall, this is not seen as a great problem. Nonetheless, a large range of capped raw scores does suggest a paper that was not well designed (too difficult).

Table 5.3: Frequency distribution of number of capped raw marks by type of unit.

Distance from the maximum raw mark to the cap	A level		GCSE untiered		GCSE Higher		GCSE Foundation	
	N	%	N	%	N	%	N	%
0	191	44.8	71	29.3	1	1.6	3	4.8
1	37	8.7	46	19.0	10	15.9		
2	20	4.7	17	7.0	2	3.2		
3	14	3.3	13	5.4	5	7.9	1	1.6
4	15	3.5	16	6.6	4	6.4	2	3.2
5	18	4.2	18	7.4	6	9.5	4	6.5
6	20	4.7	9	3.7	5	7.9	2	3.2
7	10	2.3	15	6.2	1	1.6	2	3.2
8	15	3.5	7	2.9	1	1.6	1	1.6
9	16	3.8	5	2.1	1	1.6	2	3.2
10	11	2.6	4	1.7	2	3.2		
11	5	1.2	3	1.2	1	1.6	2	3.2
12	7	1.6	6	2.5	2	3.2	1	1.6
13	5	1.2	2	0.8	3	4.8	1	1.6
14	6	1.4	2	0.8	1	1.6	6	9.7
15	7	1.6	2	0.8	4	6.4	2	3.2
16	8	1.9	4	1.7			2	3.2
17	4	0.9			3	4.8	2	3.2
18	4	0.9			3	4.8	2	3.2
19	1	0.2			2	3.2	2	3.2
20	5	1.2			1	1.6	3	4.8
21	2	0.5			2	3.2	1	1.6
22	1	0.2			1	1.6	3	4.8
23	1	0.2	2	0.8			3	4.8
24	3	0.7			1	1.6	1	1.6
25							3	4.8
26							1	1.6
27								
28							2	3.2
29							2	3.2
30					1	1.6	3	4.8
31								
32								
33								
34							1	1.6
35							1	1.6
36								
37							1	1.6

Distance from the maximum raw mark to the cap	A level		GCSE untiered		GCSE Higher		GCSE Foundation	
	N	%	N	%	N	%	N	%
38							3	4.8

Table 5.3 shows that the majority of units at all levels were capped, but that at A level the majority of units had relatively few capped marks. This was not the case at GCSE, however, where especially on the foundation tier (but elsewhere too) there were a significant number of units with a large number of capped marks.

