CAMBRIDGE ASSESSMENT

Estimates of Reliability of Qualifications

Contract number 2907 (QCA/Ofqual and Cambridge Assessment)

20th October 2010

Ofqual/11/4826

Tom Bramley & Vikas Dhawan

ARD Research Division
Cambridge Assessment

Ofqual This report has been commissioned by the Office of Qualifications and Examinations Regulation

University of Cambridge
Local Examinations Syndicate

Acknowledgements

We would like to thank: the members of our supervisory group – Tim Oates, Sylvia Green, Neil Jones and Robert Coe – for their input and advice; Ofqual's Technical Advisory Group for its feedback on the draft version of this report; OCR for allowing access to data from their processing systems and databases; and the Operational Research Team in OCR for supplying computer programs and answering queries.

Table of contents

Preface

The purpose of this report is to contribute to Ofqual's Reliability Programme. It is one of a series of pieces of commissioned work covering various aspects relating to the reliability of examinations and other assessments. Although there is necessarily some discussion of conceptual and technical reliability-related issues, the main focus of this report is on the presentation of data and analysis from live high-stakes examinations taken in England, at GCSE and A Level, using data from the June 2009 examination session made available by the awarding body OCR.

Although we recognise the depth and complexity of many of the issues, we have aimed to keep the discussion as simple as possible, with the intention of arriving at a way for exam boards and other assessment agencies to present reliability-related information in a manner that is both transparent and informative.

Introduction

Test reliability is a topic with an extensive, and sometimes highly technical, research literature. It is not easy to summarise it for a lay or indeed professional audience without presenting any formulas or equations, although this feat was achieved (in the context of assessment in the UK) by Wilmut, Wood & Murphy (1996) in a review for one of Ofqual's predecessors - the School Curriculum and Assessment Authority (SCAA). A comprehensive, up-to-date technical survey of the field (from a US perspective) can be found in Haertel (2007), and recent conceptual and practical discussions of reliability in a UK context can be found in Hutchison (2008) and Newton (2009).

Aside from the difficulties of conceptualising and estimating reliability, the issue of communicating it to the public is also of prime importance. Newton (2005a, b) discussed the public and professional understanding of assessment error, calling for greater transparency and proactivity on the part of assessment agencies. One of the first pieces of work commissioned by the newly-formed Ofqual was a survey of the public's understanding of assessment inaccuracy (Ipsos MORI, 2009), which led to the suggestion of using the word 'variation' instead of 'error' (Boyle, Opposs & Kinsella, 2009) to avoid the negative and potentially misleading connotations of the latter when its everyday meaning is confused with its statistical meaning. However, the term 'error' is used in this report in conformance with its normal usage in the educational measurement literature.

Throughout this report we have considered solely what information and analysis could be reported from the data available, and how it could most effectively be reported. We have not considered what impact the availability of such information might have on examinees, schools, and examination boards, nor its effect on the general public's confidence in the examination system as a whole. These are of course crucial issues that are being considered elsewhere in Ofqual's Reliability Programme (e.g. He, Opposs & Boyle, 2010; Chamberlain, 2010).

The paragraph below shows how Ofqual has communicated its approach to reliability:

> "Reliability, in educational assessment terms, can be defined as consistency. A high level of reliability means that broadly the same outcomes would arise were an assessment to be replicated. Given the general parameters and controls that govern the assessment process (including test/exam specification, administration conditions, approach to marking, standard setting methodology and so on), reliability concerns the impact of the factors that inevitably vary from one assessment to the next. These include:
> - the particular **occasion** (e.g. if assessed on another day, the student might have been less tired)
> - the particular **test** (e.g. if a different test/exam had been set, the student might not have been confused by the wording of an essay title)
> - the particular **marker** (e.g. if a different marker had been assigned, the student might not have been marked down for using an unusual stylistic construction)
> - the particular **standard setting panel** (e.g. if a different team of people had been involved, different grade boundaries might have been set)."
>
> (Ofqual, 2009. Ofqual's Reliability of results programme: programme of work. Annex 1).

This report provides data and analysis for the second and third of the above points – i.e. the effect of variability in outcomes arising from different test questions, and from different markers. These form Sections 1 and 2 of this report. We did not attempt to answer the question of what boundaries would be set by a different standard setting panel, but in Section 3 we carried out what might be described as a 'sensitivity analysis', which investigated what the changes in the grade distributions would be at both unit/component level and at aggregate assessment level if the grade boundaries on the units/components were set one mark higher or one mark lower. If it is accepted that the setting of grade boundaries is not an exact science (by which we mean there is no algorithm that is guaranteed to generate the 'correct' answer, because a 'correct'

answer cannot necessarily even be defined) then slight differences from the actual decisions could presumably have arisen in other circumstances, and it is of interest to see what the effect of those slightly different decisions would have been.

We did not consider the effect of occasion-related variability, partly because the data needed for this to be evaluated has not been collected (re-sit data is not relevant because the assumption behind calculations of test-re-test reliability is that no genuine learning or forgetting has taken place, which one would hope is not true), and partly because most of the causes of any such variability (e.g. fluctuations in student tiredness or motivation) are not within the control of the assessment agencies (test developers, markers, grade boundary setters etc.)

The analyses in Sections 1 and 2 relied on data obtained from on-screen marking. This meant that the kinds of units/components not marked on screen (at the time of writing of this report) were not included. Such units/components tended to be either those requiring extended written responses, such as English Literature and History; or those that were not externally assessed, such as coursework, practical examinations, assessment of portfolios, art and design work, musical performances and language oral examinations. Of course, reliability issues are as important in such units/components as in any others, but the absence of relevant data meant we could not include them in this report. For a review of teacher or school-based assessment that includes discussion of its reliability, see Stanley et al. (2009).

All the data we analysed came from a single examination board (OCR). Although different examination boards offer different specifications with slightly different emphases and styles of assessment, the similarities between them outweigh the differences. This is because they are all designed to meet ‘common subject criteria’ in terms of content, assessment objectives, and assessment structure laid down by the Qualifications and Curriculum Development Agency (QCDA). We have no reason to believe that data from other boards in similar kinds of units/components to those we investigated would yield significantly different results from those reported here.

Each section of the report is relatively self-contained and ends with its own discussion. The final section of this report contains a short summary of the main findings and conclusions from each section and suggestions of directions for future research. The appendix contains a glossary of assessment terminology used in the report.

Section 1 – Test-related variability

1.1 Classical Test Theory

1.1.1 Definition of reliability – Classical Test Theory

The concept of reliability is essentially about the variation of measurement outcomes when the measurement procedure is replicated. It thus borrows from physical measurement the idea that when an attribute of an object (such as its length, or mass, or temperature) is repeatedly measured with a measuring instrument, there is likely to be a distribution of measurements obtained, rather than the same exact measurement being produced on each occasion. The 'scatter' in this distribution arises from a variety of random influences which in certain conditions results in a Gaussian (normal) distribution of 'errors'. If the measurement instrument is accurate (i.e. non-biased) the mean of this error distribution is, by definition, zero. The standard deviation of the error distribution, referred to as the Standard Error of Measurement (SEM), indicates the precision (or reliability) of measurement. The less the scatter (the lower the SEM), the more reliable the measurement instrument. Precision/reliability therefore refers to a distribution of measurements (Kendall & Buckland, 1957, cited in Stallings & Gillmore, 1971).

Classical Test Theory uses the same idea. Its central equation is:

$$X_i = \tau_i + E_i \qquad (1)$$

where X_i is the observed test score of person *i*, τ_i is the 'true score' and E_i is the error. The true score is a constant, defined to be the expected value (average) of the observed scores over a series of independent replications, and the error is defined to be the difference between the observed score and the true score on any particular occasion (Lord & Novick, 1968).

The replications can be conceived as repeatedly giving the same test to the same individual, with intermediate 'brainwashing' to make sure that each outcome is statistically independent of the others, or as repeatedly giving different 'strictly parallel'[1] versions of the test to the same individual. As Borsboom (2005) has pointed out, neither series of replications can actually be carried out in practice, which makes the connection with physical measurement rest on a thought-experiment.

In a group of test-takers the observed scores, true scores and errors can all be treated as random variables, giving:

$$X = T + E \qquad (2)$$

As a consequence of the initial definitions and assumptions of classical test theory it can be shown that in a group of test-takers the true scores and the errors are uncorrelated, giving:

$$\sigma^2_X = \sigma^2_T + \sigma^2_E \quad (3)$$

In words: the variance of observed scores is the sum of the true score variance and the error variance.

It can further be shown that the covariance between parallel tests is equal to the true-score variance, giving:

$$\rho_{XX'} = \sigma_{XX'} / \sigma_X\sigma_{X'} = \sigma^2_T/\sigma^2_X = 1 - \sigma^2_E/\sigma^2_X \qquad (4)$$

where $\rho_{XX'}$ is called the *reliability* of the test, equal to the correlation between two parallel tests, and equal to the proportion of observed score variance that comes from true score variance. The reliability is 1 if all the observed score variance is true score variance, and 0 if it is all

[1] For the properties that define strictly parallel tests, see Haertel (2007) p69.

random error variance (as might happen, for instance, if the scores on all questions were determined by tossing a coin).

The term σ^2_E, the error variance, is an *average* error variance across the group of test-takers. It is not an assumption of the theory that every individual has the same error variance across replications. The square root of this average error variance, σ_E, is taken to be the standard error of measurement (SEM).

Substituting (3) into (4) to eliminate σ^2_T and rearranging gives:

$$SEM = \sigma_E = \sigma_X \sqrt{(1 - \rho_{XX'})} \qquad (5)$$

Equations (4) and (5) make it clear that reliability and SEM are both defined in terms of a particular group of test-takers. Thus many authors are at pains to stress that the terms 'reliability' and 'SEM' do not apply to the test per se, but to the set of test scores obtained by a particular group of test-takers. If the average error variance remains constant, test scores from a group with a wider spread of true scores (i.e. a higher true score variance) will have a higher reliability than test scores from a group with a lower spread of true scores. Thus differences in reliability do not necessarily imply anything about the quality of the test (but may imply something about how well the test was 'targeted' at a particular group of test-takers).

1.1.2 Estimating reliability – Classical Test Theory

There are several different methods for estimating the unobservable reliability, $\rho_{XX'}$. A discussion of them is beyond the scope of this report. We focus on the most commonly used estimate of reliability, Cronbach's Alpha (Cronbach, 1951). The reason it is the most commonly used is that it does not actually require any repeated testing (!) and can be obtained from the item level data (the matrix of scores of each person on each test question) collected at a single administration of the test. It is sometimes described as an index of 'internal consistency reliability', since it is effectively the average of all possible split-half reliability estimates (Haertel, 2007). Different methods of estimating reliability differ in what is considered to contribute to the error variance. Internal consistency methods do not take into account any variability over time although this is not necessarily a problem since the test itself does not vary over time, and any error variance arising over time would presumably be attributable to either the thing being measured not being stable over time, or variation over time in measurement conditions, neither of which is in the control of the assessment agency.

The formula for Cronbach's Alpha is:

$$\alpha = \frac{n}{n-1}\left(\frac{\sum_{i \neq j} \sigma_{ij}}{\sigma_X^2}\right) \quad \text{or} \quad \alpha = \frac{n}{n-1}\left(\frac{\sigma_X^2 - \sum \sigma_i^2}{\sigma_X^2}\right) \qquad (6)$$

where σ_{ij} is the covariance of scores on items i and j, n is the number of items (test questions) and σ^2_i is the variance of scores on item i. Cronbach's Alpha can be applied to tests containing polytomously scored items (i.e. multiple-mark items) as well as dichotomously scored items (i.e. 1-mark items).

Thus in terms of numerical calculation, Cronbach's Alpha is the amount of inter-item covariance expressed as proportion of total test variance and scaled by a factor of $n/(n-1)$ (which allows its maximum theoretical value to reach 1). As the number of items increases, the contribution of inter-item covariance to the total variance generally increases more rapidly than the contribution of the item variances (providing the extra items do not have any low or negative covariances with other items). Hence Cronbach's alpha tends to be higher for tests with more items.

Some authors have lamented the widespread 'misuse' of Cronbach's Alpha (e.g. Schmitt, 1996; Sijtsma, 2009). First, it is not a measure of unidimensionality – that is, high values of alpha do not indicate that the items are all 'measuring the same thing'. Second, the derivation of alpha contains an assumption of 'essential τ-equivalence', which implies that all inter-item covariances are equal. This assumption is unrealistic. If it does not hold, alpha only estimates a lower bound for the reliability. There are other reliability estimators that give better (i.e. higher) lower bounds than alpha in these conditions (Sijtsma, 2009), but they are not widely used. Third, if a different assumption in the derivation of alpha does not hold (namely, that errors are uncorrelated across items, which might not be true if several items refer to the same source material or question stem), then alpha *over*estimates the reliability (Green & Yang, 2009).

Although these objections are relevant and should be borne in mind when interpreting reliability statistics, the fact remains that Cronbach's Alpha is the de facto standard in reporting of test internal consistency reliability in psychology, education and other fields. Its widespread use and ease of calculation make it and the associated Standard Error of Measurement the natural choice for a report such as this that covers a large number of tests.

1.1.3 Cronbach's Alpha and Standard Error of Measurement in GCSE and A-level units

Before 2004[2], exam papers (scripts) from GCSEs and A-levels were posted to markers (examiners) who would mark them at home and return the marked scripts to the examination board. Each marker would generally receive all the scripts from one or several schools (centres). In recent years, on-screen marking has become more widely used by all of the main GCSE and A-level awarding bodies. In this system, the scripts are scanned, and the digital images are downloaded and marked by markers connected to the internet, using proprietary software. All the item marks are recorded electronically, as well as the paper total (which is now obtained by electronically adding the item marks, eliminating human error at this point either in the adding or the keying). Furthermore, each marker receives a quasi-random allocation of scripts from all centres[3], ensuring that each marker sees a more representative selection of the entire examination cohort.

The existence of marks at item level has several implications for quality control and quality assurance (see for example Bell, Bramley, Claessen & Raikes, 2006). OCR produces reports for each on-screen marked examination that provide test developers with information about question difficulty, discrimination, differential item functioning (DIF) and reliability. We collated some of the information from these reports (Cronbach's Alpha and standard deviation of total scores) and used it to investigate test-related reliability in a Classical Test Theory framework.

Figures 1.1 and 1.2 and Table 1.1 on the next pages show the distribution of Cronbach's Alpha in the A-level and GCSE assessments for which we could obtain the relevant data. In Table 1.1 the GCSE data has also been further subdivided into: i) GCSE units (from 'unitised' GCSEs) and GCSE components (from 'linear' GCSEs; and ii) foundation and higher tier units or components.

Table 1.1 shows that the average value for Cronbach's Alpha was slightly above 0.8 for both AS and GCSE units/components. It is interesting to note that the un-tiered GCSE units/components had a higher average value for Cronbach's Alpha than the tiered units/components. This might be expected from the discussion above – presumably pupils of all abilities take these un-tiered units/components and thus the true score variance will be greater than if pupils from a more limited range of ability took the same units/components. On the other hand, a tiered test might be expected to be targeted more appropriately at its intended cohort, increasing information (see Section 1.2) so it is not necessarily the case that a tiered unit/component will have a lower value for Cronbach's Alpha.

[2] The exam board Edexcel was the first to use on-screen marking in GCSEs and A levels in 2004. It was first introduced by AQA and OCR in 2005 and 2006 respectively.

[3] Most examination centres are schools, but some are not, hence the more general term 'centres' is usually used by exam boards.

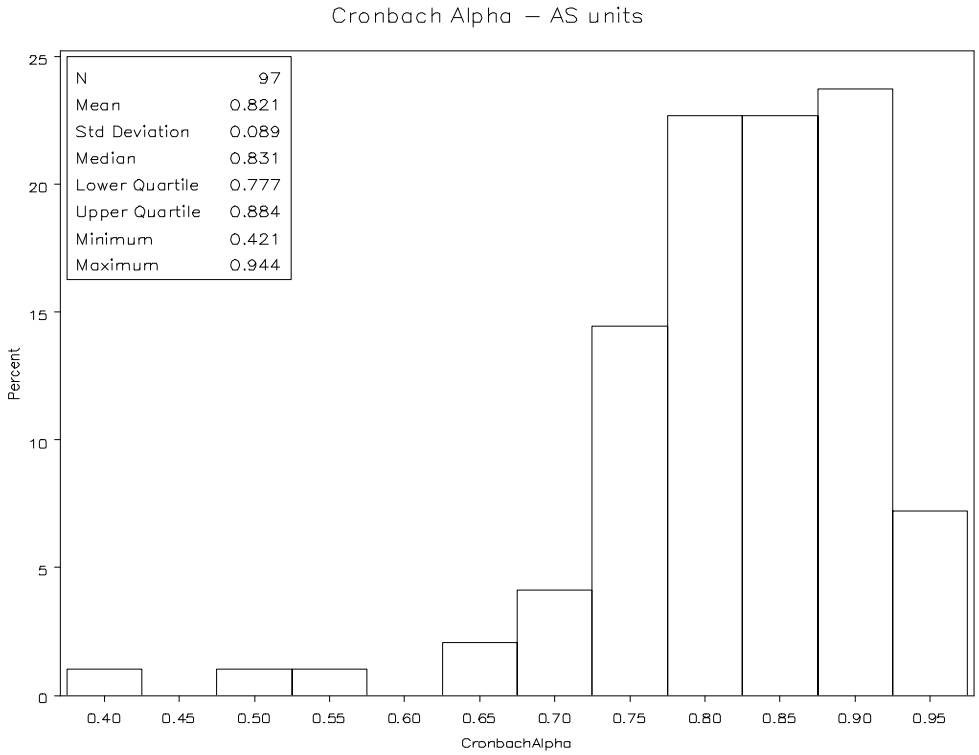

Figure 1.1: Distribution of Cronbach's Alpha in Advanced GCE units/components.

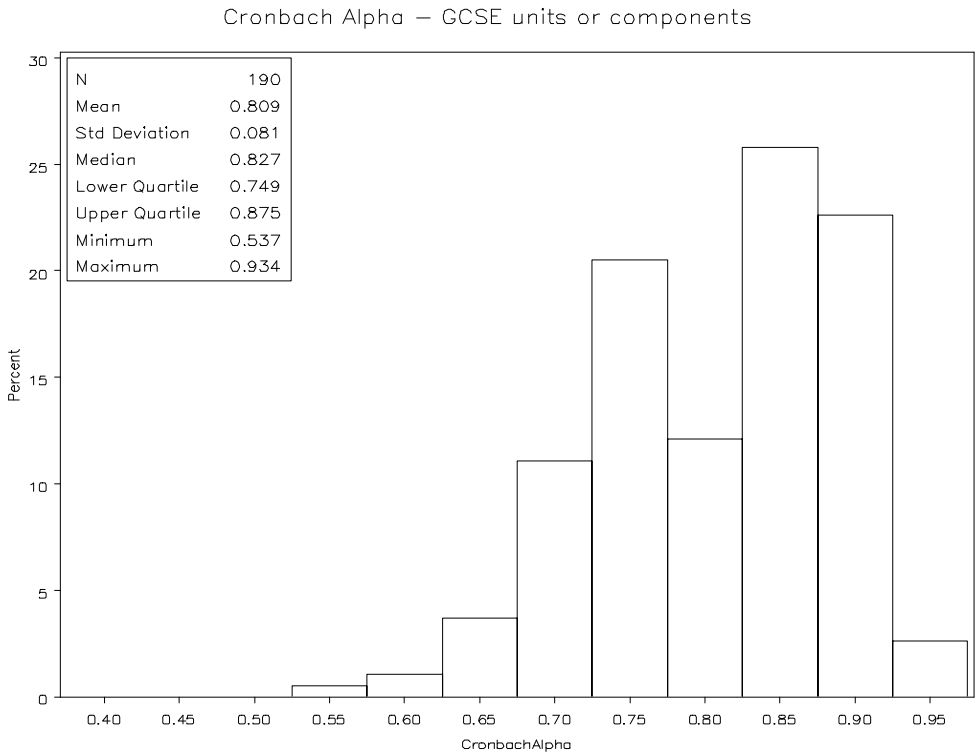

Figure 1.2: Distribution of Cronbach's Alpha in GCSE units/components.

Table 1.1: Distribution of Cronbach's Alpha.

Type	N	Mean	SD	Min	Lower quartile	Median	Upper quartile	Max
All	287	0.813	0.084	0.421	0.755	0.829	0.879	0.944
Advanced GCE units	97	0.821	0.089	0.421	0.777	0.831	0.884	0.944
GCSE units or components	190	0.809	0.081	0.537	0.749	0.827	0.875	0.934
Unitised GCSE units (all)	128	0.814	0.085	0.537	0.755	0.835	0.885	0.934
Linear GCSE components (all)	62	0.800	0.073	0.670	0.741	0.794	0.854	0.931
GCSE (foundation)	81	0.788	0.082	0.537	0.738	0.786	0.849	0.934
GCSE (higher)	79	0.807	0.082	0.585	0.745	0.813	0.886	0.932
GCSE (not tiered)	20	0.881	0.032	0.836	0.856	0.876	0.913	0.931

How should the figures in Table 1.1 be interpreted in absolute terms? In terms of classical test theory, they suggest that on average around 80% of the variability in test scores is true score variance, and 20% is random error variance. Cronbach's Alpha is used in many fields across the social sciences, and different criteria are used for evaluating reliability in different contexts. In educational testing, Frisbie (1988) suggested that the reliability coefficient should be at least 0.85 if the scores will be used to make decisions about individuals and if the scores are the only available useful information. Table 1.1 shows that the percentage of units/components meeting this criterion was between 25% and 50%. This might seem rather low, but it is important to bear in mind that these units and components were all part of larger assessments. The reliability of the assessment as a whole is perhaps of more relevance (in terms of decision-making about individuals) and thus the 0.85 guideline should probably be applied to the composite reliability (see Section 1.3) rather than that of the individual units/components.

A more serious drawback to interpreting Table 1.1 is that the units/components can and do vary along a number of dimensions – in particular the total number of marks (i.e. what the test was 'out of'), the number of items, the distribution of scores, and the weighting of the unit or component in the total assessment. This means that comparisons between different units/components are not necessarily comparisons of 'like with like'. Table 1.2 below shows the distribution of the SEM for the units and components for which we calculated Cronbach's Alpha.

Table 1.2: Distribution of SEM.

Type	N	Mean	SD	Min	Lower quartile	Median	Upper quartile	Max
All	287	4.19	1.57	1.53	3.10	3.74	5.00	11.35
Advanced GCE units	97	5.48	1.75	3.07	4.06	5.13	6.12	11.35
GCSE units or components	190	3.53	0.94	1.53	2.88	3.33	3.89	6.16

The figures in Table 1.2, however, are practically meaningless because the contextual factors of total number of marks etc. are even more essential for interpretation of SEM. The next section explores the relation of Cronbach's Alpha and SEM to test length.

1.1.4 The relation of Cronbach's Alpha and SEM to test length

According to classical test theory, a longer test will in general be more reliable than a shorter one. However, 'longer' could mean 'contains more items' or 'has a larger maximum mark'. If all the items are dichotomous, the two are the same. However, for GCSEs and A levels there is a lot of variability both within and between units/components in the mark tariffs for the items. This is illustrated in Figure 1.3 below[4]. It is clear that while there is a small general upward trend (i.e. papers with a higher maximum mark contain more items) there is a great deal of variability in the number of items for a given maximum mark. For instance, the number of items on units/components with a maximum mark of 100 ranges from 14 to 62. These differences in number of items will often be related to the nature of the subject and how it is assessed – some subjects are more often assessed by a larger number of low tariff items; others by a smaller number of high tariff items.

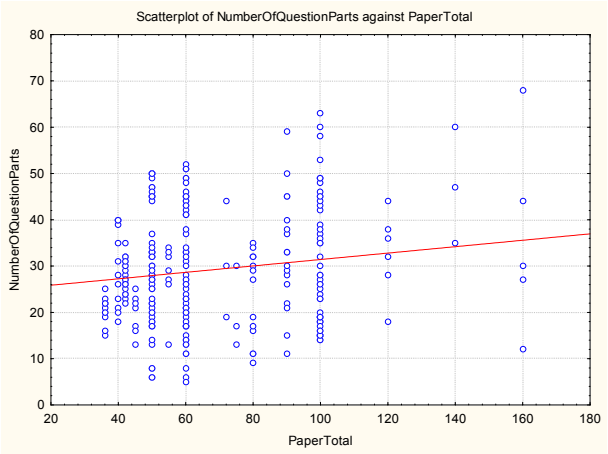

Figure 1.3: Plot of the number of items (question parts) against the maximum mark (paper total).

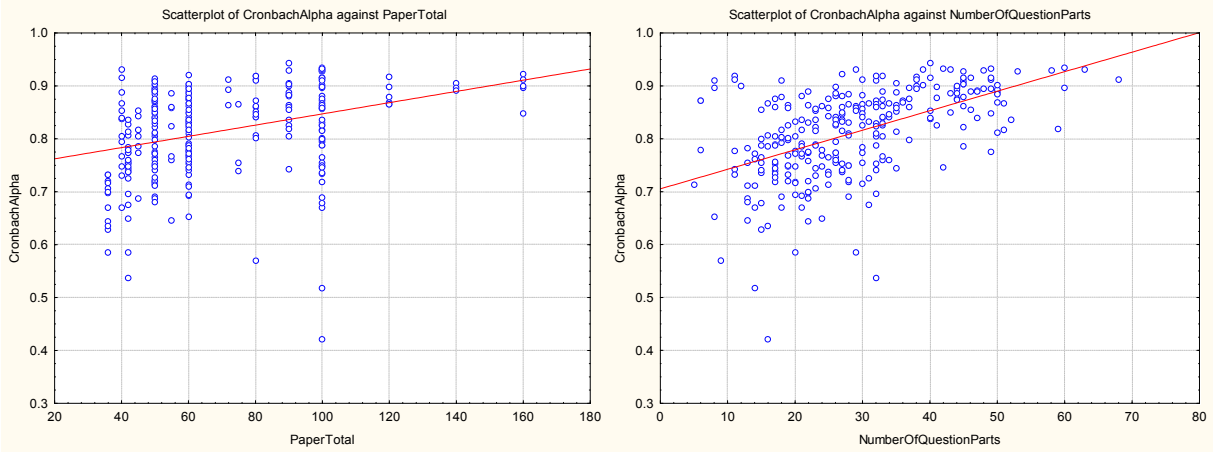

Figure 1.4: Plot of Cronbach's Alpha against the maximum mark (left) and the number of items (right).

Figure 1.4 shows a general increasing relationship between Cronbach's Alpha and both the maximum mark and the number of items, as might be expected. However, there is clearly plenty of variability among the units/components for any given maximum mark or number of items.

[4] In Figure 1.3 and similar figures, the best-fit regression line is drawn for ease of visual interpretation only.

Judging outliers by eye, there appears to be six units/components with unusually low values for Cronbach's Alpha (<0.6) given the maximum mark or number of items. They are shown below in Table 1.3.

Table 1.3. Units/components with low values of Cronbach's Alpha given maximum mark.

Type	Cronbach's Alpha	Paper Total	# of items
AS unit	0.421	100	16
AS unit	0.518	100	14
GCSE unit (foundation tier)	0.537	42	32
AS unit	0.567	80	9
GCSE unit (foundation tier)	0.585	42	29
GCSE unit (higher tier)	0.585	36	20

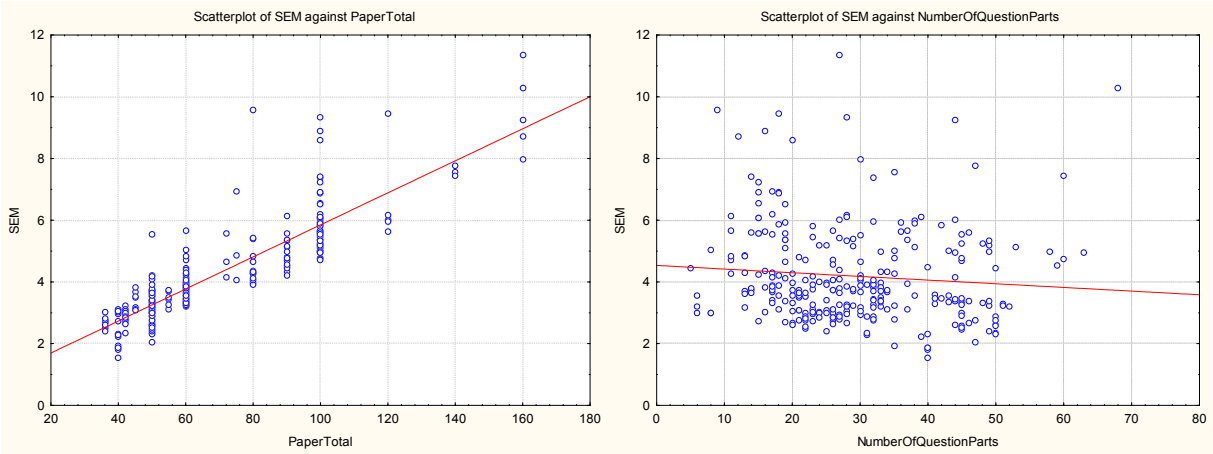

Figure 1.5: Plot of SEM against the maximum mark (left) and the number of items (right) for each of the 287 units/components.

Figure 1.5 shows that there is a very clear linear relationship between SEM and maximum mark, whilst there is almost no linear relationship between SEM and number of items. This implies that SEM on its own cannot properly be interpreted without extra information. Judging outliers by eye, there appears to be six units/components with unusually high values of SEM given the maximum mark. They are shown on the next page in Table 1.4. Only two of the units/components in Table 1.4 were also in Table 1.3, showing that different indices of test-related reliability will not necessarily 'flag up' the same units/components.

Table 1.4. Units/components with high values of SEM given maximum mark.

Type	SEM	Paper Total	# of items
AS unit	11.35	160	27
AS unit	9.56	80	9
AS unit	9.45	120	18
AS unit	9.32	100	28
AS unit	8.90	100	16
AS unit	8.61	100	20

1.1.5 The relation of SEM to the grade scale

We have argued in the above section that Cronbach's Alpha and SEM cannot be properly interpreted without extra information, particularly about the test maximum mark, but also about the number of items. A problem with making comparisons between different units/components is that each one has its own raw mark scale. The SEM has 'units' that are in raw marks – i.e. on the same local scale that is specific to each unit/component. In order to compare the different units/components in terms of SEM we would ideally need them to be on a common scale. In fact, GCSEs and A levels are reported on a common scale: the grade scale, which has the categories shown in Table 1.5 below.

Table 1.5: Grades reported at A level and GCSE.

A level		A	B	C	D	E	U		
GCSE	A*	A	B	C	D	E	F	G	U
GCSE foundation tier				C	D	E	F	G	U
GCSE higher tier	A*	A	B	C	D	E			

In general, each unit/component is graded individually. The aggregate grade for the whole assessment is determined in different ways for different assessment types. The grade outcome is what is reported to the examinee, and therefore it is desirable to find a way to relate reliability information to the grade scale. One natural way to do this is by comparing the size of the SEM with the size of a grade band (i.e. the number of marks on the raw scale that are allocated to each grade category). The grade boundaries are determined by a complex procedure that blends expert judgment with statistical information according to a code of practice produced by the regulator (e.g. Ofqual, 2009). For A level units/components, the expert panel sets the A and the E boundary, and the intermediate B,C and D boundaries are interpolated at equal intervals between them (with rounding rules that ensure that the higher grade bands contain more mark points where the division leaves a remainder). For GCSE units/components (in general[5]), the A, C and F boundaries are set by the expert panel, with the other boundaries being interpolated/extrapolated as necessary.

For the analysis below, we have taken the grade bandwidth to be the number of marks in the A-B range for A level and higher tier GCSE units/components, and the number of marks in the C-D range for lower tier GCSE units/components. We would expect units/components with higher maximum marks to have higher grade bandwidths. Figure 1.4 shows how the grade bandwidth varies with maximum mark.

[5] A few tiered GCSEs have a different structure to the one shown in Table 1.5 but they are a minority.

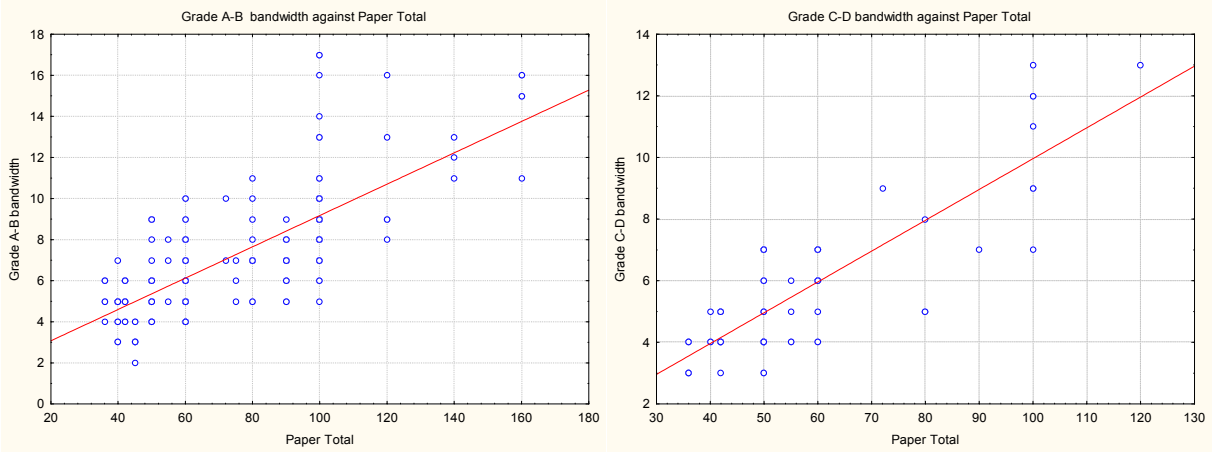

Figure 1.6: Plot of grade bandwidth against maximum mark (paper total).

The information in Figure 1.6 is interesting in its own right, even though it contains neither Cronbach's Alpha nor SEM. Returning to the analogy with physical measurement (on which classical test theory is based), consider measuring the height of a group of people and then defining points on the scale that allocated them to categories of (for example) 'very tall', 'tall', 'medium', 'short' and 'very short'. If the measurements were taken in inches then there would be a smaller 'bandwidth' than if the measurements were taken in centimetres. That is, the size of the bandwidth tells us about the precision that the measuring instrument purports to be measuring with. The bandwidth of units/components with a maximum mark of 100 ranges from 5 to 17. So the former unit/component distinguished between 5 levels of grade B examinee on its raw mark scale, whereas the latter unit/component distinguished between 17 levels of grade B examinee on its raw mark scale.

Combining the information about grade bandwidth with the SEM could thus perhaps allow more meaningful comparisons between different units/components. A unit/component with a bandwidth of 10 marks and a SEM of 5 marks is arguably equivalent in terms of reliability to a unit/component with a bandwidth of 6 marks and a SEM of 3 marks, because an examinee with a true score in the middle of a grade band would have the same likelihood of being misclassified (receiving a different grade from their 'true' grade) in both cases. For further discussion of classification consistency, see Section 1.4. The ratio of bandwidth to SEM could therefore be an appropriate index for comparing different units/components that use the same grade reporting scale.

Bandwidth:SEM ratio = Grade bandwidth (in marks) / SEM (in marks).

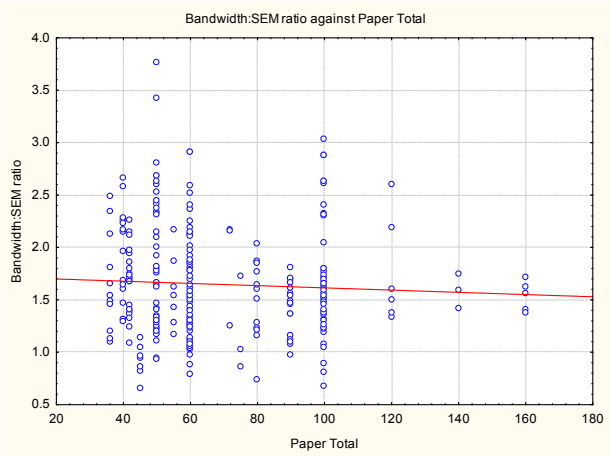

Figure 1.7: Plot of Bandwidth:SEM ratio against maximum mark.

Figure 1.7 shows that there is little or no linear relation of the bandwidth:SEM index and maximum mark. This suggests that the new index has successfully 'allowed for' differences in maximum mark among the different units/components. The six units/components with the lowest values for the bandwidth:SEM index are shown in Table 1.6 below. Three of the units/components in Table 1.6 were also in Table 1.3, and two of the units in Table 1.6 were also in Table 1.4, supporting the earlier comment that different indices of reliability will not necessarily 'flag up' the same units/components.

Table 1.6. Units/components with the lowest values for bandwidth:SEM ratio.

Type	Bandwidth:SEM	Paper Total	# of items
AS unit	0.652	45	25
AS unit	0.674	100	16
AS unit	0.732	80	9
AS unit	0.792	60	8
AS unit	0.810	100	14
AS unit	0.819	45	17

An alternative way of presenting the bandwidth:SEM index, which might be easier to communicate, would be to relate it to the probability that a person with a true score in the middle of a grade band might obtain an observed score in a different grade band. This is illustrated in Figure 1.8 below.

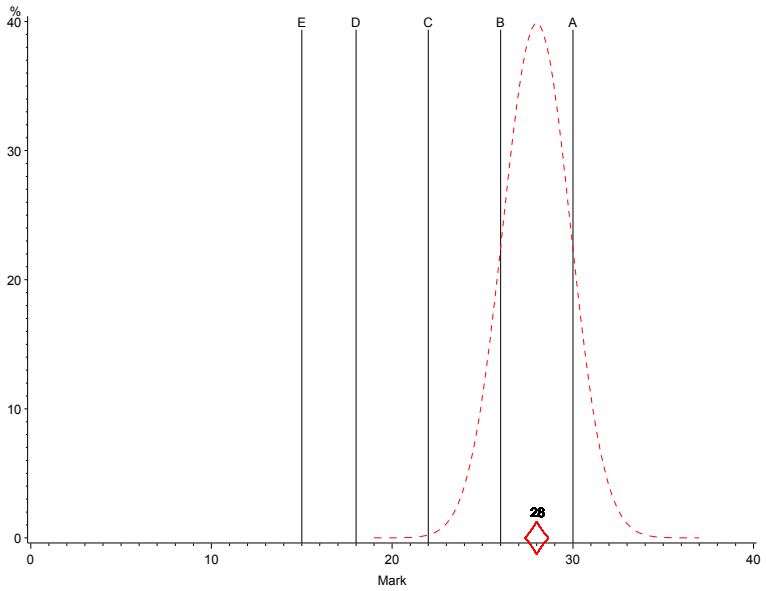

Figure 1.8: Plot of hypothetical measurement error distribution for an examinee with a true score at the middle of the grade B band.

The probability of obtaining an observed score outside the B band (i.e. obtaining a grade other than B) corresponds to the area under the curve that is outside the B band. In this case it is 0.28. Using this index ('prob_outside'), higher values correspond to lower reliability. The units/components identified as high or low on this index will be exactly the same as those identified using the bandwidth:SEM index. Figure 1.9 below plots this index for all the units/components. It is an upside-down version of Figure 1.7 with a non-linear transformation of the y-axis arising from the conversion of z-scores to probabilities from the normal distribution.

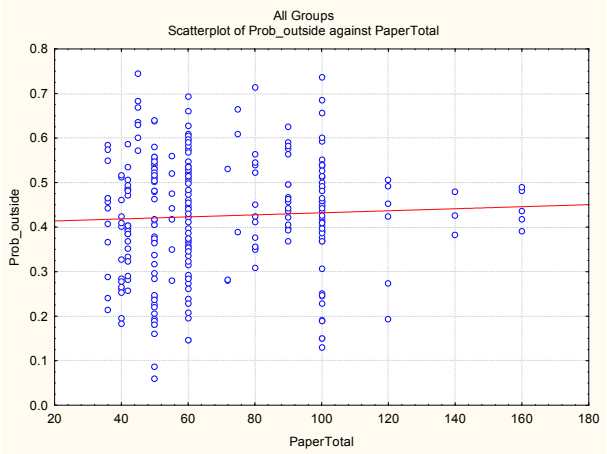

Figure 1.9: Plot of probability of observed score falling outside grade band for an examinee in the middle of the band.

1.1.6 The relation of reliability to unit/component weighting

Finally, in order to re-emphasise the point that all the reliability statistics calculated so far have related to units or components of larger assessments, we plot Cronbach's Alpha and the bandwidth:SEM index against the weighting that the unit/component had in the assessment of which it was a part. For the GCE units, we have taken the weightings in terms of the A level. They would in most cases be doubled if the overall assessment was taken to be the Advanced Subsidiary (AS level).

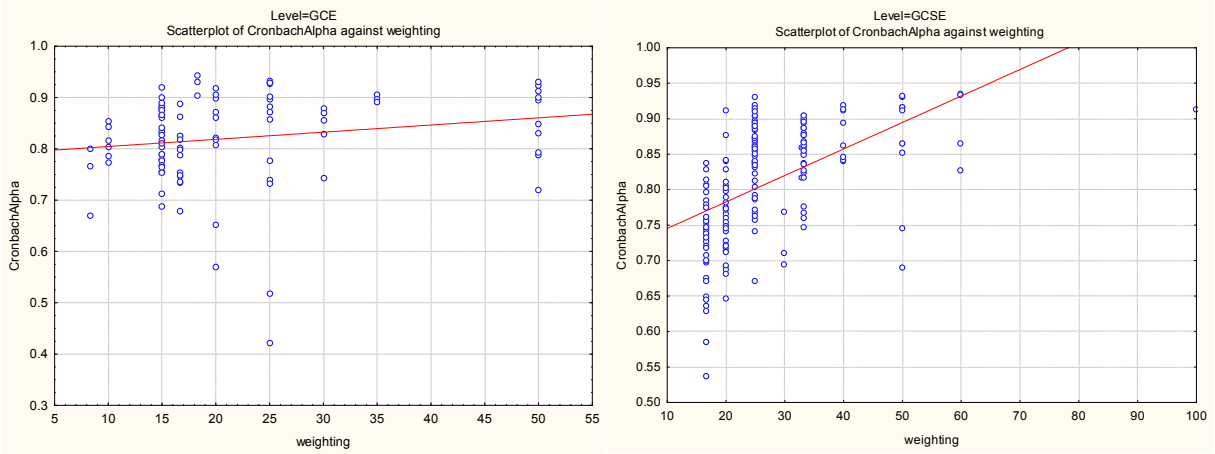

Figure 1.10: Plot of Cronbach's Alpha against unit/component weighting for GCE (left) and GCSE (right) units/components.

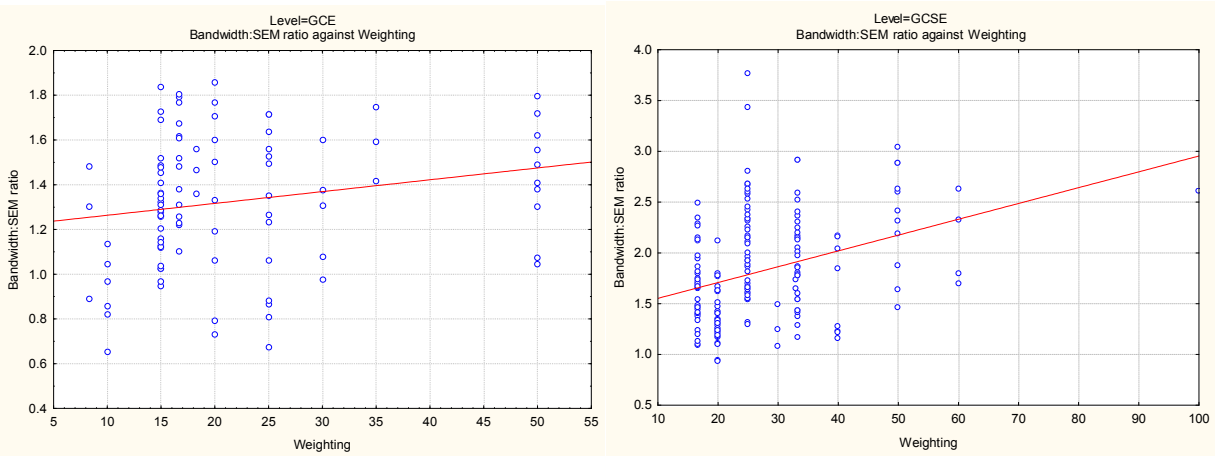

Figure 1.11: Plot of bandwidth:SEM ratio against unit/component weighting for GCE (left) and GCSE (right) units/components.

It is clear from Figures 1.10 and 1.11 that most GCE units/components were worth 35% or less of the whole (A level) assessment, and that most GCSE units/components were worth 50% or less of the whole assessment. The lowest absolute values of both Cronbach's Alpha and bandwidth:SEM ratio tended to be seen for the GCE units/components, which carried less overall weight.

1.2 Item Response Theory (IRT)

In Item Response Theory, as the name suggests, the item scores rather than the total scores are the focus of the modelling. The simplest model, the 1-parameter IRT model for dichotomous items, is also known as the Rasch model (Rasch, 1960).

In IRT models such as the Rasch model, the attribute or trait that one is trying to measure (e.g. knowledge, skill, attainment in subject X) is conceptualised as an infinite line marked out in equal-interval units, on which different objects (examinees and items in the case of educational testing), are located with respect to whether they have more or less of the measured attribute. Each examinee's location on the line is referred to as their 'ability', and each question's location as its 'difficulty'. In the Rasch model for dichotomous items, the estimated numerical values for ability and difficulty are distances in logits (log odds units) along the line relative to the local origin.[6] Lower values indicate less, and higher values indicate more of the attribute. The probability of getting an item right is modelled as a function of the difference between the examinee's 'ability' and the item's 'difficulty'. For more details about the Rasch model see, for example, Wright and Stone (1979), or Bond and Fox (2001).

The Rasch model has been extended to apply to polytomous items (those worth more than 1 mark) in what is known as the 'partial credit model' (Masters, 1982). In this model, each threshold between adjacent score categories has a separate parameter on each item. The equation for the Rasch Partial Credit Model is given below:

$$\pi_{nix} = \frac{\exp\left(x(\beta_n - \delta_i) - \sum_{k=1}^{x} \tau_{ik} \right)}{\sum_{x=0}^{m} \exp\left(x(\beta_n - \delta_i) - \sum_{k=1}^{x} \tau_{ik} \right)} \quad (7)$$

where π_{nix} is the probability of person n (with ability β_n) scoring in category x on item i (with difficulty δ_i) which has a total of $m+1$ categories separated by m thresholds (τ_{ik}). The sum of the thresholds on each item is zero (i.e. in this formulation of the model, the thresholds represent the deviation from the average item difficulty).

In 2- and 3- parameter IRT models for dichotomous items each item is characterised by further parameters representing 'discrimination' and 'guessing' (Birnbaum, 1968). These models have also been extended to apply to polytomous items, for example in the 'Generalised partial credit model' (Muraki, 1992). In all cases, the item and person parameters (and their associated standard errors) are estimated from the data by specialist software using one of several different estimation algorithms. The fit of data to model can then be evaluated by a wide variety of tests of fit, each of which is sensitive to particular aspects of deviation of the 'observed' data from the 'expected' data (i.e. the data that would be expected if the model fit well).