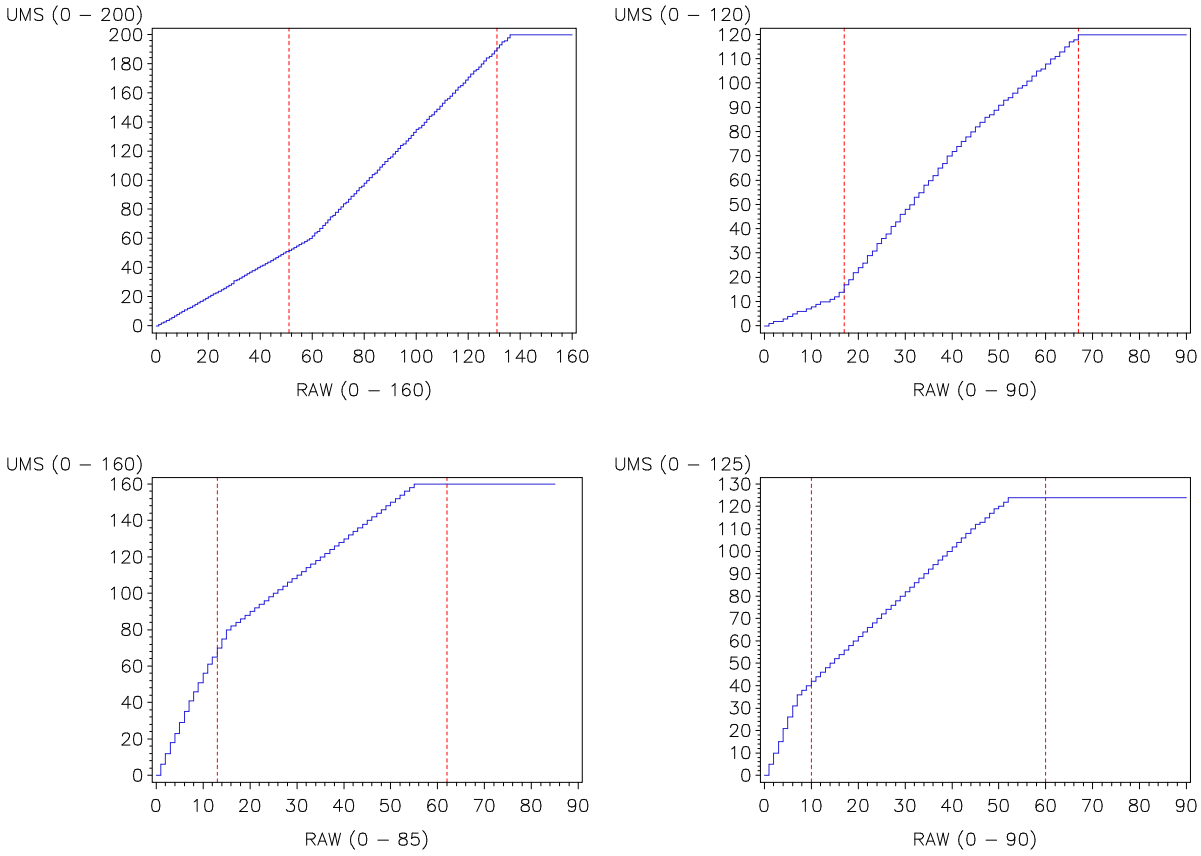

Figure 5.2. Examples of units with the largest numbers of capped marks.[14]

For the units in the lower part of Figure 5.2, more than 2.5% of examinees would have had their raw scores capped at the maximum UMS score, showing that capping is not necessarily a phenomenon affecting only a few examinees.

The above tables and graphs have shown that the UMS scaling function is very likely to create gaps (unobtainable UMS scores) and caps (ranges of raw scores that map to the maximum UMS score) at unit level for both GCSE and A level. The unit level UMS marks are reported to schools so examinees could in theory experience the scenario described above where a fellow student has received a UMS mark that was unavailable to them on the version of the unit that they took. The existence of capping also implies that it is in principle possible for an examinee to achieve a worse

[14] Top left: F887 – Portuguese: Listening, reading and writing 1 (from A level Portuguese H596). Top right: A265 – Businesses and their communications systems (from untiered GCSE Business and communication systems J230). Bottom left: B732/02 – Biology B: modules B4, B5, B6 (higher tier) (from GCSE Biology J263). Bottom right: A382/01 – Foundation Paper 2 (from GCSE Applications of Mathematics J925).

grade on the overall assessment than an examinee that has taken exactly the same units and obtained a lower aggregate raw score, although this can only arise when there is an extremely uneven pattern of performance across units. As far as we are aware, these potential sources of confusion and perceived injustice have never caused any particular problems with the 'users' of unit results. This is not to say that the possibility of gaps would be unproblematic if they were possible on the final reporting scale (i.e. the grades).

6. Summary and discussion

The purpose of a reporting scale is to aid users in interpreting test results. Section 2 of this report showed that if the users are concerned about repeatability of results (reliability) with tests containing different questions but testing the same construct, then at qualification level most A levels and GCSEs are probably reliable enough to support at least 10 grade categories along with the inference that an examinee's true grade lies approximately within ±1 grade of that obtained – in other words that it is unlikely that any grade is 'out' by more than one (the criterion suggested by Skurnik & Nuttal, 1969). At unit level this was less clearly the case – most untiered GCSE units reported more categories than the ±1 criterion would allow. Similar conclusions could be drawn from the slightly different (and less stringent) criterion of Mitchelmore (1981) that there should be a 90% probability of an examinee receiving the same grade on a parallel test. Two caveats should be applied to the optimistic conclusion about reliability at qualification level: first the estimates of measurement error did not explicitly incorporate marking error, although this is to some extent included in estimates of internal consistency reliability like Cronbach's Alpha provided that most marker error is unsystematic. Secondly, the calculations for both S&N and Mitchelmore assumed a constant error variance across the score range. It can be shown that errors around the true score are less at the extremes of the mark range than in the middle but in practice the bulk of examinees do not score near the extremes (especially on the aggregate scale of linear examinations), and the grade boundaries are also not usually located there, which may mean that the SEM is relatively constant over the operational range of raw scores.