From one perspective, Rasch models are simply special cases of more general IRT models. From this 'modelling' perspective, the issue is to find the model that best fits a given data set. From a different 'measurement' perspective, Rasch models specify requirements that the data must meet in order to yield invariant (sample-free) estimates of relative difficulty and invariant (test-free) estimates of relative ability (Wright, 1977; Andrich, 1989). The well publicised disagreements between researchers holding different perspectives are not considered further in this report. We use the Rasch Partial Credit Model for the pragmatic reason that the ability estimates obtained from the Rasch model have a one-to-one relationship with the test total score. In other words, everyone with the same total score receives the same ability estimate. This is the basis on which GCSE and A level units/components are graded, so in effect the

[6] The arbitrary origin could be the location of a particular "reference item" or "reference person", or (more usually) the average location of all the items in a test.

assumption is made that the Rasch model is a reasonable approximation for these kind of examinations. The figures below illustrate the kind of information that can be obtained from a Rasch analysis. Figure 1.12 shows histograms of the estimated person abilities and item difficulties. Figure 1.13 plots the estimated ability corresponding to each possible raw score on the test. It can be seen that this is an s-shaped curve – that is, differences in raw scores correspond to greater differences in estimated ability at the extremes of the test than in the middle. However, the relationship is approximately linear over much of the raw score range.

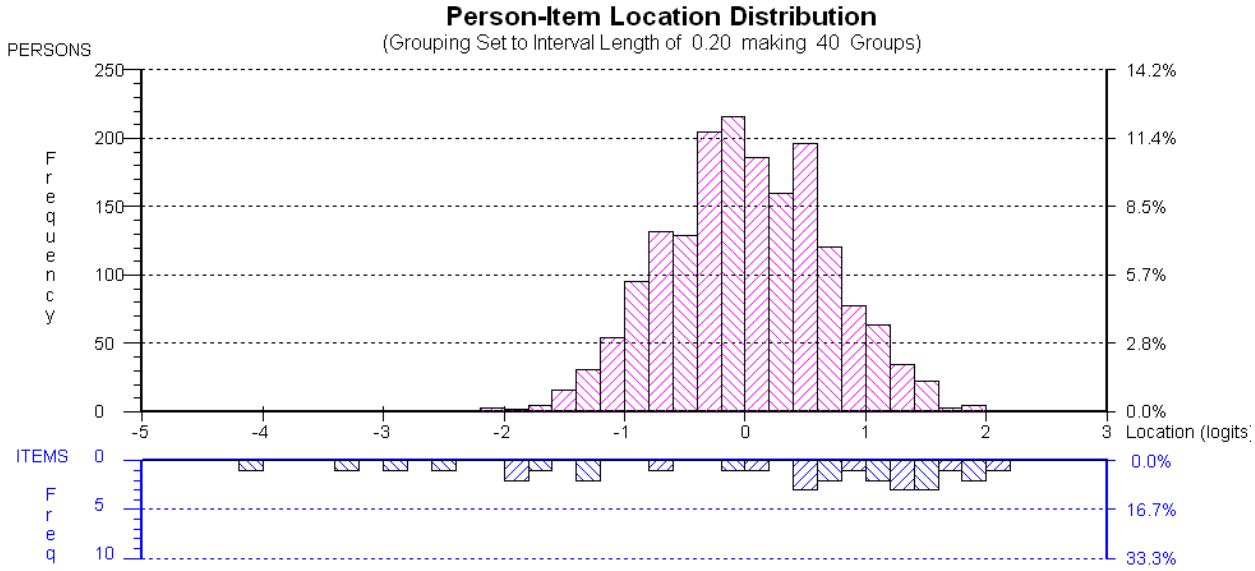

Figure 1.12: A GCSE Foundation tier paper, distribution of estimated item difficulties and person abilities.

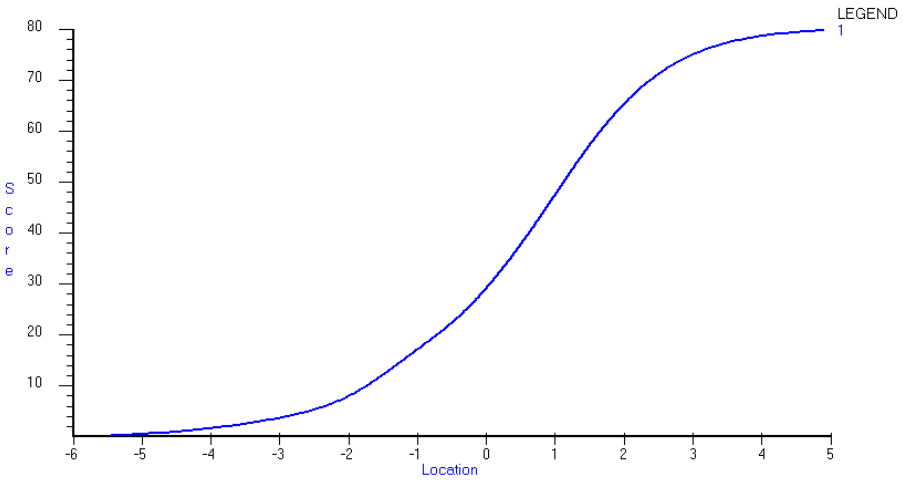

Figure 1.13: A GCSE Foundation tier paper, plot of raw score against estimated ability.

In other IRT models, it is possible for two examinees with the same total score to receive different ability estimates, depending on which items they have answered correctly. In fact, the focus in most IRT modelling is on the item parameters, and the person parameters are usually only represented by a distribution rather than estimated for each examinee (e.g. Thissen & Wainer, 2001). This is a second reason why it is easier to connect classical concepts of reliability to the Rasch model (e.g. Andrich, 1982).

1.2.1 Definition of reliability – IRT

The most natural index of precision in IRT modelling is the standard error of the person ability estimate. This is arrived at via the 'information function' $I(\beta)$, which is a concept from statistical estimation theory, the details of which are beyond the scope of this report[7]. Equation (8) below shows that the standard error is inversely proportional to the square root of the information:

$$SE(\beta) = \frac{1}{\sqrt{I(\beta)}} \qquad (8)$$

The total information conditional on ability, $I(\beta)$, is the sum of the information obtained from each item, on the usual assumption of local independence. For the dichotomous Rasch model, the modelled information in a single person-item response is:

$$I_i(\beta) = p_i(1 - p_i) \qquad (9)$$

where p_i is the modelled probability that a person with ability β will answer item i correctly (Wright & Stone, 1979). This means that information (precision) is greater, and the standard error is smaller, when:

- there are more items (because the total information is the sum of the individual item information, and item information is always positive[8]);
- the items are better targeted at the individual abilities. In the case of the Rasch model for dichotomous items, equation (9) shows that information is at a maximum when the probability of success is 0.5. (One major rationale for computerised adaptive testing is that by sequentially targeting items at estimated ability, fewer items are needed to obtain an ability estimate of equivalent precision to one obtained from a fixed-length test).

An important difference between classical test theory (as it is usually applied) and IRT is that the IRT information and standard error are conditional on ability (β) and therefore differ across the score scale. Precision is greater, and hence standard errors are smaller, in the middle of the score scale than at the extremes. It is not possible in IRT to estimate directly the ability corresponding to scores of zero or maximum marks because here there is 'infinite' uncertainty about ability – we know it is very high or very low, but not by how much. Most software can get round this constraint by adding in prior information or other assumptions. Figure 1.14 below shows the information function and the standard error.

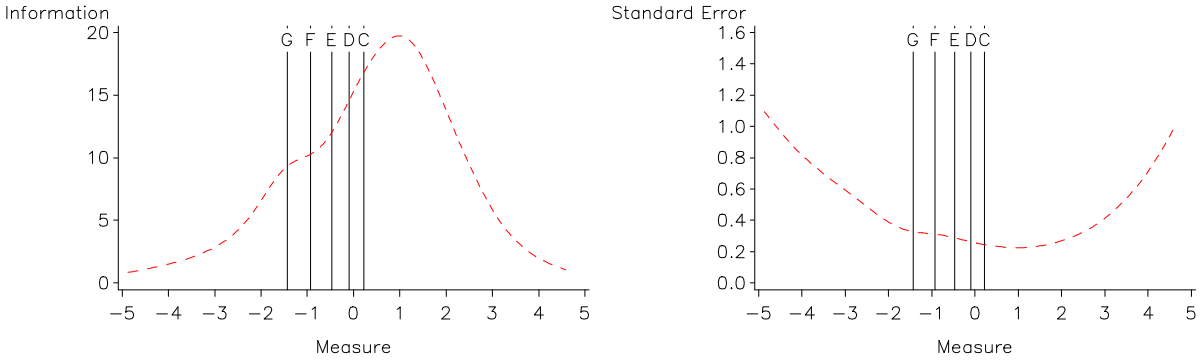

Figure 1.14: A GCSE Foundation tier paper, plot of information (left) and standard error (right) against estimated ability.

[7] There is a relatively accessible explanation on Wikipedia: http://en.wikipedia.org/wiki/Item_response_theory . For further details see Birnbaum (1968), or Yen & Fitzpatrick (2006). For a derivation of the information function for the generalized partial credit model, see Donoghue (1994).

[8] For the 1- and 2-parameter IRT models, but not necessarily the 3-parameter model (Bradlow, 1996).

Figure 1.14 shows that the test was measuring with greatest precision at a point on the scale slightly above the highest boundary (grade C). Figure 1.12 shows that there was a small cluster of items with difficulty values around this point on the scale. These two figures suggest that the test was not as well targeted as it could have been (i.e. it was too difficult). The ideal would be to have the test measuring with greatest precision in the region of the targeted grades – in this example G to C.

In contrast to classical test theory, where reliability is estimated first by Cronbach's Alpha (or some other estimate) and then an average SEM can be obtained from equation (5), in IRT the information function and conditional SEM for each level of estimated ability are obtained first as described above, and this can then be used to derive an alpha-like measure. For example, most Rasch software reports an index called the 'Person Separation Reliability' (R_β). This is calculated as (Wright & Masters, 1982, p106):

$$R_\beta = 1 - \frac{MSE_\beta}{SD_\beta^2} \qquad (10)$$

where MSE_β is the average (mean) of the squares of the person ability standard error estimates, and SD^2_β is the variance of the estimated abilities. The formula in (10) can be seen to be analogous to that in equation (4).

One main difference between classical test theory and IRT is that in practice the latter is much better at dealing with 'missing data' – i.e. situations where not every examinee has attempted the same items. When the data fit the model, an examinee's ability can be estimated from any set of appropriately calibrated items. This is the principle behind tailored testing, item banking and computerised adaptive testing, and is the main practical reason for going to the trouble of using IRT. A value for R_β can be derived in situations where there is missing data, which is not the case for Cronbach's Alpha, which requires all examinees to have attempted the same set of items. This distinction is not relevant for the purposes of this report, because all 287 units/components for which we calculated Cronbach's Alpha consisted entirely of compulsory questions.

1.2.2 Separation reliability index for twelve GCE/GCSE units/components

Table 1.7 on the next page shows the values of R_β obtained from analysing the data from twelve units/components with the Rasch Partial Credit Model using the software RUMM2020[9]. The units/components were chosen to reflect a balance of subject types and paper totals, and to fit with the work on composite reliability (Section 1.3). The values of Cronbach's Alpha estimated from the same data are given for comparison.

[9] There is a very slight difference from equation (8) in how RUMM2020 calculates the person separation reliability. The average of the standard errors of the person ability estimates is calculated by dividing the sum by N-1 rather than N, where N is the number of persons. See Rumm Laboratory (2004) page 7.

Table 1.7: Rasch separation reliability indices for twelve GCE/GCSE units/components.

Type	Part of	No. of examinees	Cronbach's Alpha	R_β	Paper Total	# of items
AS unit	A 2/4-unit GCE	4355	0.652	0.737	60	8
AS unit		4352	0.750	0.815	90	11
AS unit	A 3/6-unit GCE	625	0.813	0.861	60	27
AS unit		625	0.827	0.841	60	35
AS unit comp.		625	0.862	0.856	45	23
GCSE comp.	Foundation tier	1761	0.837	0.848	80	30
GCSE comp.		1758	0.839	0.877	80	34
GCSE comp.	Higher tier	3081	0.857	0.877	80	29
GCSE comp		3082	0.841	0.883	80	27
GCSE unit	Foundation tier	26233	0.930	0.940	100	58
GCSE unit	Higher tier	31629	0.915	0.921	100	46
AS unit	A 2/4-unit GCE	689	0.930	0.956	90	29

Comp. = Component.

It seems from Table 1.7 that in general the Rasch separation reliability R_β is slightly higher than the corresponding Cronbach's Alpha, but the relative ordering of the two sets of results is very similar.

1.3 Composite reliability

1.3.1 Background

As emphasised in Section 1.1, the units/components for which Cronbach's Alpha and SEM were estimated were only part of larger assessments. The GCSE units/components were part of a whole GCSE, and the GCE units/components were part of both AS and A level. Whilst knowing the reliability of individual units/components may be useful to the assessment agencies in their quality control processes, from the point of view of the examinee it is perhaps reliability at the level of the whole assessment that is more important.

Later in this section we present indices of composite reliability for a few selected assessments and discuss their interpretation. Before doing that, however, it is necessary to describe in some detail the structure of a typical GCSE and A level, and to explain how the units/components are aggregated and graded, in order to highlight the complexities involved. These complexities reside not so much in the reliability calculations as in defining what the 'composite' might mean in the first place.

In June 2009, the examination session from which most of the data used in this report was obtained, both A levels and GCSEs were in a period of transition in terms of assessment structure. Previously, A levels had in general consisted of six units: three AS and three A2. The majority of these were in the process of moving to a 4-unit structure – i.e. two AS and two A2. In June 2009 some examinees were taking AS units in the old 6-unit structure and some were taking AS units in the new 4-unit structure. The A2 units in the 4-unit structure were not available until January 2010. For GCSEs, the transition was from linear specifications to unitised specifications. The June 2009 data available to us contained examples of both the old and the new at both GCSE and A level. Figure 1.15 below is taken from OCR administrative documentation, available to schools and published on the website[10] and shows the structure of the AS assessment in Chemistry, on the 3-6 unit structure.

◇ CHEMISTRY		Availability		Max. Uniform Mark
Entry Codes and Rules of Combination		**January 2009 1A09**	**June 2009 6B09**	
3882	**AS GCE Chemistry (Certification)**	**!**	**!**	**300**
	For a certificate candidates must have taken the following **two** mandatory units:			
2811	Foundation Chemistry	Q	Q	90
2812	Chains and Rings	Q	Q	90
	and one from:			
	How Far, How Fast? / Experimental Skills 1			
2813A	with Coursework			
	01 Written Paper	Q	Q	120
	02 Coursework	P	P	
2813B	with Carried Forward Coursework Mark			
	01 Written Paper	Q	Q	120
	82 Carried Forward Coursework Mark	C	C	
2813C *	with Practical Examination			
	01 Written Paper	Q	Q	120
	03 Practical Examination	Q	Q	

Figure 1.15: AS Chemistry assessment structure. Extract from OCR Entry Codes 14-19 Qualifications (2009). Note: Q=Written Question Paper, P=Postal moderation of coursework, C=Carry Forward Component.

[10] Entry codes: 14-19 Qualifications 2009/10 available at http://www.ocr.org.uk/administration/documents/general.html . This report used the 2008/9 equivalent.

The units could be entered in any session where they were available (subject to any restrictions in the specification). Because we did not have any item level data from the A2 units, we focussed on the AS assessment (Chemistry 3882). Even with this simplification, there were still many different possible routes to the composite (aggregate) assessment. Typical examinees[11] could have taken the units 2811, 2812 and 2813 at any of the following sessions: January 2008, June 2008, January 2009 or June 2009. Few examinees took all three units in June 2009. The data from a given unit in any one session may include re-sit examinees who took the unit in a previous session (the question paper is different in each session). See Section 3 of this report (Tables 3.17 and 3.18) for more details on the numbers of examinees taking different combinations of units, and the numbers re-sitting.

A further complication arises in unit 2813, which comprised two components: a written paper (component 01) plus either i) coursework (component 02); or ii) carried forward coursework mark (component 82) – i.e. coursework that had been entered and marked in a previous session, which the examinee was re-using, presumably in the hope of getting a better mark on the written component (01); or iii) a practical examination (component 03). Unit 2813 in itself therefore has several 'composite reliabilities', even within a session, depending on which of the optional components (02, 82, or 03) the examinees took. A further complication is that of weighting. In unit 2813, component 01 (the written paper) had a maximum raw mark of 45 but a maximum weighted mark of 60. The weighted marks on component 01 (i.e. weighted by multiplying by 60/45) were combined with the unweighted marks on either 02, 82 or 03 to give a total mark for unit 2813 out of 120. Units 2811 and 2812 each had a maximum raw mark of 60. The final complication is that of aggregation. In unitised examinations (as opposed to linear examinations) the raw (or weighted) marks from the units are not added together. This is for the obvious reason that if examinees are entering different units in different sessions, the total raw (or weighted) mark would be meaningless, unless it somehow happened that each unit was of exactly the same difficulty at all points in the raw mark scale as the same unit in other sessions. To get round this problem, in unitised examinations the Uniform Mark Scale (UMS) is used. Each unit in each session is graded separately, and the raw (or weighted) marks are transformed to the uniform scale. This scale is essentially a more fine-grained numerical form of the grade scale with fixed boundaries corresponding to the different grades (For GCE units this is 80% for A, 70% for B etc. down to 40% for E). The number of UMS points available for a particular unit reflects the weight of that unit in the overall assessment, as set out in the specification. In the example we have been using of Chemistry 3882, we see from Figure 1.15 that units 2811 and 2812 each carried 90 UMS points, but 2813 carried 120 UMS points.

Raw (or weighted) marks at the level of the unit are converted to UMS points by a piecewise linear transformation. The marks corresponding to the grade A and grade E boundaries are 'mapped' to the corresponding marks on the UMS scale and via these two points a linear transformation is defined which is applied between the A and E boundaries, and also is extrapolated a certain distance beyond them. If UMS points are plotted against the raw marks in this region (as in Figure 1.16 later), this defines the 'main line'. The remaining points on the raw scale are mapped to the UMS scale by (at the bottom end) a linear transformation linking zero on both raw and UMS scales with the bottom of the main line, and (at the top end) a linear transformation linking maximum marks on both raw and UMS scales with the top of the main line. If the extrapolation of the main line above the A boundary reaches the maximum UMS before it reaches the maximum raw mark, then the raw score at which this happens is known as the 'cap', and all raw scores above this point are mapped to the maximum UMS score. For more details, consult either (AQA, 2009; Gray & Shaw, 2009).

It is the UMS points at unit level that are aggregated to form an overall UMS score, and it is this that determines the overall grade of the examinee. Table 1.8 below shows the raw and UMS grade boundaries for the 3882 assessment. The UMS boundaries are fixed, but the raw

[11] i.e. those studying AS / A level Chemistry in Year 12 & 13.

boundaries vary across sessions[12]. The raw boundaries in Table 1.8 came from the following units:
2811 June 2009
2812 January 2009
2813 June 2009 (Option B: Written Paper 01 + Coursework carried forward 82).

Table 1.8: Raw and UMS grade boundaries for AS Chemistry 3882.

	Max	A	B	C	D	E
2811 Raw (June 2009)	60	49	44	39	34	29
2811 UMS	90	72	63	54	45	36
2812 Raw (Jan 2009)	60	49	44	39	34	29
2812 UMS	90	72	63	54	45	36
2813 Raw (June 2009)	120	97	87	77	67	57
2813 UMS	120	96	84	72	60	48
3882 UMS	300	240	210	180	150	120

The above discussion should have made clear that it is far from straightforward to say what is meant by 'the' composite reliability of a unitised assessment. The structure of the full A-level assessment for Chemistry was even more complex, with one of the A2 units containing five different options (each with two components). Nevertheless, in the next section we attempt to calculate a value for the composite reliability of AS Chemistry (3882) and two other assessments.

1.3.2 Composite reliability formula (classical test theory)

The formula we used for calculating composite reliability is given below.

$$r_c = 1 - \left(\frac{\sum w_j^{\,2} \sigma_j^{\,2} - \sum w_j^{\,2} \sigma_j^{\,2} r_{jj'}}{\sum w_j^{\,2} \sigma_j^{\,2} + 2 \sum w_j \sigma_j w_k \sigma_k r_{jk}} \right) \qquad (11)$$

where:
r_c = reliability (internal consistency) of the composite assessment (Composite Alpha)
w_j = weight of element j in the composite total
σ_j^2 = variance (SD^2) of scores on element j
$r_{jj'}$ = reliability of element j (Cronbach's Alpha)
r_{jk} = correlation of element j with element k, where $j \neq k$.

Equation (11) is based on equation (26) in Haertel, 2007, with the denominator expanded to show how the total composite variance was calculated. It is the same as the formula for 'stratified alpha' (Haertel, 2007, equation 29), except that the 'strata' are the different units/components rather than items testing different content domains within the same test. We subsequently refer to r_c as 'Composite Alpha'. Equation (11) is of same form as equation (4), the key assumption being that errors of measurement are uncorrelated across elements of the composite. The numerator of the term in brackets is the sum of the error variances of the elements of the composite, and the denominator is the total variance of the composite. The SEM is therefore simply the square root of the numerator.

[12] It is a coincidence that the raw boundaries of 2811 in June 2009 were the same as the raw boundaries of unit 2812 in Jan 2009.

1.3.3 Composite Alpha of AS Chemistry 3882

In order to calculate Composite Alpha it is necessary to use results from the same examinees on all elements of the composite. In the case of AS Chemistry, because of the variety of possible routes, we selected the combination with the largest number of matchable examinees. This was the combination shown above in Table 1.8, which had 625 examinees. Calculating Composite Alpha was done in two stages: first by calculating Composite Alpha for Unit 2813 option B (combining the written component 01 with the coursework carried forward component 82), and then by calculating Composite Alpha for the AS assessment 3882, using the reliability for Unit 2813 calculated in the first stage.

Table 1.9: Descriptive statistics for the components of Unit 2813.

Component	Max raw	Max weighted	Weight	No. of examinees	Mean	SD	Cronbach's Alpha	Correlation
2813B 01	45	60	1.33	625	27.76	8.77	0.862	0.249
2813B 82	60	60	1.00	625	45.25	5.99	0.500**	

** arbitrary estimate.

Because we were not able to obtain item level data for the coursework component (which was not marked on-screen) we had no means of calculating a value for Cronbach's Alpha for this component. We therefore used a value of 0.5 as a 'worst case' value. Using the statistics in Table 1.9 the value for Composite Alpha was calculated to be 0.823. (Changing the estimate of Cronbach's Alpha for the coursework component 82 from 0.5 to 0.7 changed the value of Composite Alpha to 0.857).

Table 1.10 descriptive statistics for the elements of AS Chemistry 3882.

Unit	Max raw	Max UMS*	Weight	No. of examinees	Mean	SD	Cronbach's Alpha	Correl. with 2812	Correl. with 2813
2811	60	90	1.5	625	41.13	10.44	0.813	0.645	0.741
2812	60	90	1.5	625	39.41	9.74	0.827		0.655
2813	120	120	1.0	625	82.26	14.41	0.823		

*these values are taken from Figure 1.15.

Using the statistics in Table 1.10 the value for Composite Alpha for the AS assessment was calculated to be 0.924 (with a value for SEM of 10.93). This value only increased to 0.929 when the value of Cronbach's Alpha for component 82 in Table 1.9 was taken to be 0.7, and only decreased to 0.912 when it was taken to be 0! This shows that the coursework component 82 did not make a large contribution to Composite Alpha for the whole assessment. This is partly because of its relatively low specified contribution to the assessment (20%), but also because of its low correlation of 0.25 with the written component 01. However, this does not imply that coursework marks gained were of no use to the examinee, nor does it imply that the coursework did not contribute to the overall validity of the assessment, if the latter is defined as its correlation with some hypothetical criterion variable (Rudner, 2001).

The Composite Alpha calculated above refers to the composite total derived from aggregating the weighted raw marks on the three units, based on the 'intended weights' implied by the UMS points available. It does not take into account the UMS transformation that was actually used in practice when aggregating these units. Figure 1.16 below shows the UMS conversion for unit 2811 based on the information in Table 1.8 above. The red line shows the actual conversion used. The blue line shows the 'ideal' conversion – ideal in the sense that the same conversion of raw marks to UMS points applies at all parts of the mark range. The slope of the 'ideal' line is 90/60=1.5, which was the weighting factor used in Table 1.10 above.

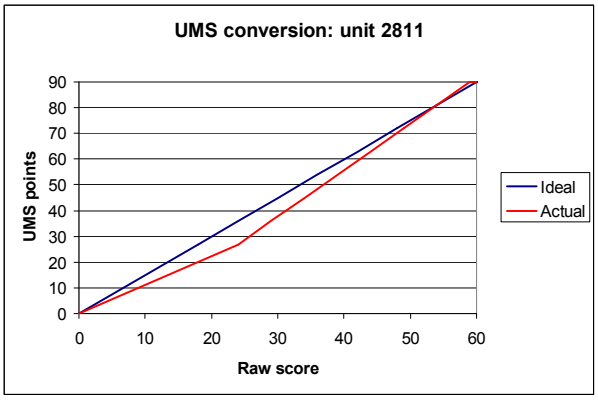

Figure 1.16: Raw mark to UMS conversion for Chemistry unit 2811 (June 2009).

One way to estimate the effect of the UMS transformation might be to calculate the weights based on the 'conversion rate' of raw marks to UMS points in the 'main line' part of the transformation. This can be done by dividing the UMS A-E distance by the raw A-E distance in Table 1.8 above, which gives the slope of the red line in the A-E range in Figure 1.16. These 'UMS weightings' and the new Composite Alpha are shown in Table 1.11 below. It is a coincidence that the UMS weightings happen to be the same for 2811 and 2812 – this would not usually happen, because it depends on where the raw grade boundaries are set.

Table 1.11: AS Chemistry 3882. Original weights, UMS weights, and new composite reliability.

Unit	Original weight	UMS weight	Composite Alpha	SEM
2811	1.5	1.8 (36/20)	0.924	13.11
2812	1.5	1.8 (36/20)		
2813	1.0	1.2 (48/40)		

Composite Alpha was unaffected by the slight changes to the weights, because their relative magnitude happened to be unchanged (although this in general would not necessarily be the case). However, the SEM increased by 2.2 points, because of the overall increase in composite score variance.

1.3.4 Composite Alpha of a 2-unit AS level.

The new 2-4 unit structure for AS/A level has simplified the situation regarding calculation of Composite Alpha. For example, there are now fewer possible routes to the composite, although the number is still considerable. In the example below, examinees had to take Units 1 and 2, which were available in January and June, giving four possible combinations. In fact, the vast majority had taken one of two combinations:
Unit 1 in January, Unit 2 in June (N=5245)
Unit 1 in June, Unit 2 in June[13].(N=4357)

Table 1.12: Descriptive statistics for this 2-unit AS level.

Unit	Max raw	Max weighted	No. of examinees	Mean	SD	Cronbach's Alpha	Correlation
Unit 1 (Jan)	60	60	5245	33.63	7.61	0.642	0.508
Unit 2 (June)	90	90	5245	43.11	11.94	0.733	
Unit 1 (June)	60	60	4357	32.86	8.56	0.653	0.601
Unit 2 (June)	90	90	4357	40.56	12.24	0.750	

[13] A large number of examinees (2352) took Unit 1 in both January and June. The correlation between their raw scores was 0.474. As discussed earlier, it is doubtful whether this is a satisfactory estimate of parallel forms reliability.

Table 1.13: Composite Alpha for this 2-unit AS level.

Unit	Max UMS	Weight	Composite Alpha	SEM	UMS weight	Composite Alpha	SEM
Unit 1 (Jan)	80	1.33	0.799	10.22	1.88	0.798	15.95
Unit 2 (June)	120	1.33			2.18		
Unit 1 (June)	80	1.33	0.820	10.58	2.29	0.819	17.65
Unit 2 (June)	120	1.33			2.18		

As with the AS Chemistry, the values of the weights have little effect on the value for Composite Alpha, but have a larger effect on the SEM.

1.3.5 Composite Alpha of a linear GCSE.

This GCSE was linear, with all components only available in the June examination session. Re-sitting examinees were allowed to carry forward the coursework mark from a previous session. We derive Composite Alpha for the foundation tier and higher tier options below, using only examinees who did not carry forward their coursework mark. Because we did not have any means of calculating a value for Cronbach's Alpha for the coursework component, these values were estimated. There was no use of the UMS scheme and each component carried its 'natural' weight in the composite – i.e. the raw marks from each were simply added together with no weighting (which is equivalent to using a value of 1 for the weight of each component).

Table 1.14: Descriptive statistics for this linear GCSE.

Component	Max raw	No. of examinees	Mean	SD	Cronbach's Alpha	Corr. with 02/04	Corr. with 05
01 Found.	80	1762	30.79	10.27	0.837	0.753	0.397
02 Found.	80	1762	31.52	10.84	0.839		0.382
05 C/work	40	1762	24.04	7.77	*0.500		
03 Higher	80	3082	51.17	14.28	0.857	0.812	0.498
04 Higher	80	3082	49.16	13.56	0.841		0.523
05 C/work	40	3082	31.56	6.15	*0.600		

*Estimated value. It seemed reasonable to estimate a higher value for the higher tier coursework because the correlations with the other components were higher, in particular the correlation with component 04 was greater than 0.5.

Table 1.15: Composite Alpha for this linear GCSE.

	Max raw	Composite Alpha	SEM
Foundation	200	0.885	8.14
Higher	200	0.920	8.57

For both tiers, the value for Composite Alpha was higher than the values for Cronbach's Alpha of the components.

1.3.6 Composite reliability in terms of grade bandwidth

As with the individual units/components, it is possible to attempt to compare the composite grade bandwidth with the composite SEM to get an index of reliability in terms of the grade scale. Obviously this is much more problematic with assessments that use the UMS scale, first

because it is not entirely clear what values should be taken as the composite weights, and second because the assessment aggregate can be arrived at in so many different ways. But for the sake of interest, the values for the two indices explained earlier (bandwidth:SEM ratio, and P(different grade) given true score in middle of band) are presented below for the assessments where we derived the composite reliability. The SEM calculated using the 'UMS weight' was taken for the unitised assessments.

Table 1.16: Composite reliability in terms of grade bandwidth

Assessment	Maximum mark	Grade bandwidth	SEM	Bandwidth: SEM	P (different grade)
AS Chemistry 3882	300	30	13.12	2.29	0.25
2-unit AS level (1)	200	20	15.95	1.25	0.53
2-unit AS level (2)	200	20	17.65	1.13	0.57
Linear GCSE Foundation tier	200	14	8.14	1.72	0.39
Linear GCSE Higher tier	200	24	8.57	2.80	0.16

Table 1.17: Summary of unit/component and composite reliabilities.

Assessment	Unit/component	Alpha	SEM	Bandwidth:SEM	P (different grade)
AS Chemistry 3882	2811	0.813	4.51	1.33	0.51
	2812	0.827	4.05	1.48	0.46
	2813	*0.823	*6.07	*1.65	*0.41
	Composite	0.924	13.12	2.29	0.25
2-unit AS level (1)	Unit 1 (Jan 09)	0.641	4.56	1.10	0.58
	Unit 2 (June 09)	0.733	6.16	0.97	0.63
	Composite	0.798	15.95	1.25	0.53
2-unit AS level (2)	Unit 1 (June 09)	0.653	5.04	0.79	0.69
	Unit 2 (June 09)	0.750	6.13	0.98	0.62
	Composite	0.819	17.65	1.13	0.57
Linear GCSE Foundation tier	01	0.837	4.15	1.20	0.55
	02	0.839	4.34	1.15	0.57
	05 (coursework)	*0.500	*5.49	*0.73	*0.72
	Composite	0.885	8.14	1.72	0.39
Linear GCSE Higher tier	03	0.857	5.40	2.04	0.31
	04	0.841	5.41	1.85	0.35
	05 (coursework)	*0.600	*3.89	*1.54	*0.44
	Composite	0.920	8.57	2.80	0.16

* entirely or partly estimated.

Table 1.17 shows that in all cases the composite reliability, as calculated by Composite Alpha, was higher for the composite than for the elements comprising it. In terms of grade bandwidth it was also always the case that the composite had a higher bandwidth:SEM ratio than the individual elements. This data suggests that all elements (and composite) of the 2-unit AS level had lower reliability than might be desired, whichever index is used. Also it is interesting to note that the reliability indices for the linear GCSE were higher for the higher tier than the foundation tier. The difference appears small when Composite Alpha is used, but appears larger when the two bandwidth-related indices are used.

1.3.7 IRT composite reliability

All the problems associated with calculating composite reliability using classical test theory, described in detail in the previous sections, still apply, with extra problems discussed below.

1. The SEM varies with estimated ability – this could be avoided by simply using an average error variance, as in He (2009).
2. Each IRT analysis of an individual unit/component creates a scale with its own origin and unit – this could be tackled be rescaling the ability measures from the components so that they have the same means and standard deviations (He, ibid.).
3. The assumption that all the components are measuring the same trait may not hold – this could be addressed by using a multidimensional IRT model, although this would create further problems in terms of interpreting the multidimensional outcome.
4. There is a lack of a meaningful aggregate ability scale. Essentially the current system for unitised assessments already has two scales – the raw mark scale and the UMS scale. Involving a third – the IRT scale – is likely to create more confusion than it removes. He (ibid.) suggests that a composite ability estimate can be obtained by creating a weighted sum of the component ability estimates, but does not say how the weights should be obtained.

It seems to us that the proposed fixes to the first and fourth problems merely mimic what the classical composite alpha achieves, and that the proposed fix to the second problem potentially introduces invalidity if the means and standard deviations on the components *should* in fact differ.

Regarding the third problem, the assessment structure of GCSEs and A levels is often designed to ensure that different units/components measure different knowledge and skills and are hence multidimensional, although it can always be argued that the resulting composite is in some sense unidimensional if the results from each unit/component can be meaningfully aggregated.

The literature on IRT composite reliability is sparse and somewhat unconvincing. One method for deriving a value is given in Childs, Elgie, Gadalla, Traub and Jaciw (2004). Even if this relatively complex procedure could be followed successfully, the validity of the results would still depend on the extent to which the IRT model had fitted in the first place.

A brief analysis of the problem from a Rasch perspective is given in Wright (1994). From this perspective, if the parts of the composite are all measuring the same thing then they should be analysed together (i.e. in a single analysis, as though all the items formed one large test). But, as Wright says:

> *"When, however, the differences among part-measures are significant enough (statistically or substantively) to become interesting, what then is the logic for either a part-measure mean or even a whole-test measure?"* (Wright, 1994).

Wright (ibid.) shows that the same composite raw score may give different composite ability estimates, depending on the part-scores. Furthermore, the composite variance estimated from part-scores will be further affected by differential measurement error in the part-tests, and by uneven patterns of misfit.

In short, it seems to us that if the process of estimating Composite Alpha under classical test theory had strained credibility, the process of coming up with a number representing composite reliability in IRT for GCSEs and A levels would stretch it beyond breaking point. See the discussion in Section 1.5 for further comment.

However, for the sake of interest, two possible ways to calculate composite IRT reliability based on the results of separate IRT analysis of the components are described below.

Method 1 (mimic the classical composite reliability analysis):
- Carry out an IRT analysis of each unit/component.
- Transform the scales arising from each individual analysis so that the mean and SD of the distribution of estimated ability for the common examinees is the same.
- Derive the equivalent means, variances, reliabilities, and between-unit/component correlations as used in the calculation of classical composite alpha.
- Using the relative weights implied by the UMS weighting scheme as specified in the assessment structure, calculate a composite IRT alpha using equation (11).

Table 1.18: IRT composite reliability for AS Chemistry 3882 – Method 1.

Unit	Max raw	Max UMS	Weight	No. of examinees	R_β	Correl. with 2812	Correl. with 2813	Composite IRT reliability
2811	60	90	3/2	625	0.865	0.620	0.748	
2812	60	90	3/2	625	0.847		0.616	0.939
2813*	45	60	4/3	625	0.863			

*the coursework component of this unit has been ignored.

Method 2 (calculate a composite ability estimate for each examinee):
- Carry out an IRT analysis of each unit/component.
- Transform the scales from each individual analysis so that the mean and SD of the distribution of estimated ability for the common examinees is the same.
- Calculate a weighted average ability estimate for each examinee (composite ability), using the relative weights implied by the UMS weighting scheme as specified in the assessment structure.
- Calculate a information-weighted average error variance estimate (composite error) for each examinee, also factoring in the component weightings used in the previous step.
- Calculate the variance of composite ability and the average composite error across all examinees.
- Calculate a composite separation reliability using equation (10).

Table 1.19: IRT composite reliability for AS Chemistry 3882 – Method 2.

Variance of weighted average composite ability estimates	0.7734
Average error variance of the estimated composite abilities	0.0395
Composite IRT reliability	0.949

The result from Method 2 was 1 percentage point higher than the result from Method 1, and both were higher than the IRT reliabilities of the individual units/components.

We did not attempt to carry out a single IRT analysis of all three components together, because it would not have been possible to allow for the different weightings of the components. If we had, each total raw score would have been associated with a single ability estimate and associated standard error. Using Method 2 above meant that examinees with the same composite raw score could have different estimated composite abilities and different standard errors, as shown in Figure 1.17 on the next page.

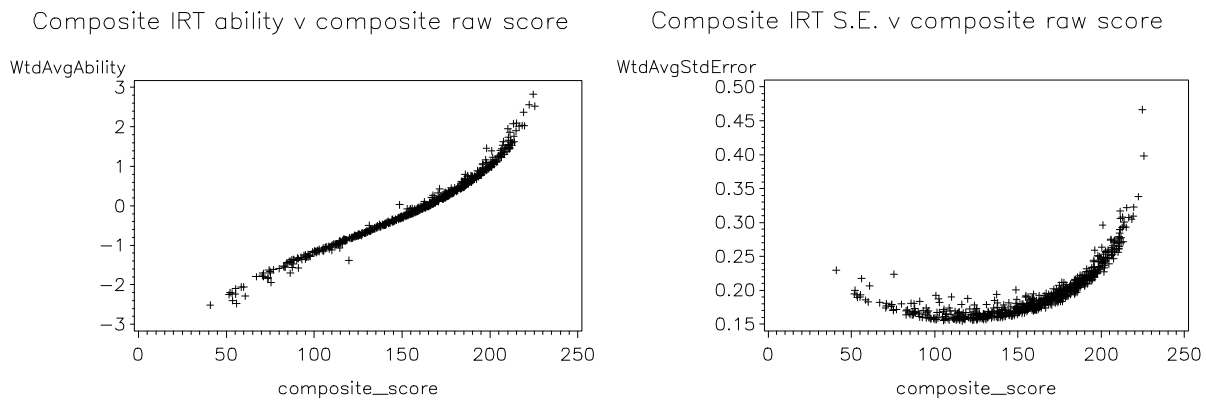

Figure 1.17: AS Chemistry 3882 – plot of Composite IRT ability (left) and composite IRT SEM (right) against composite raw score[14].

The variable SEMs for examinees with the same composite raw score illustrate the point made by Wright (1994). They are potentially diagnostic of different profiles of scores across the three units/components. For example, in the right panel of Figure 1.17, the examinee with a composite score of ≈75 but with a composite SEM of ≈0.23 had obtained (unweighted) raw scores on the three units/components of 5 out of 60, 40 out of 60 and 6 out of 45. It is reasonable to conclude that there is more uncertainty about the 'true' ability of this examinee than of others with a more even profile.

[14] This composite raw score was calculated from the unit/component raw scores using the weights in Table 1.18. It did not involve the UMS scale.

1.4 Classification consistency

In a GCSE or an A level, the outcome for the examinee is generally a grade, rather than a raw score. Since 2002, A level examinees have been told their UMS score on each unit, which is essentially a more fine-grained version of the grade scale. UMS scores will also be reported to GCSE examinees in the new unitised specifications.

Given that the grade is the focus of reporting, arguably it is more appropriate to evaluate reliability at this level too. The issues associated with imposing a relatively coarse grade scale onto a finer-grained underlying mark scale have been discussed several times over the decades in the context of UK examinations (e.g. Skurnik & Nuttall, 1968; Please, 1971; Cresswell, 1986). These discussions are summarised by Bradshaw and Wheater (2009):

> *The total number of grades in an assessment influences the reliability of the grades. By decreasing the number of grades in an assessment, the reliability of the grades is increased, but there is also a loss of information, with two main effects. These are that a misclassification has a more serious implication for a student, and that students and their teachers learn less from the results of an assessment. Although there has been research into the reporting of error and uncertainty, there have been no clear answers as to how the reliability of grades could be reported in a meaningful way.* Bradshaw and Wheater (2009, p13-14)

Intuitively, the loss of information that occurs when marks become grades seems unfair because two examinees who scored, say, 6 marks apart might receive the same grade on a unit/component, whereas two examinees who scored 1 mark apart might receive different grades if they were either side of a grade boundary. The counter-arguments are that i) using grades avoids spurious precision (because they are more reliable in the sense of repeatable); ii) it allows aggregations to be made across different units/components; and iii) it provides a 'common currency' for comparisons to be made across different assessments. A fourth argument for using grades – that they represent qualitatively different categories of achievement (implying that examinees 1 mark apart either side of a grade boundary are in fact more different than examinees 6 marks apart within a grade) – is possible, but as far as we are aware has not been widely deployed, probably because it is unsupportable in the context of GCSEs and A levels.

When grades become the focus of attention, the reliability question becomes 'To what extent would the grade outcomes be the same if the test or assessment were to be replicated?' The same issues discussed earlier about what stays the same and what differs in the hypothetical replication still apply. Here we assume it to be consistency of classification on two parallel forms of the test. The question can be framed in a variety of ways:

- in terms of the examinee – the probability that a given individual would get the same grade;
- in terms of all the examinees – the proportion that would get the same grade;
- in terms of the examinees with a given grade – the proportion that would get the same grade.

These are questions about classification *consistency*, as opposed to classification *accuracy*, which is usually defined in terms of whether an observed outcome is the same as the 'true' outcome (e.g. Livingstone & Lewis, 1995; Wheadon & Stockdale, 2010; but see also Hutchison & Benton, 2009; Newton, 2009; and Bramley, 2010).

It should be noted that this kind of information about repeatability of grades is not necessarily related to traditional indices of reliability like Cronbach's Alpha. Considering an extreme scenario where all the examinees are of very high ability compared to the test – if all the examinees score well above the top grade boundary, their grades are likely to be highly repeatable, yet the value for internal consistency reliability could be low because the true score variance was low. (This is sometimes observed with language examinations for subjects like

Gujarati where although the exam is in theory aimed at second-language speakers, most of the examinees are native speakers).

In an earlier section (1.1.5) we have already considered one way expressing reliability information in terms of the grade scale – by relating the SEM estimated via Cronbach's Alpha to the size of the grade bandwidth in marks. However, to answer the three questions about classification consistency posed above, ideally the conditional SEM obtained from an IRT method would be used, because this varies across the score range and should thus give a more accurate answer.

The approach we adopt is the one described in Stearns & Smith (2008). Although there are potential criticisms of this approach (see the discussion in Section 1.5) it has the advantage of relative simplicity and follows naturally from the results of the IRT analysis described above in Section 1.2. Each examinee has an ability estimate corresponding to their raw score, and an associated standard error of measurement. There is also an ability estimate associated with each grade boundary – namely the ability estimate corresponding to that raw score. On the assumption that measurement errors are normally distributed, the distance of an examinee's estimated ability from a given grade boundary in terms of the measurement error distribution is:

$$z = \frac{\beta_n - GB}{SE_{\beta_n}} \quad (12)$$

where β_n is the estimated ability of examinee n, $SE_{\beta n}$ is their estimated asymptotic standard error, and GB is the ability corresponding to the grade boundary (Stearns & Smith, 2008, p308). This z-score can then be converted into the probability that an examinee would be the same side of the grade boundary on re-testing, using the standard tables of cumulative normal probabilities corresponding to z-scores. If there is only one boundary (e.g. a pass-fail test), this is also the probability that the examinee would receive the same grade classification on re-testing. If there is more than one boundary, the z-score distances from each boundary can be calculated and hence the probability of being reclassified into the same grade on re-testing.

The observed score/ability frequency distribution can then be used to calculate the proportion of all examinees who would be consistently classified on re-testing, and also the proportion of examinees within each grade who would be consistently classified on re-testing. This is illustrated below for two of the units for which we carried out an IRT analysis.

1.4.1 Classification consistency for a component of an AS unit.

Figure 1.18 below shows the ability scale on the x-axis and on the y-axis the probability of classification into each of the different available grades at each point on the ability scale. The graph shows that on the left, at very low abilities (raw scores) the probability of being re-classified into the lowest category approaches 1, but this falls to zero as ability increases. The probability of being re-classified into each subsequent grade then rises and falls as ability increases until at the top end of the ability scale, the probability of being re-classified into the top category (grade A) approaches 1. Within each grade band, it is usually more likely that an examinee would be reclassified within that band than in any other particular band, but overall the probability of consistent classification could be less than 0.5, as Figure 1.18 shows for the E and B grade bands.

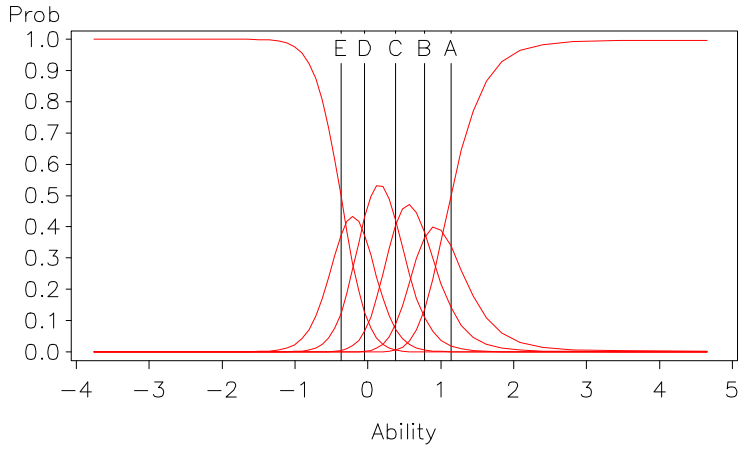

Figure 1.18: A component of an AS unit – probability of being consistently classified.

Figure 1.19 below shows the observed frequency distribution (blue line) and the estimated percentage of examinees at each ability who would be consistently classified (red line). Where the lines overlap, all the examinees would be estimated to be classified consistently. It is clear that this only happens at the lowest and highest ends of the ability scale. Graphs like this give a visual impression of where classification inconsistency has the greatest impact, and make it clear that we can expect more inconsistent classification on tests where the grade boundaries fall in the main part of the score distribution. This particular unit is seen to have had a high ability set of examinees – the modal mark was above the grade A boundary.

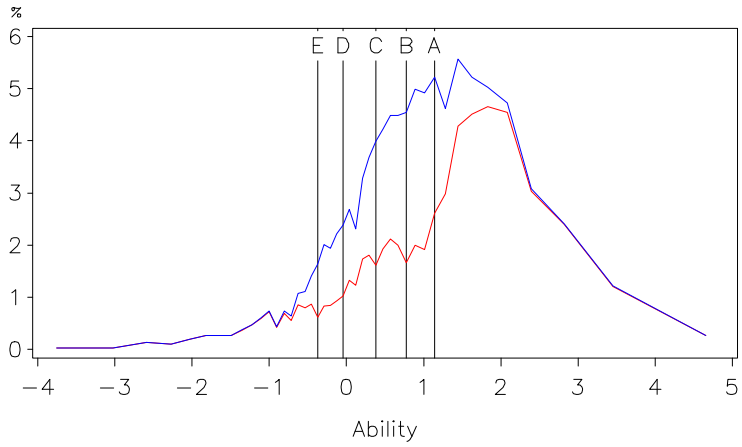

Figure 1.19: A component of an AS unit – percentage of examinees consistently classified.