If the scaling function maps a large number of raw score categories into a smaller number of scale score categories, then information is lost – the scale scores make fewer discriminations amongst the examinees than the original scores. Section 3 of this report illustrated what has previously been shown in the research literature: namely that it is not possible to compensate for measurement error by reporting in coarser categories because the rounding/grouping error adds to the measurement error in all but specially contrived cases. Our simulation based on a 169-category raw score scale and a reliability of 90% explored two statistical indicators of information retention: ρ^2 (the squared correlation between original and rounded/grouped scores) and *p_agree* (the proportion of pairs of examinees for whom the same inference about their relative achievement would be made based on the raw scores and the rounded/grouped scores). For both indicators, reporting on a 51-category scale lost almost none of the information. For scales of the length reported for GCSEs and A levels (i.e. around 7 to 10 categories) the ρ^2 statistic remained above 95% (as found by Shaw et al. 1987), but the *p_agree* statistic was somewhat lower at around 85%. This suggests that the coarser scales may be very useful for 'users' who want to carry out statistical analyses of examination data, but perhaps that a significant amount of information may be lost for those who want to make comparisons among a relatively small number of individuals, or individuals from a particular segment of the overall ability range. For all but the coarsest scales considered in our simulation (4 categories and 2 categories) the information lost to measurement error was greater than that lost to rounding/grouping error.

One argument for shorter scales is that it is easier for users to 'build up a picture' of what examinees in a given scale category know and can do. This can either be through explicit 'performance descriptors' or through experience. While not disagreeing with this, we suggested it might be also possible for users to attach meaning to scores reported on a percentage scale, which would have the advantage of retaining nearly all the information in the raw scores. We explored this in the Section 4 of this report. Of course, percentage scales are not comparable from one version of the exam to the next, but with a return to linear examinations at GCSE and A level it might be possible to use statistical equating techniques to link raw scores back to a reference raw score scale and thence to a percentage scale. This would allow scores to be interpreted in terms of the reference year's examination. It would add equating error to the measurement error and rounding / grouping error already considered, but this error is already present in any system that attempts to achieve comparability of reporting scales across different versions of examinations.

With statistical equating methods it is possible that for a given examination and a given number of scale score categories there may be some scale scores that are not mapped to by any raw score – we termed these 'gaps' in the reporting scale for obvious reasons. We equated actual (not simulated) data from the linear GCSE RS exam in the four years 2007-2010 to 2006, and found that reporting results on a percentage (101-category) scale did not create any gaps. However, increasing the number of categories to 111 created at least one gap in the 2008 scale, and further increases to 126 and 151 categories created gaps for nearly all years.

This raises the question of whether the existence of gaps is a problem for a proposed scaling function. One of Dorans (2002)'s criteria for a satisfactory scaling function was that there should not be more scale score points than raw score points, suggesting that it is a problem. We considered a hypothetical scenario where one applicant was beaten to a place on a course by another applicant with a scale score one higher than theirs that had not been obtainable on the version of the exam they had taken, and noted that if the equating is perfectly accurate then it is still likely that the applicant with the higher score is slightly better than the one with the lower score. Many have made the point that longer scales could encourage users to take small (i.e. substantively meaningless) differences in scale score too seriously in their decision making. Cresswell (1986) noted that if a coarser reporting scale is used, the loss of information from that specific test could have the beneficial effect of encouraging those taking selection decisions to rely on more varied sources of information (assuming that the other sources of information are valid).