The first column of Table 1.20 on the next page shows the different grades available. The second column shows how many marks on the raw score scale were allocated to each grade. (The 'all' row has a bandwidth of 61 because the maximum mark for the unit was 60, giving 61 possible scores on the test, including zero). The third and fourth columns show the number and percentage respectively of examinees in each grade category. The fifth column shows the estimated percentage (of the total) of examinees consistently classified (effectively the area under the red line within each grade band as a percentage of the total area under the blue line in Figure 1.16). Finally, the sixth column shows the estimated percentage of examinees within each grade band consistently classified (effectively the area under the red line within each grade band as a percentage of the area under the blue line within each grade band in Figure 1.19).

Table 1.20: A component of an AS unit – percentage of examinees consistently classified.

Grade	Grade bandwidth (marks)	No. of examinees	% of examinees	Estimated % of total consistent	Estimated % consistent within grade
All	46	2986	100.00	61.66	61.66
A	11	1115	37.33	30.47	81.64
B	3	432	14.46	5.57	38.54
C	4	513	17.19	7.65	44.50
D	5	428	14.33	7.12	49.69
E	4	233	7.80	3.21	41.14
U	19	265	8.89	7.63	85.84

It is clear from inspection of the figures and Table 1.20 that most of the grade A and U examinees would be consistently classified, but that only around half of those in the intermediate grades would be. The value of 61.7% summarises overall classification consistency.

1.4.2 Classification consistency for a Higher tier GCSE unit.

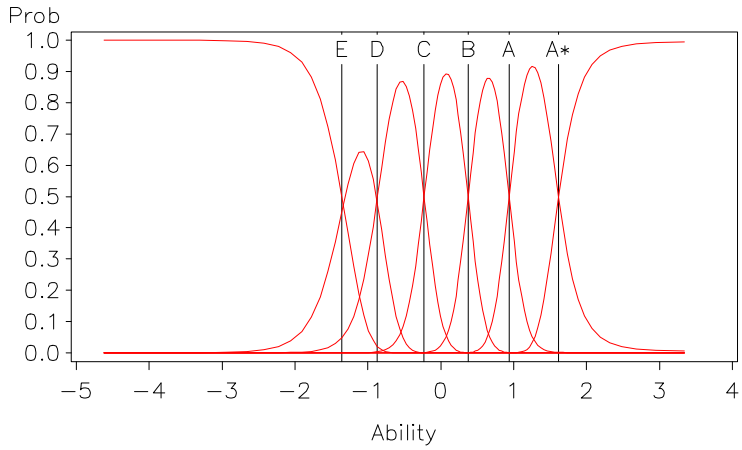

Figure 1.20: A Higher tier GCSE unit – probability of being consistently classified.

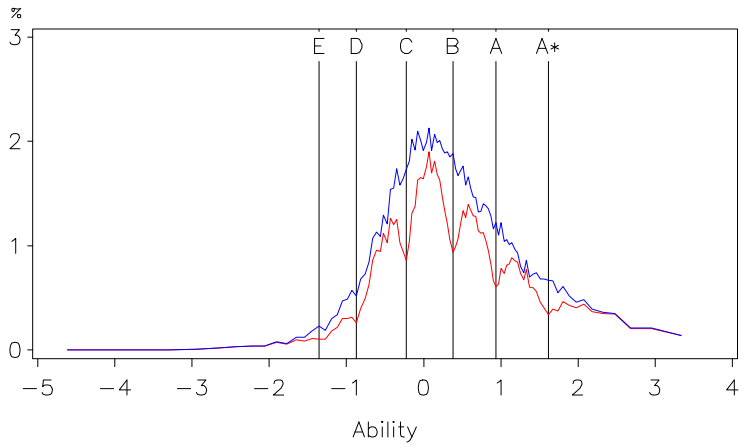

Figure 1.21: A Higher tier GCSE unit – percentage of examinees consistently classified.

Table 1.21: A Higher tier GCSE unit – percentage of examinees consistently classified.

Grade	Grade bandwidth (marks)	No. of examinees	% of examinees	Estimated % of total consistent	Estimated % consistent within grade
All	101	31646	99.98*	74.82	74.84
A*	15	1802	5.70	4.55	79.82
A	17	4907	15.50	11.77	75.96
B	17	8135	25.71	18.98	73.83
C	17	10501	33.17	25.06	75.56
D	14	5260	16.61	12.37	74.49
E	7	822	2.59	1.52	58.87
U	14	219	0.70	0.56	79.92

*Percentages do not sum to 100 because of rounding.

The results for this higher tier GCSE unit provide an interesting contrast with the AS unit. Figure 1.21 and Table 1.21 show clearly that the probability of consistent classification would be high across all grades except for grade E[15]. However, the overall figure for classification consistency of 74.8%, while higher than that for the AS (61.7%), is perhaps not as much higher as might be expected from the graphs. One reason for this could be that there were 7 rather than 6 grade categories on the GCSE unit's scale, although this ought to be more than compensated for by the fact that the mark scale was much longer (100 marks compared with 45). Another reason is the location of the score distribution, which for the GCSE unit was better matched with the range of scores covered by the grade boundaries than the AS unit with the mode being in the middle of the grade C band rather than above the grade A boundary. In particular, a much lower percentage of the GCSE unit's examinees were in the extreme categories.

[15] The grade E boundary is below the target range for this paper, but is there as a 'safety net' for examinees who might otherwise receive a grade U classification. Its bandwidth is specified to be half the C/D bandwidth.

1.4.3 Classification consistency for all twelve units/components that were IRT analysed

Table 1.22 below shows the overall classification consistency index for all twelve units/components for which we carried out an IRT analysis (these are the same units/components in the same order as Table 1.7). The other global reliability indices are included in Table 1.22 for comparison. Table 1.23 shows the rank-order correlation among the four reliability indices.

Table 1.22: Overall reliability indices for twelve GCE / GCSE units/components.

Type	Part of	Paper Total	# of items	N	Alpha	R_β	band width: SEM	Class. con. %
AS unit	A 2/4-unit GCE	60	8	4355	0.652	0.737	0.79	52.4
AS unit		90	11	4352	0.750	0.815	0.98	51.9
AS unit	A 3/6-unit GCE	60	27	625	0.813	0.861	1.14	64.5
AS unit		60	35	625	0.827	0.841	1.23	59.8
AS unit comp.		45	23	625	0.862	0.856	0.96	61.7
GCSE comp.	Foundation tier	80	30	1761	0.837	0.848	1.21	62.5
GCSE comp.		80	34	1758	0.839	0.877	1.16	62.9
GCSE comp.	Higher tier	80	29	3081	0.857	0.877	2.04	73.6
GCSE comp.		80	27	3082	0.841	0.883	1.85	71.1
GCSE unit	Foundation tier	100	58	26233	0.930	0.940	2.41	73.0
GCSE unit	Higher tier	100	46	31629	0.915	0.921	3.04	74.8
AS unit	A 2/4-unit GCE	90	29	689	0.930	0.956	1.36	72.1

Comp. = Component.

Table 1.23: Relationship between the four indices of reliability (Spearman's rho).

	Alpha	R_β	Bandwidth: SEM	Class. con. %
Alpha	1.000	0.881	0.697	0.767
R_β		1.000	0.767	0.879
Bandwidth: SEM			1.000	0.860
Class. con. %				1.000

The four indices all show a positive relationship with each other, as shown in Table 1.23. As expected, the two indices that do not take into account the grade boundaries (Alpha and R_β) are more closely related to each other than to the others. The two indices that do take into account the grade boundaries (bandwidth:SEM and classification consistency) are more closely related to each other than they are to Alpha, but surprisingly classification consistency has a slightly higher rank correlation with R_β than it does with bandwidth:SEM ratio. Of course, with only twelve values being correlated it would not be sensible to read too much into these correlations.

1.5 Discussion

In the previous sections we have provided a variety of analyses of test-related reliability, which could be categorised along three independent facets:

- those based on Classical Test Theory v those based on Item Response Theory
- those based on test scores v those based on grades (classification consistency)
- those based on individual units or components v those based on the aggregate (composite) assessment.

In this section we discuss some of the issues and implications in terms of the underlying theory, its application in practice, and finally in terms of the wider aim of assessment agencies and the regulator communicating with the public about measurement error in assessments.

The Ofqual-commissioned reports by He (2009), Johnson & Johnson (2009) and Hutchison & Benton (2009) have given the conceptual basis of reliability a more thorough and comprehensive discussion than is possible here. However, it seemed to us worthwhile to address some points that have not been explicitly covered in the above reports.

The first is the contrast between CTT and IRT. Are they just different tools for getting the same job done, the choice of which depends on the preference of the analyst as much as features of the situation? The analysis in Mellenbergh (1994) treats them in a unified framework, distinguishing between continuous test score models, discrete test score models, continuous item score models and discrete item score models. He shows that the population-dependent concept of reliability (normally used in CTT) and the population-independent concept of information (normally used in IRT) can be applied to all four kinds of model. Other authors (e.g. Holland & Hoskens, 2003; Bechger, Maris, Verstralen & Beguin, 2003) have also shown the connections between CTT and IRT. On the other hand, it can be argued that CTT and IRT methods are based on different underlying philosophies of measurement (Borsboom, 2005). Classical test theory talks about measurement errors, but arguably has little to do with measurement. The raw scores it analyses are simply treated as though they *are* measurements. The 'true score' of CTT is not the same as the 'trait score' or 'construct score' of IRT, even though they are easily confused – a confusion that keeps returning "like an alien in a B-movie", to use the vivid phrase of Borsboom (ibid.). Because in CTT the true score is defined as the long run average of the examinee's observed score over independent replications, every conceivable test has a true score. CTT does not specify what it means for a test to measure an attribute. In IRT the trait can be conceived as the 'common cause' of the item responses (Borsboom, ibid.) and issues of validity can be separated from issues of reliability by using the IRT model to test substantive hypotheses about the underlying structure of the attribute. Confirming that the data fits the model (or that the model fits the data) is a prerequisite to treating the estimated measures as measures.

CTT and especially its conceptual development into Generalizability Theory are really about sampling rather than about placing examinees on a scale, as Johnson & Johnson (2009, p20) make clear. While a conceptualisation based on sampling might be appropriate for national monitoring surveys of achievement that actually use formal sampling of both examinees and items, it is less clear that it is appropriate for quantifying test-related measurement error in examinations that do not formally sample either items or examinees[16] (or markers). The emphasis in IRT (and particularly in the Rasch model) on estimating the location of each person on the trait with an associated amount of uncertainty seems to capture better intuitively what measurement is about. Physical measuring devices usually are supplied with some indication from the manufacturer of the ± limits representing the degree of accuracy that can be expected. However, the validity of the instrument (i.e. that it measures what it purports to measure) can be more safely assumed than in the realm of psychological or educational testing.

[16] The items and examinees are of course a 'sample' in the sense of being a *subset* of those that might conceivably have been written, or entered for the examination, but that is not the sense of 'sample' intended here.

While our conceptual preference is for an IRT approach, we must concede that as we have used it here, it has merely been a convenient way to generate a conditional SEM (i.e. conditional on ability, and varying across the score range). No investigation of the fit of the Rasch model to the data was carried out, nor were any other IRT models explored. Lack of fit can in some cases be addressed by removing misfitting items or examinees and re-analysing – but in terms of the estimated standard error of examinee ability only the removal of items would have a noticeable effect, and this would mean that the ability estimates would then be based on a different set of items to the ones on which the examinees' scores were obtained in the actual examination. Wright (1995) suggests a formula for inflating the reported 'ideal' standard errors in order to allow for misfit. The idea has some intuitive appeal – there is presumably more uncertainty about the true ability of an examinee who has succeeded relatively often on difficult items and failed relatively often on easy items than there is about the ability of an examinee with the same raw score whose pattern of item scores fits the model. However, this inflation formula does not seem to be widely used, as far as we are aware.

Although CTT appears to make fewer demands on the data than IRT, in the sense that there are no explicit tests of model fit, it still involves assumptions about how the data are generated and, in particular, in relating calculated values such as Cronbach's Alpha to theoretical concepts like reliability. In most of the GCSEs and GCE units/components that we analysed, the 'items' were sub-parts of larger questions. If scores among sub-parts of the same larger question tend to be more related to each other than to scores on other questions, this would inflate the apparent reliability for both Cronbach's Alpha and the IRT separation reliability R_{β}. It would be possible to investigate this further by combining the sub-parts into question totals and re-calculating the reliability indices (Marais & Andrich, 2008).

A second issue worth mentioning is the interpretation of the standard error of measurement (SEM). In CTT this is an average value, so arguably it should not be reported in a way that encourages individuals to treat it as applying to them. A further problem is that strictly the SEM should be interpreted as applying to the distribution of observed scores around the true score, rather than vice versa. There are ways to derive a confidence interval for the true score but these involve regressing the observed score towards the mean of a suitable reference population (e.g. Harvill, 1991). If the reliability in the reference population is relatively high and the observed score is relatively close to the mean, the usual practice of using the SEM to form a confidence interval around the observed score is a reasonable approximation.

The same point applies to IRT estimates of conditional SEM (e.g. Wright, 1995). The method we used to derive the classification accuracy in Section 1.4 not only treated the estimated abilities as the true abilities, but also assumed that estimated abilities are normally distributed around the true ability. Neither of these assumptions is strictly accurate. There are some sophisticated ways to get round these problems (e.g. Emons, Sijtsma & Meijer, 2007; Verstralen & Bechger, 2008). Simulating response patterns and data sets based on the estimated IRT parameters and deriving indices of classification accuracy and consistency from these simulations would be another possibility. However, such approaches can be complex and computationally intensive, and would rely on the data fitting the IRT model. Stearns & Smith (2008) found that their relatively simple method outperformed some more complex ones, and our data probably did not fit the model well enough to justify using a more complex method. For a much more detailed investigation of classification accuracy and consistency using data from GCSEs and A levels, see the report by Wheadon & Stockford (2010).

A third issue is how to conceptualise composite reliability for GCSEs and GCEs. The overview in He (2009), while comprehensive in terms of methods covered, does not really address the problems that arise in practice from the way these assessments, especially the modular ones, are structured. They were not designed to make life easy for the psychometrician – which is not necessarily a bad thing. Nonetheless, the large amount of flexibility in choice of units, when to take them, and the possibility of re-sits, combined with the complexities arising from the use of the Uniform Mark Scale with potentially different effective weights applying to different versions

of the same unit, make the calculation of reliability indices seem like a number-crunching exercise that becomes increasingly disconnected from reality. Our results seemed to confirm the expectation that the composite is more reliable than its individual components, but how much the absolute numerical values of the composite reliability indices should be taken seriously is open to question.

The GCSE and GCE awarding bodies deliver thousands of examination units/components each year. These units/components are not designed to fit a particular IRT model, and the various components of a complete assessment are often designed to test different knowledge and skills. If information about reliability is to be routinely generated for every unit/component (and the composite assessment) where possible, it is probably unrealistic at the current time to expect IRT methods to be used. The much-criticised Cronbach's Alpha, for all its flaws, can at least be calculated from the data in a batch job as easily as the mean or standard deviation. It is then possible to interpret it in a comparative way, as we have shown in Section 1.1.4, by plotting it against features of the units/components that are relevant to interpreting it properly, such as the number and type of items, the length (maximum available mark) of the unit/component, and the weighting of the unit/component in the overall assessment. We would suggest that inspection of plots like this could be the most useful way to identify units/components where there might be a problem with reliability – for example a unit/component that had a noticeably lower value for Cronbach's Alpha than other similar units/components. On the face of it, low values for alpha suggest a relatively large amount of random noise in the data. The task then for the assessment agency would be to explain why the value was lower. It may well be that there is a perfectly reasonable explanation – or on the other hand, the low value could diagnose an inherent problem in the assessment that should be addressed. Hutchison & Benton (2009, p40-41) make the important point that Cronbach's Alpha takes account of non-systematic inter-marker variability. If, as they claim, this is the main source of between-marker difference, then one explanation for low values of Cronbach's Alpha could be unreliable marking.

The imposition of a grade scale on the raw mark scale creates further problems and opportunities for investigating reliability. We have suggested one index of reliability, the grade bandwidth:SEM ratio, that could be used in addition to Cronbach's Alpha. Higher values of this index indicate more reliability. Lower values are caused by either a narrower grade bandwidth, or a higher SEM. In fact, we have suggested that the size of the grade bandwidth might be a statistic worth reporting in its own right. It would seem desirable to have grade bands that are wide, and that cover the full range of the available marks – but there are competing factors, such as the desire to have the lowest grade represent a significant amount of achievement, and to have marks at the top end that 'stretch the most able'. Both these factors would tend to compress grade bandwidths.

It would also be possible to invert the bandwidth:SEM ratio, in which case it would become effectively an SEM in grade units. This might be easier to understand – but could invite misleading interpretations. The largest value for this inverted index in our data would have been $\approx$1.5. A 95% confidence interval around a true score would then be $\approx \pm$ 3 grades – i.e. the entire A to U range, which would probably not inspire much confidence. Keeping the index in the form we have suggested and plotting it for all the units/components would allow outliers to be identified and investigated without inviting possibly spurious interpretations at the individual unit/component level or examinee level.

The graphical and tabular presentations of classification consistency information in Section 1.4 are an appealing way communicate about reliability. However, the interpretation of their absolute values does depend on the accuracy of the estimated SEM. If they were to be used for comparative purposes, or just to give a general impression of reliability, then it might be possible to use a crude estimate of the SEM calculated via Cronbach's Alpha to calculate classification consistency. We conclude this part of the report by presenting on the next page a 1-page summary of information from a unit/component, all based on CTT statistics. It should be possible to produce this kind of information as a routine 'batch job'.

Table 1.24: Example summary statistics table.

Maximum mark for paper	45
Number of examinees	2986
Mean mark	30.55
SD mark	8.12
Cronbach's Alpha	0.85
Standard error of measurement (SEM)	3.10
A-B grade bandwidth (marks)	3
Bandwidth : SEM ratio	0.97

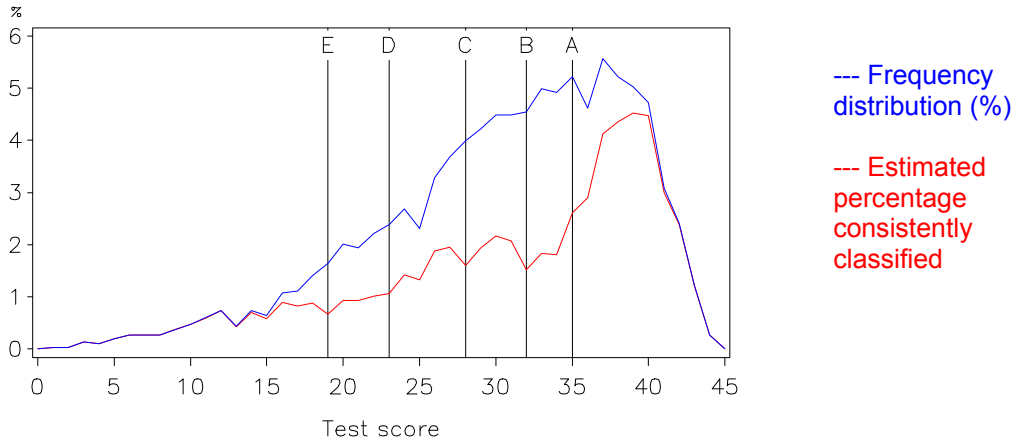

Figure 1.22: Example classification consistency plot based on average SEM derived via Cronbach's Alpha.

Table 1.25: Example classification consistency table based on average SEM derived via Cronbach's Alpha.

Grade	Grade boundaries	Grade bandwidth (marks)	Number of examinees	% of examinees	Estimated % consistently classified	(IRT for comparison)
All		46	2986	100.0	61.7	(61.7)
A	35	11	1115	37.3	79.9	(81.6)
B	32	3	432	14.4	35.6	(38.5)
C	28	4	513	17.2	45.2	(44.5)
D	23	5	428	14.3	53.2	(49.7)
E	19	4	233	7.8	45.3	(41.1)
U	0	19	265	8.9	87.4	(85.8)

The above two tables and graph summarise the outcomes from this single component of an AS unit. They highlight much of the information that might be of interest. In particular, showing the observed distribution of scores with the grade boundaries superimposed makes clear how well the test was targeted at its intended 'audience'. In this particular instance it seems that the paper was well targeted, but that the examinee cohort was of high ability.

The classification consistency statistics are generally similar to those obtained via the IRT analysis. It is a coincidence that the overall figure is the same – this would not generally be the case. Using the average SEM from CTT overestimates the error in the middle of the score range, but underestimates it at the extremes. Future research could examine the conditions under which the two give similar or different results, and determine whether it would be too misleading to use the CTT-based approximation.

1.6 References for Section 1

Andrich, D. (1982). An index of person separation in latent trait theory, the traditional KR-20 index, and the Guttman scale response pattern. *Education Research & Perspectives, 9(1)*, 95-104.

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

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

Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Beguin, A. A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement, 27*, 319–334.

Bell, J. F., Bramley, T., Claessen, M., and Raikes, N. (2006). Quality control of marking: some models and simulations. Paper presented at the annual conference of the British Educational Research Association (BERA), Warwick, September 2006.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison Wesley.

Borsboom, D. (2005). *Measuring the mind: conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University Press.

Boyle, A., Opposs, D. & Kinsella, A. (2009). No news is good news? Talking to the public about the reliability of assessment. Paper presented at the 35th International Association for Educational Assessment (IAEA) Annual Conference in Brisbane, Australia, 13–18 September, 2009. http://www.ofqual.gov.uk/files/2009-09-iaea-no-news-is-good-news.pdf Accessed 08/01/10.

Bradlow, E.T. (1996). Negative information and the three-parameter logistic model. *Journal of Educational and Behavioral Statistics, 21(2)*, 179-185.

Bradshaw, J., & Wheater, R. (2009) *International survey of results reporting*. NfER. Ofqual/10/4705. http://www.ofqual.gov.uk/files/Ofqual_10_4705_International_Survey_of_Results_Reporting_08_03_10_(2).pdf Accessed 16/9/10.

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.

Chamberlain, S. (2010). Public perceptions of reliability. AQA. Ofqual/10/4708. http://www.ofqual.gov.uk/files/Ofqual_10_4708_public_perceptions_reliability_report_08_03_10.pdf Accessed 16/9/10.

Childs, R.A., Elgie, S., Gadalla, T., Traub, R., & Jaciw, A.P. (2004). IRT-linked standard errors of weighted composites. *Practical Assessment, Research & Evaluation, 9(13)*. Retrieved December 17, 2009 from http://PAREonline.net/getvn.asp?v=9&n=13

Cronbach, L.J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika, 16(3)*, 297-334.

Donoghue, J.R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 31(4)*, 295-311.

Emons, W.H.M., Sijtsma, K., & Meijer, R.R. (2007). On the consistency of individual classification using short scales. *Psychological Methods, 12(1)*, 105-120.

Frisbie, D.A. (1988). Reliability of scores from teacher-made tests. *Educational Measurement: Issues and Practice, 7(1)*, 25-35.

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.

Green, S.B., & Yang, Y. (2009). Commentary on coefficient alpha: a cautionary tale. *Psychometrika, 74(1)*, 121-135.

Haertel, E.H. (2007). Reliability. In R. L. Brennan (Ed.), *Educational Measurement*. (pp. 65-110). ACE/Praeger series on higher education.

Harvill, L.M. (1991). An NCME Instructional Module on Standard Error of Measurement. *Educational Measurement: Issues and Practice, 10(2)*, 33-41.

He, Q. (2009). *Estimating the reliability of composite scores*. Coventry: Ofqual. http://www.ofqual.gov.uk/files/2010-02-01-composite-reliability.pdf Accessed 16/02/10.

He, Q., Opposs, D. & Boyle, A. (2010). *Public perceptions of unreliability in examination results in England: a new perspective.* Paper presented at the 36th International Association for Educational Assessment (IAEA) Annual Conference in Bangkok, Thailand, 22-27 August 2010. http://www.ofqual.gov.uk/files/2010-08-public-perceptions-of-unreliability-in-exam-results.pdf Accessed 16/9/10.

Hutchison, D. (2008). On the conceptualisation of measurement error. *Oxford Review of Education, 34(4)*, 443-460.

Hutchison, D., & Benton, T. (2009). *Parallel universes and parallel measures: estimating the reliability of test results.* Slough: NfER. Ofqual/10/4709 http://www.ofqual.gov.uk/files/2010-02-01-parallel-universes-and-parallel-measures.pdf Accessed 16/02/10.

Ipsos MORI (2009). *Public perceptions of reliability in examinations.* A research study conducted for Ofqual. Final report May 14th 2009. http://www.ofqual.gov.uk/files/2009-05-14_public_perceptions_of_reliability.pdf. Accessed 12/02/10.

Kendall, M.G., & Buckland, W.R. (1957). *Dictionary of statistical terms*. Edinburgh: Oliver and Boyd.

Livingston, S.A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement, 32(2)*, 179-197.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Marais, I., & Andrich, D. (2008). Effects of varying magnitude and patterns of response dependence in the Unidimensional Rasch Model . *Journal of Applied Measurement, 9(2)*, 105-124.

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

Mellenburgh, G. (1996). Measurement precision in test score and item response models. *Psychological Methods, 1(3)*, 293-299.

Muraki, E. (1992). A generalized partial credit model: application of an EM-algorithm. *Applied Psychological Measurement, 16*, 159-176.

Newton, P.E. (2005a). The public understanding of measurement inaccuracy. *British Educational Research Journal, 31(4)*, 419-442.

Newton, P.E. (2005b). Threats to the professional understanding of assessment error. *Journal of Education Policy, 20(4)*, 457-483.

Newton, P.E. (2009). The reliability of results from national curriculum testing in England. *Educational Research, 51(2)*, 181-212.

Ofqual (2009). Ofqual's Reliability of results programme: programme of work. Annex 1.

Ofqual (2009). GCSE, GCE and AEA code of practice, April 2009. http://www.ofqual.gov.uk/files/2009-04-14-code-of-practice.pdf Accessed 08/01/10.

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

Rudner, L.M. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice, 20(1)*, 16-19.

Rumm Laboratory Pty. Ltd. (2004). Interpreting RUMM2020. Part 1: dichotomous data.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8(4)*, 350-353.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika, 74(1)*, 107-120.

Stallings, W.M., & Gilmore, G.M. (1971). A note on "accuracy" and "precision". *Journal of Educational Measurement, 8(2)*, 127-129.

Stanley, G., MacCann, R., Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of teacher assessment: evidence of what works best and issues for development*. Oxford: Oxford University Centre for Educational Assessment. Report commissioned by QCA. QCA/2686.

Stearns, M., & Smith, R.M. (2008). Estimation of decision consistency indices for complex assessments: model based approaches. *Journal of Applied Measurement, 9(3)*, 305-315.

Verstralen, H., & Bechger, T. (2008). *Classification accuracy of educational tests*. Arnhem: CITO. Draft report to NAA.

Wheadon, C., & Stockford, I. (2010). Classification accuracy and consistency in GCSE and A level examinatnions offered by the Assessment and Qualifications Alliance (AQA) November 2008 to June 2009. Ofqual.

Wilmut, J., Wood, R., & Murphy, R. (1996). *A review of research into the reliability of examinations*. London: SCAA.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14(2)*, 97-116.

Wright, B.D. (1994). Part-test (subtest) measures vs whole-test measures. *Rasch Measurement Transactions, 8(3)*, 376 http://www.rasch.org/rmt/rmt83f.htm. Accessed 23/2/10.

Wright, B.D. (1995). Which standard error? *Rasch Measurement Transactions, 9(2)*, 436 http://www.rasch.org/rmt/rmt92n.htm Accessed 18/2/10.

Yen, W.M., & Fitzpatrick, A.R. (2007). Item Response Theory. In R. L. Brennan (Ed.), *Educational Measurement.* (pp. 111-153). ACE/Praeger series on higher education.

Section 2 – Marker-related variability

2.1 Introduction

Variability in examination outcomes attributable to the markers is probably the aspect of reliability of most concern to the public. Occasionally a school will write to complain if they think that an exam question was not on the syllabus (in the specification). Complaints that the grade boundaries have been set in the wrong place are even rarer, and tend to arise at a whole-system level, as in the 'crisis' of 2002 (Tomlinson, 2002; Baird 2007). But the vast majority of result enquiries and appeals from examinees or their schools are about marking reliability. It seems that people can more readily accept that they might have got a different result if they had done the exam on a different day, in different conditions, or with different questions, but they do not like the idea that the performance they actually produced on the day might have been worthy of a better (it is usually this way round!) mark or grade than the one it received.

Agreement between examination markers can be seen as just a particular case of the general topic of inter-rater agreement, which arises in many fields of human endeavour, including for example medical diagnosis, job interviewing, and consumer satisfaction . The research literature is correspondingly large, and it is beyond the scope of this report to summarise or synthesise it. For a comprehensive review that focuses on the issues that arise in marking of GCSE and A-level type examinations, see Meadows & Billington (2005).

2.1.1 Conceptualising marker agreement

There are many more ways of conceptualising and quantifying marker reliability than there are for test score reliability. There does not seem to be a standard use of terminology or statistical indicators, which can sometimes make it difficult to compare results from different studies. Some of the issues involved in conceptualising marker reliability, and in choosing an appropriate statistic to quantify it, are discussed in Bramley (2007). He suggested using 'agreement' as the most general (high-level) term. This then raises the question 'agreement with what?' which has a different answer in different situations, depending what kind of examinee response is being marked.

In the case of an entirely objective (e.g. multiple-choice) question where there is an unambiguously correct answer then the agreement is between the marker's mark and the correct mark. In this scenario Bramley (ibid) suggested the use of the term 'accuracy' to describe marker agreement.

Most exam questions in GCSEs, and many in A levels, could be described as 'short-answer questions', requiring from one word to around three lines of text, or a calculation, or labelling of a diagram, etc. Rules or guidelines for determining the mark to give to the examinees' answers to such questions are specified in the mark scheme, a document that is created in tandem with the question paper and finalised by senior examiners after consideration of a sample of examinees' responses once the examination has been taken 'live'. A variety of ways of classifying and coding these questions and mark schemes has been proposed (see for example Massey & Raikes, 2006; Bramley, 2008; Suto & Nadas, 2009; Black, Suto & Bramley, submitted). For most question / mark scheme / examinee response combinations it is usually possible to say unambiguously what the correct mark should be (as an extreme example if the examinee has not written an answer, the correct mark is zero!) In such cases, the term 'accuracy' is still an appropriate term to use for marker agreement.

For questions requiring longer responses or essays the mark scheme is more likely to indicate examples of valid points that might be made, or more general descriptors of different levels of achievement, sometimes with exemplification. Bramley (2007) suggested simply using the general term 'agreement' for such questions, bearing in mind that sometimes a degree of interpretation of the mark scheme is required and that an examiner's interpretation might legitimately differ from that of the senior examining panel. 'Agreement' here now means

'Agreement with what the Principal Examiner (or senior examining panel) has specified to be the correct mark, or would specify to be the correct mark if they were to mark the response'.

Bramley (ibid.) suggested reserving the term 'reliability' for situations where either the classical or IRT conceptions of reliability are appropriate, i.e. where reliability is conceived as a ratio of 'true' variance to total (true + error) variance. Again, descriptions of these approaches are outside the scope of this report. Johnson & Johnson (2009) give an excellent overview of how Generalizability Theory can be used to conceptualise and quantify marker reliability, and Myford & Wolfe (2003, 2004) show how marker reliability is dealt with from a Rasch perspective. It is worth noting, however, that in both these approaches, there is no 'correct' mark. In the CTT (and GT) approach the 'true' score is an average over replications, in this case replications across markers. In the Rasch approach the emphasis is on estimating the examinee's location on the latent trait as accurately as possible, making best use of the information from all the markers (raters, judges) in order to do this.

> *"Thus accurate measurement depends not on finding the one "ideal" judge but in discerning the intentions of the actual judges through the way in which they have replicated their behaviour in all the ratings each has made."* (Linacre, 1994, p41).

Both GT and IRT approaches require multiple raters (markers) in order to estimate the necessary variance components for a reliability coefficient. In examinations like GCSEs and A levels, it is simply not feasible to have multiple markings of all the examinees' answers in an operational setting. As Meadows & Billington (2005, p58) state: "Awarding bodies struggle to recruit enough examiners to mark scripts once, let alone twice". Investigations into marker reliability that would yield the necessary information to calculate reliability coefficients have thus tended to take place as separate research exercises.

2.1.2 Quantifying marker agreement

The way marker agreement is defined has implications for the choice of indicator used to quantify it. In general quantifying marker agreement by a single number loses information, particularly for longer, more subjectively marked answers, because markers can differ in a number of ways (Bramley, ibid.):

- their interpretation of the latent trait (i.e. what is better and what is worse)
- severity / leniency (a systematic bias in the perceived location of the responses on the trait)
- scale use (a different perception of the distribution of responses on the trait)
- erraticism (the extent to which their marks contain random error).

For example, methods of quantifying agreement that rely on correlation (for example the correlation between two markers marking the same set of scripts) will be sensitive to the first and fourth of these, but not the second and third.

From a measurement perspective, the ideal scenario for quantifying marker agreement is when two or more markers mark the same piece of work[17] without knowledge of what marks the other markers have given to it (referred to as 'blind' double or multiple marking). The marks can then be treated as independent in the statistical sense which is usually an assumption of most common methods of analysing agreement.

If the total score for each examinee is conceived of consisting of a 'true score' plus a random error component attributable to the marker, then the correlation between the scores from each marker is an estimate of reliability, analogous to the correlation of scores on two parallel tests (see Section 1). This estimate of reliability is not sensitive to differences between the two markers in absolute level of marks, or in how spread out they are (the correlation coefficient is

[17] henceforth referred to as a 'script' – meaning the examinee's responses to the questions on the examination.

the covariance between z-scores, which are standardised to have the same mean and standard deviation).

If the question of interest is whether the two markers are interchangeable (i.e. function like two identical measurement instruments) then the correlation is not an appropriate indicator of agreement. Altman & Bland (1983) and Bland & Altman (1986) recommend plotting the differences between the two markers against the average of their marks in order to assess interchangeability. A plot like this shows whether the data appear to scatter at random around the line of zero difference (i.e. sensitivity to differences in absolute value), and whether there is any systematic relationship between difference and average mark (i.e. sensitivity to differences in spread of marks). The mean and standard deviation of the differences provide the simplest way of quantifying the disagreement.

In assessing the reliability of GCSE and A levels, both ways of quantifying agreement are of interest. Both in situations where agreement can be fairly described as 'accuracy' or is better described as 'agreement with the Principal Examiner (PE)' we are interested in discovering the extent to which the differences between the markers' marks and the correct or 'definitive' marks have a mean and SD of zero. By analogy with Cronbach's Alpha in Section 1, it is also of interest to quantify the proportion of variance in marks that can be attributed to differences among the markers.

Although it is possible to investigate marker agreement or reliability at a variety of grain sizes from the smallest part-question upwards, for consistency with the first section we will focus on marker agreement / reliability at the level of the total score on a unit/component.

2.2 Results from research studies

There is sometimes the perception that UK examination boards are either not interested in investigating marker reliability, or 'cover up' their findings (e.g. Newton, 2005b; Black, 2003; Macintosh, 2000). It is no longer true to say that little research has been done – on the contrary as can be seen below there has been a considerable amount – but it is fair to say that much research involving multiple marking has not had the calculation of a reliability coefficient for a particular examination as its primary purpose. The focus of the research has tended to be on discovering the factors affecting levels of marker agreement (e.g. Bramley, 2007; Suto & Nadas, 2008, 2009; Greatorex & Bell, 2008; Black, Curcin & Dhawan, 2010) with a view to understanding how marking reliability can best be monitored and improved (e.g. Bell, Bramley, Claessen & Raikes, 2006; Black, Suto & Bramley, submitted). For example, there has been research into the comparative agreement of markers marking on paper compared with on screen (Johnson, Nadas & Bell, 2010); the effectiveness of training or feedback (Shaw, 2002; Greatorex & Bell, 2008), the effect of different types of exemplar training script (Baird, Greatorex & Bell, 2004), the effect of different types of marker standardisation procedures (Baird et al, 2004; Raikes, Fidler & Gill, 2009), the effect of different levels of marker expertise (Suto & Nadas, 2008; Suto, Nadas & Bell in press), the consistency over time of markers (Pinot de Moira, Massey, Baird & Morrissy 2002), the effect of different models of double-marking (Vidal-Rodeiro, 2007), and comparison between different specifications in the same subject area (Fowles, 2009).

The widest ranging published analysis of marking reliability in UK examinations comes from studies carried out by Roger Murphy 30 years ago (Murphy, 1978, 1982), who covered 20 subjects at O- and A-level. In these studies a senior examiner re-marked scripts (with original marks and annotations removed) from a sample (N=100 or 200) taken from the live examination. This re-mark was then correlated with the original mark. There were higher correlations on exams containing structured, analytically marked questions than on exams containing essays, and in general the less subjective the mark scheme, the higher the correlation. The correlations ranged from 0.73 for A level English (Paper One) to 1.00 for Mathematics O-level (Paper Two).

Murphy noted that the correlations were generally higher than might have been anticipated and took this as a "commendation of the measures that are taken to standardise the marking of GCE examinations." (Murphy, 1982, p63). However, as has been shown in many places (e.g. Altman & Bland, 1983; Gill & Bramley, 2008) a large correlation can be obtained even when there are some large discrepancies between the two sets of marks. Murphy (1982) mentioned that the 'average mark change' for the subjects investigated ranged from 0.8 to 6.7, but no detailed analysis of the differences was presented, which gives some support to the idea that examination boards are reluctant to put this sort of information into the public domain.

More recently, Newton (1996) carried out a study of marking reliability in GCSE Mathematics and GCSE English using a similar design to that of Murphy, but with two senior examiners instead of one re-marking the sampled scripts. For Mathematics, the correlations were all very high – from 0.992 to 0.997. Between 30% (Higher Tier) and 50% (Foundation Tier) of all scripts received exactly the same mark on re-marking, and 'virtually all' of the differences between original and re-mark were within 4 marks. For English, the correlations ranged from 0.81 to 0.95. The majority of original / re-mark differences were within seven marks, and around 90% of the differences were within 7 marks on the Foundation tier and within 12 marks on the Higher tier. Once again, little information about the distribution of differences was presented in tables or graphs.

In a study for which the main focus was marker agreement at item level, Massey & Raikes (2006) also reported correlations at whole script level resulting from blind double-marking by three or four markers of around 300 scripts in single units/components of five different examinations, including three A levels. Again the results were similar to those of Murphy and Newton – A-level Sociology had an average correlation of 0.866, A-level Economics 0.745, and A-level Chemistry 0.992. No information about the distribution of differences was presented.

Both Murphy and Newton investigated the difference between blind re-marking (where the second marker is unaware of the marks and annotations of the first marker) and non-blind re-marking, where the marks and annotations of the first marker are visible. Murphy (1979) found that being aware of the marks of the first marker raised the correlation from 0.87 to 0.94 in an un-named O-level examination. This corresponded to a decrease in average absolute mark difference from 5.51 marks to 2.74 marks on an examination with a maximum possible mark of 80[18]. (It should be noted that different scripts were marked in each condition, but there was no reason to believe that this invalidated the findings since they had been sampled in the same way). In Newton's study, all the English re-markers could see the original annotations, but not the marks awarded. In the Mathematics study, for a sub-set of the scripts the original annotations were not removed. There was a small (0.15 mark) difference in the expected direction (i.e. a smaller difference for scripts with annotations visible) but this was not statistically significant.

If a script is marked more than once this raises the issue of how to reconcile the two (or multiple) marks to arrive at a single mark for an examinee. Vidal Rodeiro (2007) compared three double-marking models in two GCSE examination components (Classical Greek and English). The re-marks were compared with the original marks. In Classical Greek, the correlation rose from 0.954 to 0.994 when comparing blind re-marking with non-blind remarking. This corresponded to a decrease in average absolute mark difference from 2.16 to 0.67 marks (the maximum mark for the paper was 60). In English the correlation rose from 0.695 to 0.935, corresponding to a decrease in average absolute mark difference from 4.49 to 1.84 marks (the maximum mark for the paper was 40). Vidal Rodeiro's study was unusual in reporting the distribution of obtained differences. In the Classical Greek, 78% of the blind re-marks were within ±3 marks of the original mark, but in the English only 62% of the blind re-marks were within ±4 marks of the original mark, on a paper with a much lower maximum mark.

[18] The maximum mark was not given in the text but inferred from the scale of the axes in Figure 1 in Murphy (1979).

2.3 Results from live monitoring in paper-based marking system

In the paper-based system of marking examinations, which applied to all examinations until around 2005, monitoring the quality of marking is carried out by a hierarchical sampling-based procedure where a Team Leader (TL) is responsible for monitoring the quality of the marking by the Assistant Examiners (AEs) in their team. This monitoring is achieved by the TL re-marking a sample of each of their team's allocation of scripts, at one or more points in the marking process. In OCR, the TL records their own mark and the mark of the AE for each script sampled from that examiner on a specially designed form (F1)[19]. TLs sample scripts from members of their marking team on three occasions:

- Standardisation sample – these scripts should cover the range of attainment. In most cases this should contain photocopied scripts which are common to all examiners for the paper.
- Batch 1 – Each examiner sends 20% of their apportionment (once they have completed 40%) to their team leader who selects a sample from this. Team leaders should re-mark at least 10 scripts.
- Batch 2 – Team Leader contacts examiner after receiving Batch 1 sample, telling them which centres to include in their final sample. This sample should include scripts from at least 2 centres and contain at least 40 scripts.

The Regulator's Code of Practice, (e.g. Ofqual, 2009) gives further details.

As far as we are aware there is no published information systematically describing or summarising the discrepancies between TL and AE marks arising from this sampling process across different subjects[20]. Because of the large number of examination units/components processed in this way in a given examination session, and because the information on the F1 forms is not recorded electronically (in OCR at least), this data is not readily available for analysis. However, the study reported in Bramley (2008) used data from a large one-off data collection exercise that sampled from this routine monitoring of live examinations. The study reported on marker agreement at question level associated with different features of the questions and mark schemes, and did not focus on differences at the level of the whole scripts. The information from this study at script level is reported for the first time below, after the data collection process has been described in some detail.

The aim was to capture data from as many examiners as possible within the subjects sampled (see Table 2.1 below). This was done by identifying a sample of examiners whose F1 forms would be collected, then obtaining a sample of the physical scripts from the examinees listed on each form. The guidelines for this sampling are explained below. Once the scripts had been obtained, temporary staff were hired to key in the marks awarded by AE and TL, at part-question level.

The aim was to cover two question papers at GCSE level and two at A-level from each of the large-entry subject areas. The analysis reported here used data from 22 OCR units/components taken in the June 2006 examination session. The following sampling plan was agreed to be the most effective use of the resources available for the study (2400 scripts was the approximate limit that could be sampled):

Samples of F1 scripts were drawn that met all the following conditions (see the examples that follow for clarification):

- A minimum of 100 scripts from each unit/component;
- At least five scripts from each AE sampled;
- An identical number of scripts from each AE sampled;

[19] Actually in OCR the form is known as the SEM (Standardisation of Examiner Marking) form, but that acronym would cause unnecessary confusion in this report!

[20] Pinot de Moira et al (2002) reported on a large study of A level English taken from live monitoring, but only presented the non-blind mark re-mark correlation (0.97) and some multilevel modelling analysis – not the simple distribution of differences.

- Only scripts from the last batch were used. (Some components e.g. in science, maths may give examiners feedback after Batch 1 sampling, with an instruction to review their previous marking);
- At least two AEs from every team on the panel were sampled;
- An identical number of AEs was sampled from each team;
- The number of AEs sampled was maximised (i.e. more examiners was deemed better than more scripts per examiner, providing the minimum of five scripts was not breached).

Example A: A small-medium entry subject with, say, 10 AEs in two teams of five. In this case all 10 AEs would be sampled, with at least 10 scripts per examiner. This would maximise the number of examiners and meet the requirements for total number of scripts (minimum 100) and identical numbers of scripts per examiner.
Example B: A medium entry subject with, say, 19 AEs in three teams: 2×6 AEs, 1×7 AEs. The best solution here would be to use 18 examiners, excluding one at random from the group of seven. 6 scripts per AE would be sampled (giving 108 scripts in total). This would meet the requirement of maximising the number of examiners whilst having the same number of AEs from each team. The criteria regarding numbers of scripts would also be met.
Example C: a large entry subject with, say 66 AEs in eleven teams of six. The minimum solution would be to sample two examiners from each team, and select five scripts per examiner (110 scripts).

When an examiner without a Batch 2 F1 was encountered, scripts from their Batch 1 sample were used instead. This was to avoid biasing the sample by excluding such examiners - for example, if there was no Batch 2 sample because the Batch 1 sample was so good that the TL did not take a Batch 2 sample, or because the TL ran out of time and only did a Batch 2 sample for the more unreliable examiners, then we would have excluded the better examiners from the sample. The resulting keyed data set was very large because it contained a record for each examinee × item × AE/TL mark. A rough calculation based on 22 units/components, with an average of 100 examinees per unit/component and 30 items per unit/component gives 60,000 records.

The subjects and units involved are shown in Table 2.1, along with information about the maximum mark for the paper, the number of whole questions and sub-questions (items), the distribution of differences between AE and TL, the mean and standard deviation of these differences. For purposes of comparison with other work on marker reliability that has only considered correlations we also provide the correlation between examinees' total scores from the AE and the TL, although the interpretation of these correlations is somewhat problematic since several TLs and many AEs are represented in each correlation.

The AE/TL correlations were all very high, ranging from 0.999 to 0.964, and compare favourably with the results cited in Section 2.1. However, it is important to note that in this sampling-based monitoring system, the second marker (TL) is not marking 'blind' – they can see the original marks and annotations written on the script by the AE. As seen in Section 2.1, this is likely to make an appreciable difference to the agreement statistics. Nonetheless, the very high correlations shown in Table 2.1 are as high or higher than the correlations obtained in the non-blind conditions in the research cited in section 2.1, which would tend to support, or at least not run counter to, any claim that marking of public examinations has become more reliable over the years.

In all of the units/components sampled, more than 50% of marks were within ± 1 mark of the TL's mark. In all but two subjects (GCSE Media Studies and A2 History) more than 90% of the marks were within ±4 of the TL's mark. The distribution of the mark differences between the two markers in Table 2.1 illustrates well how some individual discrepancies can look alarmingly large, even when the overall correlation is above 0.95.

Table 2.1: Live marking, paper-based hierarchical monitoring system (data from June 2006). Examination units/components and marker agreement statistics.