A user's perspective on the 'gaps' problem may depend on how they conceptualise the assessment situation. If the raw score scale is seen as in some sense fundamental or primary then creating a reporting scale with more categories than raw categories may seem like attempting to create information from nothing. However, we would suggest that in many cases it is possible, fruitful even, to consider an examinee's level of achievement to be an abstract continuous quantity that the particular test is aiming to estimate, in other words that it is in some sense more fundamental than the raw score on a given test. If it is conceived as continuous, it is appropriate to use a very large (infinite!) number of scale categories. Thus tests that are calibrated with IRT methods and presented either as fixed length or adaptive tests may indeed often report the resulting estimates of examinee achievement on a scale with many more potential categories than raw score points available in the items administered. Even without making use of IRT, it is still possible to conceive of a 'universe' or 'domain' of test questions from which the selection appearing on any particular examination is a sample. The abstract quantity of interest to be estimated is the examinee's 'domain score' on a hypothetical test comprising all the questions in the domain. This hypothetical test would necessarily have an extremely long raw score scale. Alternatively, it could be considered to be the examinee's mean score on all questions in the

domain. In either case we can see beyond the particular raw score scale of a given test and conceive of a scale with many more categories. For example, even in the simple scenario of equating one 168-mark exam to another (see section 3), the pair of exams comprise one potential 336-mark examination. Once one raw score scale has been equated to the other, unless there are pairs of scores that coincide exactly under the equating function then 338 distinct values result. This can be seen by imagining interleaving Columns 1 and 4 of Table 4.1. From this perspective therefore there is nothing inherently wrong with having a reporting score scale with many more categories than are actually available in raw score terms on an individual test.

Whether long scales could ever be acceptable to the general public is debatable. Section 5 of the report considered the UMS scale and found that, at unit level, the majority of scaling functions create gaps in the UMS scores at both GCSE and A level. In some cases this is inevitable (when the maximum UMS score is greater than the maximum raw score); in others it is contingent on the location of the grade boundaries (when the difference between adjacent boundaries in raw marks is less than the corresponding difference in UMS score points). To our knowledge, there have never been any complaints from users about this aspect of the UMS scaling process, but we do recognise that UMS scores are mainly used as a means to an end – namely to obtain the final grade (on the coarse A*-U scale) on the aggregated units. It is still possible though that teachers and students use UMS scores to make decisions about resits, choices of other units etc. and in a few of these cases anomalous scaling functions may have an impact on overall outcomes.

One form of scaling that we have not considered in this report is scaling with the aim of stabilising (i.e. making constant) the standard error across the score range. For example, Kolen (1988) described an arcsine transformation that could stabilise error variance on tests for which a binomial or compound binomial model of error is appropriate (tests comprising dichotomous items). We did not explore that further here, partly because GCSE and A level units/components do not (mostly) comprise only dichotomous items, but mainly because the communication of error is a contentious and complex topic (see Zwick, Zapata-Rivera & Hegarty (2014) for some recent experiments). Embedding it into the reporting scale is not necessarily a better way of providing information about error than reporting it separately (for example as error bars around an examinee's score).

In conclusion we would say that in many respects, the current reporting scales for GCSE and A level do strike a reasonable balance in terms of preserving information and allowing for simplicity of communication. However, we question whether the 'meaning' of the grades in any subject can be clearly stated in terms of knowledge and skills implied, and thus believe there is still potential for exploring longer scales (such as a percentage of marks gained) that could both increase interpretability and preserve more information about relative performance. Reporting on such a scale would be greatly facilitated by the return to linear examinations and the reduction of choice among units and components, because then it would be possible to use statistical equating techniques to link aggregate raw score scales before applying the scaling function. One issue that would need to be resolved is how to deal with tiered examinations. It has been noted elsewhere (e.g. Bramley, 2017) that having overlapping grades (i.e. scale scores) available on tests of different difficulty creates interpretational problems even with the current system. This problem would not go away if longer scales were used. It may even appear to be exacerbated – but this could in fact have the beneficial effect of drawing attention to the issue which currently can be 'swept under the carpet' to a certain extent. The stranger-looking UMS scaling functions that can arise in tiered modular GCSEs (see Figure 5.2) are an example of what can happen behind the scenes in order to ensure that the overall reporting scale has (or appears to have) its desired properties of between-tier, year-on-year and between-board comparability.

References

Andersson, B., Branberg, K., & Wiberg, M. (2013). kequate: The kernel method of test equating. Retrieved from http://CRAN.R-project.org/package=kequate.

AQA. (2009). Uniform Marks in GCE, GCSE and Functional Skills Exams and Points in the Diploma. http://store.aqa.org.uk/over/stat_pdf/UNIFORMMARKS-LEAFLET.PDF. Accessed 28/04/14.