							% of examinees with this difference between AE and TL									
Type	Unit/component was part of	Tier	Paper total	# Qs	# items	# scripts	<-7	-7 to -5	-4 to -2	-1 to +1	+2 to +4	+5 to +7	>+7	Mean diff.	SD diff.	Corr.
GCSE	Media Studies	Both	60	4	11	82	2.4	7.3	22.0	58.5	8.5		1.2	-0.85	2.50	0.981
GCSE	History	n/a	50	6	6	171		0.6	5.8	69.6	17.0	4.7	2.3	0.80	2.22	0.964
GCSE	Design and Tech.	Found.	50	5	34	105	1.0		4.8	84.8	9.5			0.17	1.29	0.986
GCSE	Mathematics	Inter.	100	23	52	99			9.1	89.9	1.0			-0.12	0.88	0.998
GCSE	Biology	Higher	100	10	52	94			12.8	71.3	16.0			0.05	1.45	0.995
GCSE	Chemistry	Higher	100	10	50	119		0.8	15.1	62.2	20.2	1.7		0.24	1.72	0.992
GCSE	Physics	Higher	100	10	39	92		2.2	10.9	82.6	3.3		1.1	-0.24	1.70	0.995
GCSE	Geography	Found.	90	6	68	119		2.5	16.8	71.4	8.4	0.8		-0.32	1.76	0.992
GCSE	German (Reading)	Found.	50	44	44	144				100.0				-0.01	0.19	0.999
GCSE	English Language	Found.	63	3	5	109		3.7	17.4	67.9	7.3	2.8	0.9	-0.28	2.32	0.977
GCSE	English Literature	Higher	30	12	12	131		1.5	13.7	82.4	2.3			-0.59	1.07	0.977
AS	Design and Tech.	n/a	54	5	47	101			25.7	71.3	3.0			-0.69	1.25	0.992
A2	Social Science	n/a	50	4	10	100	2.0	1.0	9.0	71.0	13.0	3.0	1.0	0.16	2.45	0.965
A2	History	n/a	120	30	30	68	2.9	8.8	16.2	57.4	11.8	2.9		-0.72	2.95	0.986
AS	French	n/a	80	6	28	109			1.8	92.7	5.5			0.29	1.04	0.997
AS	Geography	n/a	75	3	17	102		1.0	22.5	59.8	14.7	2.0		-0.20	1.89	0.978
AS	English Literature	n/a	60	16	32	99	3.0	1.0	18.2	67.7	7.1	1.0	2.0	-0.36	2.63	0.968
A2	Media Studies	n/a	90	9	9	99		2.0	18.2	55.6	17.2	4.0	3.0	0.45	2.68	0.987
A2	Biology	n/a	60	5	23	115		0.9	13.0	66.1	19.1	0.9		0.10	1.71	0.981
A2	Chemistry	n/a	60	5	25	78			11.5	80.8	7.7			-0.18	1.24	0.992
AS	Physics	n/a	60	7	24	96			4.2	95.8				-0.11	0.84	0.997
A2	Mathematics	n/a	72	9	20	109		1.8	7.3	75.2	14.7	0.9		0.18	1.68	0.986

Key: #Qs= Number of questions on the paper, #items=Number of part-questions on the paper, #scripts=Number of examinees' scripts used, Found.=Foundation tier, Inter.=Intermediate tier, AE=Assistant Examiner, TL=Team Leader, diff.=difference (AE mark minus TL mark), corr.=Pearson correlation between AE and TL.

On average, the differences were close to zero – in other words there did not seem to be any systematic bias across the different subjects for AEs to be more severe or lenient than the TLs. In all except five units/components, the absolute value of the mean difference was less than 0.5 marks. In four of the five exceptions to this, the mean difference was negative, suggesting that where there was a bias, AEs tended to be more severe than the TL.

The number of scripts sampled was relatively low, so conclusions about the levels of agreement that would be observed if it were possible for the TL to re-mark every single script can only be very tentative. If the SD of the differences is taken as an estimate of the SEM arising from marker variability, this ranges from approximately 1 mark to 3 marks per unit/component – a small value compared with the SEM derived from Cronbach's Alpha – but again it must be emphasised that this SEM is likely to be a serious underestimate because of the lack of independence of the re-mark from the original mark.

It should be pointed out that the data in Table 2.1 comes from single units/components of assessments, so the effect of marker differences on the final grade based on the aggregate of several units would be considerably less, as shown in Section 1 and also by Gill & Bramley (2008). Furthermore, this data came from the live monitoring of marking, and would have been used in the quality control process to give feedback to markers, reassign marking, re-mark poorly marked work, etc. This means that the marks on the scripts used in this research were not necessarily the final marks awarded to the examinee at the end of the process.

It is possible that not all TL-sampled scripts would have been available for this study – for example if they were involved in result enquiries or appeals processes. It is likely that such scripts would have formed a very small proportion of the total, but we might expect such scripts to show more evidence of discrepancies between AE and TL. If this is true, then the statistics reported here would overestimate very slightly the true levels of agreement.

This live setting gave the advantage of no possible artefacts (e.g. time lags, the need for extra or special training, the use of photocopied scripts) which might be introduced in a specialised 'research' setting. On the other hand, it removed the opportunity for experimental control of the scripts and examiners that were sampled. We relied on the fact that the sample of units was large and representative of written papers in general qualifications.

Although we know that the TL marking was non-blind, we do not know how each TL approached their second-marking task. Some may have seen it as a reviewing/endorsing task; others as a re-marking task. This might have varied between (and possibly within) units/components. Nevertheless, the fact that the data in this study came from the routine monitoring process in a live examination session means that the information on marker agreement obtained accurately reflects the 'real' information available for marker monitoring and quality control in the paper-based system.

2.4 Results from live monitoring in the on-screen marking system

The system for monitoring on-screen marking is very different from that for monitoring marking in the paper-based system[21]. Seed scripts, for which the 'definitive' mark on each item has been established by a panel of senior examiners, are inserted into each AE's marking allocation at an approximate rate of 1 in every 20. Because the allocation of all scripts to be marked is randomised (i.e. scripts are not batched together by centre or part-centre as in the paper-based system), AEs cannot tell which scripts are the seed scripts. TLs are able to review, on-screen, all the marks given to *all* scripts (i.e. not just the seeded ones) marked by members of their team. They can also view reports showing how the marks given to the seeded scripts compare with the definitive marks. If the discrepancies on the seed scripts fall outside agreed 'tolerances' (which vary according to the type of question paper and the total mark range) then corrective action is taken. This might involve feedback to the marker on specific misunderstandings or misapplications of the mark scheme, and/or asking them to re-visit and if necessary re-mark certain scripts, or in extreme cases preventing the marker from continuing marking.

From the point of view of comparing marker agreement statistics in the on-screen system with the paper-based system, there are several important differences to be aware of:

- the second-marking is blind, because the AEs are not aware of the definitive marks;
- the same seed scripts are marked by all markers;
- the seed scripts are marked at consistent intervals throughout the marking process rather than at specific points in the process;
- the definitive mark is established by the PE and senior examiners, not the TL (in the paper-based system the TL's mark is not referred to as the definitive mark).

The analysis reported below used data from seed scripts used by OCR in monitoring on-screen marking in June 2009. We were able to extract seed data for 276 units/components of the total of 287 marked on screen. Of these, 259 had used seeding at the level of the whole script. 17 had used seeding at part-script level because the examination was marked at part-script level – for instance in some GCSE Science papers, the Biology, Chemistry and Physics questions were allocated to different markers.

The number of seed scripts used to monitor marking in any given unit/component varied considerably – from 1 to 25. The majority of units/components (190) used between 10 and 25 seed scripts. In general, fewer seed scripts tended to be used on units/components with very few markers (sometimes there was just one marker apart from the PE).

Because markers are not required to 'sign up' to mark the same number of scripts, different markers receive different sized allocations of scripts to mark. Given an approximate 'seeding rate' of 1 in 20 scripts, this means that some markers might not mark all the possible seed scripts, while others with a larger allocation might mark some more than once. The allocation mechanism ensures that seed scripts are allocated in the same fixed order for all markers, appearing at random within each block of 20 scripts. When a marker has marked all the seed scripts that they are eligible to mark they will cycle through them again in the same order. This should ensure that in general there is not a difference of more than one in how often a given marker has marked different seed scripts. In other words, they should have marked some once and some not at all, or some once and some twice, or some twice and some three times, etc. However, the 'eligible to mark' criterion disturbs this pattern in some cases. TLs are not eligible to mark seed scripts which they had provisionally marked prior to standardisation. Also, no marker should mark scripts from centres where they have an 'interest' (e.g. because they are a teacher there), and this includes seed scripts.

[21] The process described here is what happens in Scoris™, the system for on-screen marking used by OCR in partnership with RM. The system is continually being revised to take account of research findings, practical experience, and technical innovation. The specific operational details described in this section referred to the system as it was in June 2009.

Tables 2.2 to 2.4 illustrate the kind of scenarios we were dealing with. The rows list the seed scripts s1 to sN and the columns give the markers m1 to mN[22]. The entries show how often each marker marked each seed script. The column totals show how many seed scripts were marked by each marker. Multiplying this number by 20 thus gives an approximation of the number of scripts in total that would have been marked by that marker. The row totals show how many times each seed script was marked. The sum of the row or column totals in the bottom right-hand corner is the number of what we refer to later as 'marking events' at the script level.

In Table 2.2 there were only two markers other than the PE. The fact that each marker saw a different subset of seed scripts suggests that each had provisionally marked some of them prior to standardisation. In Table 2.3 there were 20 seed scripts, but of these, five were only allocated to one or two markers, because only they had a large enough allocation of scripts to get to the last few seed scripts. Table 2.4 suggests that each marker had what seems like an incredibly high marking load if the column totals represent 1/20th of their marking. Further investigation revealed that this unit was clerically marked at a marking centre.

Table 2.2: Example 1 - number of times each seed script was marked by each marker.

Script	m1	m2	Total
s1	0	3	3
s2	0	3	3
s3	0	2	2
s4	0	2	2
s5	3	0	3
s6	3	0	3
s7	3	0	3
s8	3	0	3
s9	4	0	4
s10	0	2	2
Total	16	12	28

Table 2.3: Example 2 - number of times each seed script was marked by each marker.

Script	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	Total
s1	0	1	0	0	0	0	0	1	0	0	2
s2	0	1	0	0	0	0	0	0	0	0	1
s3	1	1	1	1	1	1	1	1	1	1	10
s4	1	1	1	1	1	1	1	1	1	1	10
s5	0	1	0	0	0	0	0	0	0	0	1
s6	1	2	1	1	1	1	1	1	1	1	11
s7	1	1	1	1	1	1	1	1	1	1	10
s8	0	1	0	0	0	0	0	0	0	0	1
s9	1	2	1	1	1	1	1	1	1	1	11
s10	1	1	1	0	1	1	1	1	1	1	9
s11	1	1	0	1	0	0	0	1	0	0	4
s12	1	0	1	1	1	1	1	1	1	1	9
s13	1	0	1	1	1	1	1	1	1	1	9
s14	1	1	1	1	1	1	1	1	1	1	10
s15	1	1	1	0	1	1	1	1	1	1	9
s16	1	1	1	1	1	1	1	1	1	1	10

[22] The scripts and markers are presented in serial order according their identifiers within the system, so there is no significance in the ordering in tables 2.2 to 2.4.

Script	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	Total
s17	1	1	1	1	1	1	1	1	1	1	10
s18	1	2	1	1	1	1	1	1	1	1	11
s19	0	1	0	0	0	0	0	1	0	0	2
s20	1	1	1	0	1	1	1	1	1	1	9
Total	15	21	14	12	14	14	14	17	14	14	149

Table 2.4: Example 3 - number of times each seed script was marked by each marker.

Script	m1	m2	m3	m4	m5	m6	m7	m8	Total
s1	4	3	7	6	6	7	6	6	45
s2	4	4	7	6	6	7	7	0	41
s3	3	3	6	6	5	6	6	5	40
s4	3	3	6	6	5	7	6	5	41
s5	3	3	6	6	5	6	6	4	39
s6	4	4	7	6	6	7	7	0	41
s7	3	3	7	6	5	7	6	4	41
s8	4	4	7	6	6	7	7	0	41
s9	3	3	7	6	5	7	6	6	43
s10	4	4	7	6	6	7	6	0	40
s11	3	3	6	6	5	7	6	5	41
s12	4	4	7	6	6	7	6	0	40
s13	4	4	7	6	6	7	7	0	41
s14	3	3	7	6	5	7	6	5	42
s15	4	4	7	6	6	7	7	0	41
s16	4	4	7	6	6	7	6	6	46
s17	4	4	7	7	6	7	7	0	42
s18	3	3	7	6	5	7	6	5	42
s19	4	4	7	6	6	7	6	0	40
s20	4	4	7	6	6	7	7	0	41
Total	72	71	136	121	112	138	127	51	828

There is obviously a large number of ways in which the marker agreement data from the seed scripts on each unit/component could be analysed. Using the data from Example 2 above, four different graphical summaries are presented in Figures 2.1 to 2.4 below.

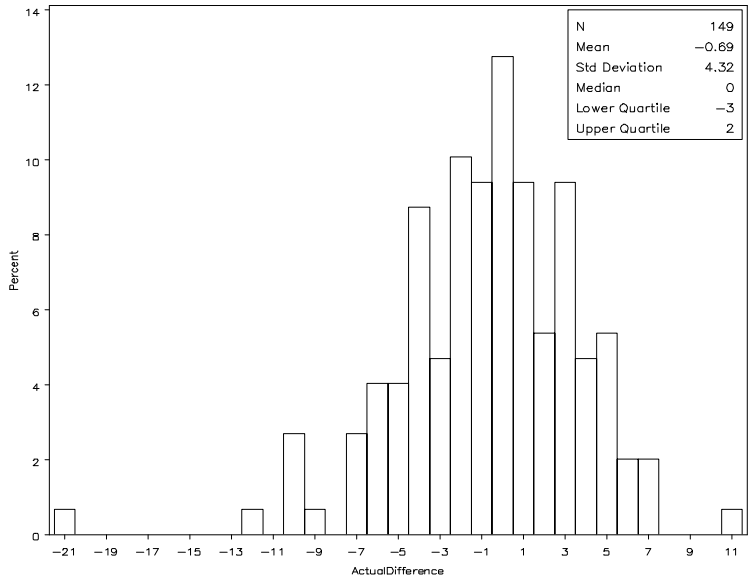

Figure 2.1: Distribution of differences between marker's mark and definitive mark at whole script level. N is the number of 'marking events' as defined above.

Figure 2.1 shows that the median (and mode) of the differences was zero, and gives a good visual impression of the spread. 50% of the differences were within a 5 mark range around zero, from -3 to +2. The SD of the difference distribution, 4.32, can be taken as a rough indicator of a SEM around the definitive mark, but bearing in mind the heterogeneous allocation of seeds to markers shown in Table 2.3. The fact that the mean difference is less than zero shows that on average, the markers were slightly severe – i.e. tending to give lower marks than the definitive mark.

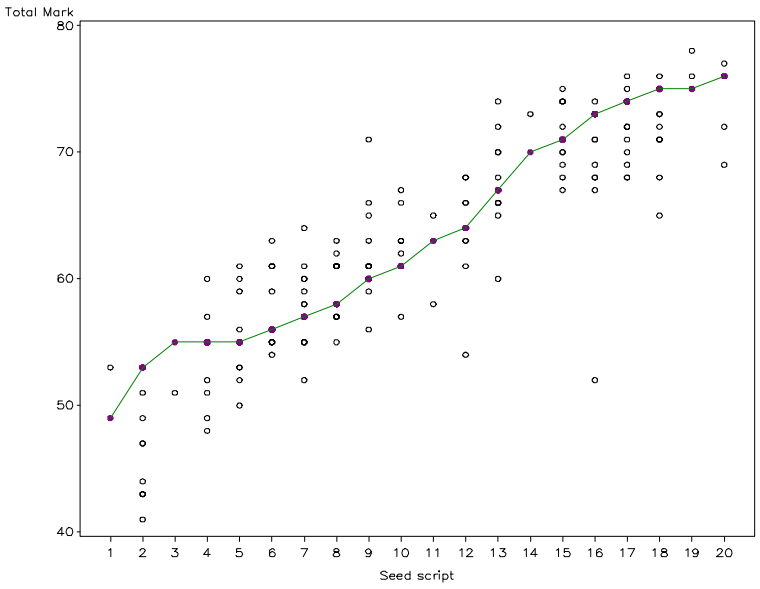

Figure 2.2: Distribution of marks awarded to each seed script. The green line shows the definitive marks[23].

In Figure 2.2 the seed scripts are arranged in ascending order of definitive mark, and the marks awarded to each are shown. A plot like this is fairly easy to interpret, showing essentially the same information as Figure 2.1 but with the different scripts identified. This plot could identify

[23] In Figures 2.2. and 2.3 the seed scripts are presented in order of definitive mark, so s1 to s20 do not correspond to the same scripts as in Table 2.3.

particular scripts that had been ‘difficult to mark’ for whatever reason – for example those with a wide spread, and/or those with marks skewed to either side of the definitive mark, such as Script 2 where all the awarded marks were less than the definitive mark.

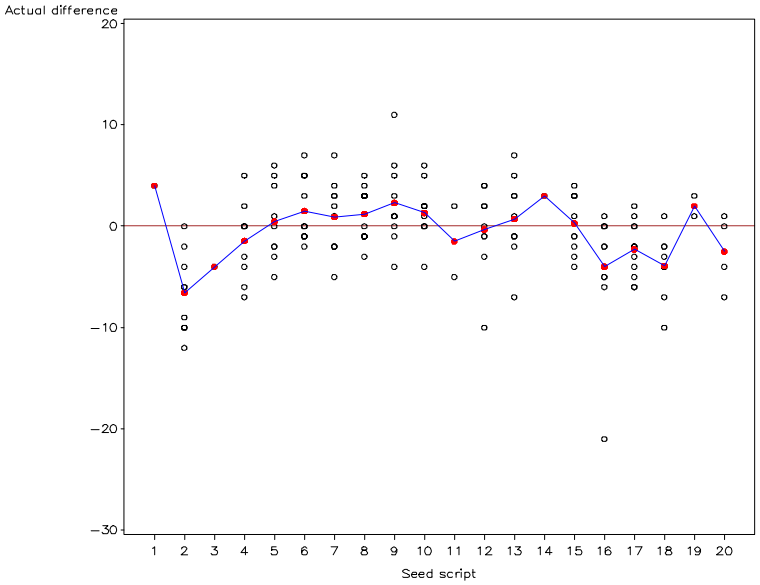

Figure 2.3: Distribution of differences between awarded and definitive mark for each seed script. The red dots connected by the blue line show the mean difference for each seed script.

Figure 2.3 effectively presents the same information as Figure 2.2, but de-trended by subtracting the definitive mark from each script’s awarded mark. The horizontal line at zero therefore now corresponds to the definitive mark, and the blue line shows whether the mean difference for each seed script was positive or negative. The seed scripts are in order of definitive mark, as in Figure 2.2. A plot like this makes it easier to identify consistent patterns of severe or lenient marking, and to see if there is any relationship between marker agreement and script ‘quality’, as defined by the definitive mark. For example, if the blue line connecting the means sloped downwards from left to right, being above the zero line for scripts with low definitive marks and being below the zero line for script with high definitive marks, this could suggest that markers were not using the full range of marks, being lenient with the worse scripts and severe with the better ones.

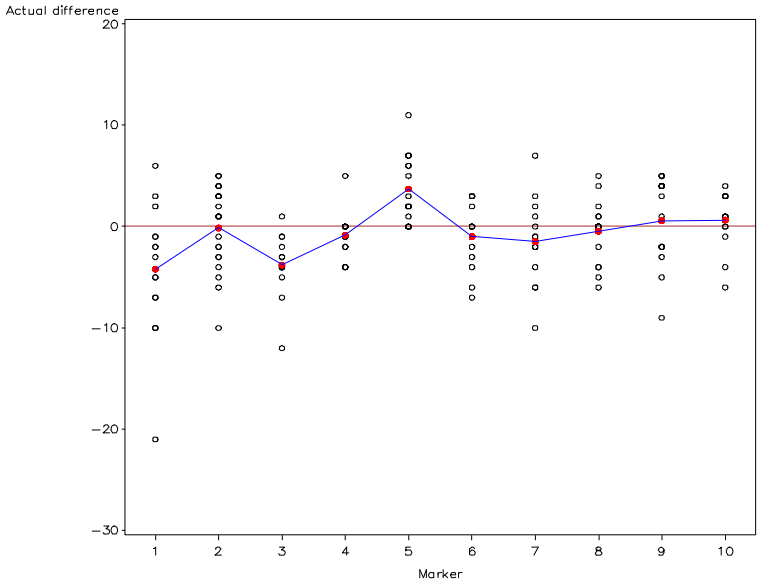

Figure 2.4: Distribution of differences between awarded and definitive mark for each marker. The blue line shows the mean difference for each marker.

Figure 2.4 presents exactly the same 'dots' as Figure 2.3, but this time grouped by marker (the markers are not arranged in any particular order). A plot like this can obviously identify markers whose marking was out of line with the definitive mark, and in which direction. For example, it appears from Figure 2.4 that marker 5 tended to be lenient and marker 3 tended to be severe, but in general all markers were neither lenient nor severe. It would obviously be too space-consuming to present this kind of graph for every unit/component in a report such as this, but as with the test-related reliability graphs and tables presented in Section 1, they could be produced in batch jobs and used operationally by examination boards for quality control purposes.

Table 2.5 below shows summary statistics for a representative selection of units/components, aiming to include most of those that were mentioned in Section 1 of this report. We have included some more robust quantile statistics along with the means and standard deviations in case some extreme outlying differences were not 'genuine' (e.g. resulting from mis-keys, system errors, data corruption, etc.).

It is important to emphasise some contrasts between Table 2.5 and Table 2.1 earlier:

- the units/components in Table 2.5 were marked on-screen, not on paper;
- the units/components marked on-screen tend not to include long answer or essay questions where there might be expected to be less marker agreement;
- the double (multiple) marking in Table 2.5 was done blind, in contrast to Table 2.1;
- the number of 'marking events' in Table 2.5 is much more variable than in Table 2.1;
- in Table 2.5 there is multiple marking, including some same-marker-same-script repetitions, of a smaller number of seed scripts than in Table 2.1 where every script represented a separate 'marking event' and was thus only marked twice.

Bearing these differences in mind, the correlations in Table 2.5 are generally slightly lower than in Table 2.1, as might be expected from blind versus non-blind multiple marking. However, the absolute values of the correlations all seem very high, and are in general higher than those reported in the studies by Murphy (1979, 1982). (But since they are not the same subjects, this is perhaps not too meaningful a comment).

The mean and median differences in Table 2.5 are all close to zero, with the exception of the two GCE Home Economics units, where the four markers seemed severe on average compared with the definitive mark. One of these units (G002) and one of the GCE Business studies units (F292) had the two highest SDs and interquartile ranges of differences. Interestingly, these units

had also been flagged in Section 1 as being among those with relatively high levels of test-related variability in analyses which used data from marking of all scripts, not just the small subset designated as seeds. One implication of this is that, as suggested in the discussion in Section 1.5, low values of Cronbach's Alpha can (sometimes) be attributed to unreliable marking.

Table 2.6 gives a more detailed breakdown of the distribution of differences, for the same units/components as shown in Table 2.5. In contrast to the test-related variability in Section 1, where an average SEM was derived indirectly from Cronbach's Alpha, here we have the actual distribution of differences. This meant it was possible to calculate the proportion of marks that would fall within a grade band for a 'definitive mark' in the middle of a grade band[24]. This is shown in the final column of Table 2.6. This is arguably the best way to make comparisons among the different components, because allowing for the grade bandwidth tends to take into account differences in paper total, and mark differences that would result in grade changes would be more important to the examinee than those that would not.