Baird, J-A., Hayes, M., Johnson, R., Johnson, S. & Lamprianou, J. (2013). *Marker effects and examination reliability: a comparative exploration from the perspectives of generalizability theory, Rasch modelling and multilevel modelling.* Report ref Ofqual/13/5261 Coventry: Ofqual. http://ofqual.gov.uk/documents/marker-effects-and-examination-reliability/ Accessed 02/06/14.

Benton, T. (2006). *Exploring the importance of graders in determining pupils' examination results using cross-classified multilevel modelling*. Paper presented at the European Conference on Educational Research, Geneva, September 2006.

Benton, T. (2014). Calculating the reliability of complex qualifications. *Research Matters: A Cambridge Assessment Publication, 18,* 48-52.

Bramley, T. (2010). A response to an article published in Educational Research's Special Issue on Assessment (June 2009). What can be inferred about classification accuracy from classification consistency? *Educational Research, 52*(3), 325-330.

Bramley, T. (2012). *What if the grade boundaries on all A level examinations were set at a fixed proportion of the total mark?* Paper presented at the Maintaining Examination Standards seminar, London.

Bramley, T. (2013). *The case for scale scores: reporting outcomes in the reformed GCSE.* Cambridge Assessment report.

Bramley, T. (2017). Some implications of choice of tiering model in GCSE mathematics for inferences about what students know and can do. *Research in Mathematics Education, 19(2),* 163-179. doi:10.1080/14794802.2017.1325775.

Bramley, T., & Dhawan, V. (2012). Estimates of reliability of qualifications. In D. Opposs & Q. He (Eds.), *Ofqual's Reliability Compendium* (pp. 217-319). Coventry: Office of Qualifications and Examinations Regulation.

Bramley, T., & Dhawan, V. (2013). Problems in estimating composite reliability of 'unitised' assessments. *Research Papers in Education, 28*(1), 43-56.

Bramley, T., & Vidal Rodeiro, C. L. (2014). *Using statistical equating for standard maintaining in GCSEs and A levels.* Cambridge Assessment Research Report. Cambridge, UK: Cambridge Assessment.

Cresswell, M. J. (1986). How many examination grades should there be? *British Educational Research Journal, 12*(1), 37-54.

Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: how and why. *Journal of Educational Measurement, 39*(1), 59-84.

Ebel, R. L. (1969). The relation of scale fineness to grade accuracy. *Journal of Educational Measurement, 6*(4), 217-221.

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

Goodman, L. A., & Kruskal, W. H. (Eds.). (1979). *Measures of association for cross-classifications.* New York: Springer-Verlag.

Hutchison, D., & Benton, T. (2012). Parallel universes and parallel measures: estimating the reliability of test results. In D. Opposs & Q. He (Eds.), *Ofqual's Reliability Compendium* (pp. 419-458). Coventry: Office of Qualifications and Examinations Regulation.

Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement, 25*(2), 97-110.

Mitchelmore, M. C. (1981). Reporting student achievement: how many grades? *British Journal of Educational Psychology, 51*(2), 218-227.

Ofqual. (2011). *GCSE, GCE, Principal Learning and Project Code of Practice.* Coventry: Ofqual.

Ofqual. (2012). *GCSEs and A levels in summer 2012: our approach to setting and maintaining standards.* Coventry: Ofqual. http://ofqual.gov.uk/documents/gcses-and-a-levels-in-summer-2012-our-approach-to-setting-and-maintaining-standards/ Accessed 29/05/14.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming and equating. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 221-261). Phoenix, Arizona: The Oryx Press.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Robinson, C. (2007). Awarding examination grades: current processes and their evolution. In P. E. Newton, J. Baird, H. Goldstein, H. Patrick & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards.* (pp. 97-123). London: Qualifications and Curriculum Authority.

Shaw, D. G., Huffman, M. D., & Haviland, M. G. (1987). Grouping continuous data in discrete intervals: information loss and recovery. *Journal of Educational Measurement, 24*(2), 167-173.

Skurnik, L. S., & Nuttall, D. L. (1968). Describing the reliability of examinations. *Journal of the Royal Statistical Society: Series D (The Statistician), 18*(2), 119-128.

Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment, 19*(2), 116-138.