It is a potential limitation of the data that some seed scripts were marked more than once by the same marker, because in theory if subsequent markings were not independent of the first mark then this could lead to inflated reliability estimates. However, we would suggest that this is likely to have only a very small, perhaps negligible, impact on the data as a whole. This is because of the large number of scripts that would have intervened between each repeated encounter of a seed script by a single marker. With 15 seed scripts in use, spaced approximately 20 scripts apart, a marker would encounter the same seed script for a second time only after marking $\approx$300 intervening scripts. Many markers' allocations would not even be this large. Our data suggested that where markers were seeing the same seed script several times, they were clerical markers working at a marking centre, marking questions with highly prescriptive mark schemes. It is doubtful that in these circumstances any seed scripts would be sufficiently memorable to lead to a statistically detectable effect on repeated encounters.

[24] For bandwidths that were an even number of marks wide, an average was taken based on the two possible mid-points. We assume that (as with the classical SEM in Section 1) the same approximate 'error' distribution applies at all points of the mark range.

Table 2.5: OCR June 2009 – summary of distribution of differences between definitive and awarded mark for seed scripts in 21 selected units/components.

Unit/component was part of	# scripts	# markers	# items	Paper Total	# MEs	Corr.	Mean	SD	Median	IQR	5th pctl	95th pctl
GCSE Psychology (Found.)	15	7	30	80	81	0.908	-0.85	3.46	0	4	-6	4
GCSE Psychology (Found.)	19	5	34	80	76	0.941	-0.34	2.67	0	3	-6	4
GCSE Psychology (Higher)	20	10	29	80	149	0.878	-0.69	4.32	0	5	-7	5
GCSE Psychology (Higher)	20	9	27	80	144	0.946	-1.39	3.69	-1	4.5	-8	5
GCE Biology	12	8	17	45	118	0.982	-0.28	1.25	0	1	-2	2
GCE Chemistry	20	6	27	60	134	0.988	0.04	1.52	0	2	-2	3
GCE Chemistry	23	6	37	60	118	0.986	0.30	1.06	0	1	-1	2
GCE Chemistry	25	4	23	45	123	0.972	0.41	1.23	0	1	-1	3
GCE Physics	18	2	25	45	44	0.957	0.09	1.34	0	2	-1	2
GCSE Science (Higher)	20	8	22	42	828	0.998	-0.01	0.28	0	0	0	0
GCSE Additional Science (Found.)	20	19	32	42	594	0.990	0.00	0.49	0	0	-1	1
GCSE Additional Science (Found.)	20	7	29	42	727	0.992	-0.06	0.47	0	0	-1	1
GCE Accounting	10	5	9	80	69	0.997	-0.62	2.10	-1	3	-4	3
GCE Accounting	19	8	18	120	100	0.987	-0.79	3.35	-0.5	5	-6	4
GCE Business Studies	14	15	8	60	214	0.895	-1.30	4.93	-1	7	-10	7
GCE Business Studies	16	26	11	90	368	0.890	0.38	6.33	0	8	-11	10
GCE Critical Thinking	12	50	17	75	771	0.958	-0.29	3.33	0	3	-6	5
GCE Electronics	10	2	29	90	28	0.969	0.68	2.92	1	4	-4	4
GCE Home Economics	10	4	14	100	34	0.931	-2.71	4.56	-3	4	-12	5
GCE Home Economics	10	4	16	100	33	0.888	-2.33	7.16	-2	10	-17	8
GCSE Additional Maths	18	25	49	100	481	0.982	0.00	1.44	0	2	-2	2

Key: # items= number of part-questions on the exam paper. # MEs= number of 'marking events' where a seed script was marked by a marker. Includes repeated markings of the same seed script by the same marker. Corr. = Pearson correlation between awarded mark and definitive mark across all marking events. IQR=Inter-quartile range. N^{th} pctl=N^{th} percentile.

Table 2.6: OCR June 2009 – distribution of differences between definitive and awarded mark for seed scripts in 21 selected units/components.

			% of MEs with this difference between marker and definitive mark							
Unit/component was part of	Paper Total	# MEs	<-7	-7 to -5	-4 to -2	-1 to +1	+2 to +4	+5 to +7	>+7	% within grade bandwidth
GCSE Psychology (Found.)	80	81	1.2	16.1	27.2	32.1	18.5	2.5	2.5	54.3
GCSE Psychology (Found.)	80	76	.	9.2	19.7	50.0	17.1	4.0	.	68.4
GCSE Psychology (Higher)	80	149	4.7	10.7	23.5	31.5	19.5	9.4	0.7	83.9
GCSE Psychology (Higher)	80	144	6.9	6.3	30.6	38.2	11.8	6.3	.	83.3
GCE Biology	45	118	.	1.7	12.7	78.8	6.8	.	.	78.8
GCE Chemistry	60	134	.	.	17.9	65.7	15.7	0.8	.	91.0
GCE Chemistry	60	118	.	.	3.4	85.6	11.0	.	.	98.3
GCE Chemistry	45	123	.	.	1.6	82.1	16.3	.	.	82.1
GCE Physics	45	44	.	2.3	2.3	84.1	11.4	.	.	62.5
GCSE Science (Higher)	42	828	.	.	0.4	99.4	0.2	.	.	100.0
GCSE Additional Science (Found.)	42	594	.	.	0.5	97.8	1.7	.	.	97.8
GCSE Additional Science (Found.)	42	727	.	.	0.7	98.9	0.4	.	.	99.4
GCE Accounting	80	69	.	2.9	23.2	58.0	15.9	.	.	88.4
GCE Accounting	120	100	3.0	10.0	25.0	36.0	24.0	2.0	.	96.0
GCE Business Studies	60	214	10.8	15.4	20.1	26.6	15.4	7.0	4.7	33.1
GCE Business Studies	90	368	9.2	12.5	12.8	23.9	18.8	10.6	12.2	43.5
GCE Critical Thinking	75	771	3.0	6.7	17.0	49.2	16.7	6.9	0.5	69.0
GCE Electronics	90	28	.	3.6	17.9	35.7	39.3	3.6	.	78.6
GCE Home Economics	100	34	14.7	8.8	35.3	23.5	11.8	5.9	.	48.5
GCE Home Economics	100	33	21.2	12.1	18.2	21.2	9.1	9.1	9.1	28.8
GCSE Additional Maths	100	481	0.2	0.8	9.6	78.2	11.0	0.2	.	99.8

Key: ME = ‘marking event’ where a seed script was marked by a marker.

In order to make comparisons with the results in Section 1, we treated the SD of the differences as an approximate SEM attributable to marker variability. Figure 2.5 plots the test-related SEM calculated via Cronbach's Alpha in Section 1 against this marker-related SEM for the 254 units/components where we had a value for each index. The red line is an identity line, so points above the line represent units/components where the test-related SEM was higher than the marker-related SEM. Figure 2.5 makes it clear that, for the kind of unit/component marked on-screen, different questions contribute more to score unreliability than different markers, because the vast majority of points are above the line. Putting it another way, for this kind of unit/component, an examinee would be more likely to receive the same score (or grade) if their performance in the examination was marked by a different marker than if they took a parallel paper with a different set of questions – not a particularly surprising finding. There was also a general trend for units/components with a higher test-related SEM to have a higher marker-related SEM (Spearman's rho = 0.78).

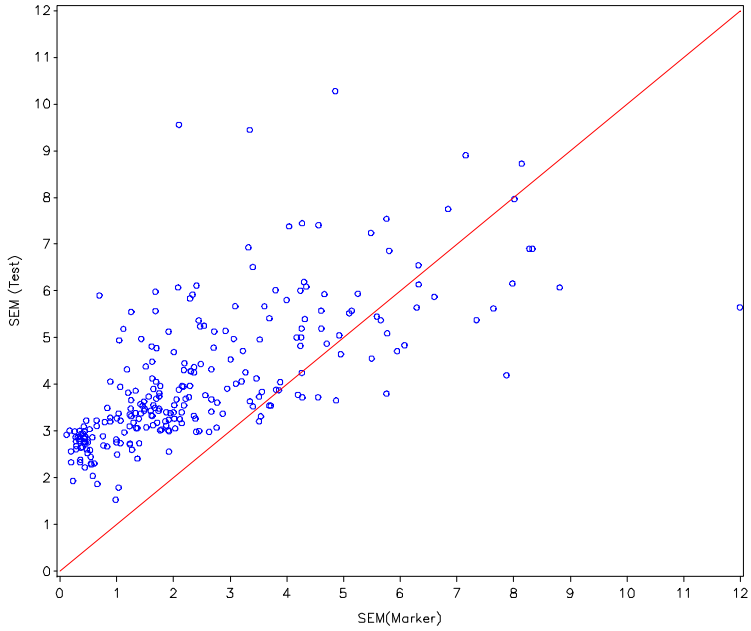

Figure 2.5: Plot of test-related SEM against marker-related SEM.

As in Section 1, we calculated the grade bandwidth:SEM index (using the marker-related SEM instead of the test-related SEM) in order to make fairer comparisons among the units/components.

Table 2.7: Summary of distribution of the Bandwidth:SEM index for marker reliability across all units/components.

	N	Mean	SD	Min	LQ	Median	UQ	Max
GCE	84	2.60	1.22	0.81	1.56	2.40	3.37	6.74
GCSE	170	6.30	6.37	0.89	1.98	3.22	9.06	34.18
All	254	5.08	5.54	0.81	1.88	2.90	5.35	34.18

Table 2.7 shows that the GCSE units/components had a higher (i.e. better) value of the index than the GCE units, but the difference between median and mean for the GCSE units/components shows the skew of the distribution of the index, arising from the long tail of high values. This long tail can be explained by many of the GCSE units/components comprising very objectively-marked questions with virtually no marker-related variability. The overall value of 1.88 for the lower quartile of the bandwidth:SEM index can be interpreted as suggesting that for 75% of the units/components, an examinee with a 'true score' in the middle of a grade band

would have had at least a 94% probability of receiving a score in the same grade band from a different marker.

Table 2.8: Summary of distribution of median difference between awarded and definitive mark at whole script level across all units/components.

	N	Mean	SD	Min	LQ	Median	UQ	Max
GCE	85	0.09	1.63	-6.0	0	0	0	6
GCSE	170	-0.13	0.66	-3.5	0	0	0	2
All	255	-0.06	1.09	-6.0	0	0	0	6

Table 2.8 is interesting because it shows that, across all units/components, not only was the median of the median difference between awarded and definitive mark equal to zero, the lower and upper quartiles were too. This is good evidence that, for the type of examination currently marked on screen, systematic severity or lenience of all the markers relative to the definitive mark is relatively rare. (The min and max values of -6 and +6 in Table 2.8 reduce to -3 and +2 when units/components with fewer than 30 marking events are excluded).

Although the focus of this report is on marker agreement at the level of the whole script, it is perhaps relevant to report a brief summary of agreement at the item level. We define a ‘marking event’ now to be an instance of each item on a seed script being marked by a marker, and use exact agreement between marker’s mark and definitive mark as our index of reliability.

Table 2.9: Exact agreement between marker and definitive mark at item level.

	# units / components	# items	# seed items	# marking events	# exact agreement	% exact agreement
GCE	88	2384	24546	404557	350354	86.60
GCSE	188	5120	102993	2588459	2442668	94.37
All	276	8104	127539	2993016	2793022	93.32

Note: # items is the total number of part-questions on the units/components. # seed items is the sum of # items multiplied by the number of seed scripts used per unit/component.

Table 2.9 shows that of nearly 3 million marking events at the item level, over 90% had exact agreement between marker and definitive mark. The percentage was lower for GCE units/components than for GCSEs because they have relatively fewer low-tariff objective questions. See Massey & Raikes, 2006; Bramley, 2008; Suto & Nadas, 2009; and Black et al. 2010 for further discussion of the factors affecting exact agreement at the item level.

Table 2.10: Distribution of % exact agreement between marker and definitive mark at item level across all units/components.

	N	Mean	SD	Min	LQ	Median	UQ	Max
GCE	88	81.80	14.46	44.91	73.26	87.05	92.06	100
GCSE	188	88.79	12.96	47.77	80.62	95.30	98.98	100
All	276	86.56	13.82	44.91	78.99	91.48	97.99	100

The values in Table 2.10 are lower than those in Table 2.9 because of the variation in number of marking events across units/components. It tended to be the case that the units/components with the largest entries also had the most short, objective items and hence the largest numbers of marking events – and that these marking events were more likely to result in exact agreement than those on other units/components.

2.5 Variance components analysis of seed script marks

It is perhaps of interest to try to quantify in some way whether the differences between awarded mark and definitive mark arise mainly because markers differ systematically in their levels of severity, or because seed scripts differ systematically in how severely or leniently they are marked. In terms of Figures 2.3 and 2.4 above, is the blue line connecting the means closer to horizontal for markers, or for seed scripts?

Variance components analysis[25] decomposes the total variance of the differences, where each marking event contributes one observation, into a row effect (between-seeds) and a column effect (between-markers). If the same marker has marked the same scripts more than once, then it is also possible to separate out a marker-script interaction from the residual error. Both seed scripts and markers are treated as 'random effects' – that is, they are assumed to be sampled at random from a wider potential population of interest (this is certainly a reasonable assumption for seed scripts, less so for markers). Rather than estimating a particular 'effect' for each seed script and each marker, a variance components analysis just estimates the distribution (mean and variance) of these effects, on the following assumptions:

- there is a linear relationship between the effects and the dependent variable;
- means are zero and variances are constant across the range of the dependent variable;
- random effects are independent and identically distributed;
- residuals are independent and identically distributed.

Table 2.11: OCR June 2009 – estimated variance components of differences between definitive and awarded mark for seed scripts in 21 selected units/components.

Unit/component was part of	Total estimated variance	Seed component (%)	Marker component (%)	Interaction component (%)	Residual error (%)
GCSE Psychology (Found.)	12.12	38	5		57
GCSE Psychology (Found.)	7.65	15	22		62
GCSE Psychology (Higher)	19.29	26	23	26	25
GCSE Psychology (Higher)	14.31	25	20	51	4
GCE Biology	1.55	18	5	0	77
GCE Chemistry	2.31	28	2	0	70
GCE Chemistry	1.22	26	12	32	29
GCE Chemistry	1.63	29	9	12	51
GCE Physics	2.33	6	47		47
GCSE Science (Higher)	0.08	0	0	7	93
GCSE Additional Science (Found.)	0.24	17	0	5	77
GCSE Additional Science (Found.)	0.22	3	0	2	95
GCE Accounting	4.87	29	20	13	39
GCE Accounting	11.38	51	8	0	42
GCE Business Studies	25.08	45	9	0	47
GCE Business Studies	42.13	33	21	15	32
GCE Critical Thinking	11.91	23	5	26	47
GCE Electronics	9.13	63	0		37
GCE Home Economics	23.19	41	0	27	32
GCE Home Economics	49.76	28	17	49	6
GCSE Additional Maths	2.27	26	3	32	40

[25] See http://faculty.chass.ncsu.edu/garson/PA765/variancecomponents.htm for a relatively straightforward overview, and http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/statug_varcomp_sect001.htm for computational details.

The results in Table 2.11 should be taken as illustrative only. It is very unlikely that all the assumptions listed above would be met. The analysis 'designs' were frequently unbalanced, with some cells (see for example Tables 2.2 to 2.4) containing more observations than others, and often with some cells empty. In particular, it is probably unrealistic to regard the interaction term as separate from the residual error. Where there is no entry in the 'interaction' column, this is because the analysis did not converge and was re-run without an interaction term.

However, even bearing all these caveats in mind, it is probably still reasonable to make two general observations:

- Most of the variance in differences between awarded and definitive mark is not attributable to systematic differences among the markers or the seed scripts (the sum of variance components for 'interaction + error' was greater than the sum of variance components for 'seed + marker' in 16 out of 21 cases in Table 2.11);
- Systematic differences among seed scripts are relatively larger than systematic differences among markers (the variance component for 'seed' was larger than the variance component for 'marker' in 18 out of 21 cases in Table 2.11).

The first of these observations supports the claim of Hutchison & Benton (2009) that non-systematic inter-marker variability is the main source of between-marker difference, implying that one explanation for low values of Cronbach's Alpha could be unreliable marking.

Once again it is important to emphasise that the units/components marked on screen are not representative of all examination units/components. In particular, those such as English Literature or History, which involve longer essays, are not included. The above generalisations, if justifiable, could only apply to the type of unit/component currently marked on-screen.

2.6 Discussion

In contrast to test-related variability, it is much more feasible to carry out the replications necessary to quantify marker-related variability. Because of the expense of organising blind double- or multiple-marking, most existing estimates are based on research exercises. However, with the advent of on-screen marking, statistics arising from blind double-marking are now becoming routinely available. These statistics are admittedly based on the very small number of 'seed' scripts per examination used for monitoring marking, but can be calculated across the full range of examinations marked on screen, thus giving a useful snapshot of the whole system.

It is a striking feature of most of the published research on marking reliability of GCSEs and A-levels that simple information on the distribution of differences at whole-script level between mark and re-mark, or between mark and 'definitive' mark, has generally not been presented in any detail. Instead, the preference of researchers seems to have been to report either simple correlation coefficients, or to delve immediately into more complex statistical analyses that are correspondingly more difficult for the layman to interpret.

We recommend therefore, that while more complex statistical analyses are certainly desirable and potentially enlightening, examination boards should consider ways that the 'raw data' arising from blind multiple-marking of seed scripts can be most effectively presented. We have given some suggestions of possible formats in Figures 2.1 to 2.4 and Tables 2.5 and 2.6. Our analysis has been based largely on the mean and SD of the distribution of differences between awarded mark and definitive mark at whole-script level, because this is both intuitively fairly straightforward, and is more closely related to classical test theory. We recognise that in some research reports the statistic of choice has been the mean absolute (unsigned) difference between awarded mark and definitive mark, and that, as a single statistic, this is more concise than the mean + SD of the actual (signed) differences. However, using the mean absolute difference loses the information about relative severity and lenience, and is less easy to connect to classical test theory.

Our main findings from analysing the seed script data are:

- Marker-related variability in scores was relatively less than the test-related variability reported in Section 1;
- On average, markers tended to be neither severe nor lenient compared to the definitive mark;
- Systematic differences in severity among markers made the smallest contribution to score variability – less than systematic differences among seed scripts and much less than random (non-systematic) error.

It is important to note that these findings only apply to the kind of unit/component currently marked on screen. At the time of writing, this does not include units/components requiring predominantly extended responses or essays. However, research into the validity and reliability of on-screen marking for essay-based examinations is underway (e.g. Johnson et al. 2010) and the early indications are that, at least in terms of reliability, there is little difference between on-screen and paper-based marking. This suggests that in the future essay-based examinations may also be marked on screen and it will be interesting to see whether the above findings generalise.

In the meantime, the information in Table 2.1 gives a picture of the level of agreement in the non-blind second marking carried out in live monitoring in the paper-based marking system. The extended-response and essay subjects (e.g. English Literature, History and Media Studies) do indeed seem to show the most variability, but the differences in question paper totals and the fact that the second marking is non-blind make it difficult to make comparisons with other subjects, or with the blind marking of seed scripts.

One interesting issue that has not yet been discussed is the extent to which the small sub-set of scripts chosen to be seeds are representative of all scripts. It is probably safe to assume that they are reasonably representative, although it is likely that scripts with very low totals, consisting of mostly blank responses, would not be used. However, given that the purpose of the seed scripts is to allow monitoring of marker performance 'in real time', is there a rationale for choosing seed scripts that exhibit particular features? For example, if the PE believes that a particular question is likely to generate responses that are 'difficult to mark', they might want to select scripts that exhibit 'problematic' responses to check that the markers have understood how they are to apply the mark scheme in these cases. If 'problematic' scripts are over-represented in the seeds then the data on multiple-marking of these scripts would probably underestimate the levels of agreement that would be observed in the non-seed scripts.

A further issue is the extent to which the 'definitive' mark agreed by the panel of senior examiners for each seed item is indeed the correct mark. For the vast majority of cases it almost certainly is, but items where the modal (most frequently occurring) mark awarded by the markers differs from the definitive mark raise the possibility that the 'definitive' mark might be wrong. Black et al. (2010) and Black & Curcin (2010) introduce the concept of the 'DIMI' (definitive mark incongruent with modal mark). Their investigation of the small percentage (4.6%) of seed item responses in their study identified as DIMIs showed that in many cases, especially those requiring extended responses, it was not possible to say what the correct mark should be, and that where it was possible, the modal mark was at least as likely to be correct as the definitive mark. If, as seems likely, units/components with questions requiring longer written responses start to be marked on screen, for such units/components it may be more appropriate to talk about a 'definitive range of marks' for seed items, only considering marks that fall outside this range to indicate marking error.

Finally, it should be noted that all the data presented in this section came from live monitoring of the operational examinations while it was taking place, whether from Team Leaders inspecting samples of (paper) scripts from markers in their team, or monitoring (online) performance on seed scripts where the definitive mark had been determined prior to marking. In both cases, the monitored scripts represent only a very small proportion of the total number of scripts marked in a live examination session. The statistics on marker agreement based on this monitoring do not take into account any further 'quality control' procedures, such as scaling the marks of examiners found to be consistently severe or lenient (paper-based system only), or re-marking the scripts of markers who were stopped from marking because their work was deemed to be unsatisfactory. If school/examinees are unhappy with the results, they can make an 'enquiry about results' which may lead to a re-mark, and in extreme cases where disputes cannot be resolved there is an official appeals procedure. The aim of all these processes is to ensure that the final grades awarded to examinees reflect as accurately as possible their performance in the examination.

2.7 References for Section 2

Altman, D.G., & Bland, J.M. (1983). Measurement in medicine: the analysis of method comparison studies. *The Statistician, 32,* 307-317.

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

Baird, J., Greatorex, J., & Bell, J.F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice, 11(3)*, 331-348.

Bell, J. F., Bramley, T., Claessen, M., and Raikes, N. (2006). Quality control of marking: some models and simulations. Paper presented at the annual conference of the British Educational Research Association (BERA), University of Warwick, September 2006.

Black, P. (2003). Testing, testing: listening to the past and looking to the future. *School Science Review, 85*(311), 69-77.

Black, B., Suto, W.M.I. & Bramley, T. (submitted). The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement.

Black, B. & Curcin, M. (2010). *Group dynamics in determining gold standard marks for seeding items and subsequent marker agreement.* Paper presented at the annual conference of the British Educational Research Association (BERA), University of Warwick, September 2010.

Black, B., Curcin, M. and Dhawan, V. (2010). *Investigating seeding items used for monitoring on-line marking: factors affecting marker agreement with the gold standard marks.* Paper presented at the annual conference of the International Association for Educational Assessment (IAEA), Bangkok, Thailand, August 2010.

Bland, J.M., & Altman, D.G.1. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, (i), 307-310.

Bramley, T. (2007). Quantifying marker agreement: terminology, statistics and issues. *Research Matters: A Cambridge Assessment Publication, 4*, 22-28.

Bramley, T. (2008). *Mark scheme features associated with different levels of marker agreement.* Paper presented at the British Educational Research Association (BERA) annual conference, Heriot-Watt University, Edinburgh, September 2008.
http://www.cambridgeassessment.org.uk/ca/Our_Services/Research/Conference_Papers
Accessed 7/04/10.

Fowles, D. (2009). How reliable is marking in GCSE English? *English in Education, 43(1),* 50-67.

Gill, T., & Bramley, T. (2008). Using simulated data to model the effect of inter-marker correlation on classification consistency. *Research Matters: A Cambridge Assessment Publication, 5*, 29-36.

Greatorex, J., & Bell, J.F. (2008). What makes AS marking reliable? An experiment with some stages from the standardisation process. *Research Papers in Education, 23(3)*, 233-255.

Johnson, S., & Johnson, R. (2009). *Conceptualising and interpreting reliability*. Assessment Europe. Ofqual/10/4709.

Johnson, M., Nadas, R., & Bell, J.F. (2010). Marking essays on screen: an investigation into the reliability of marking extended subjective texts. *British Journal of Education Technology 41(5)* 814-826.

Linacre, J.M. (1994). *Many-Facet Rasch Measurement.* (2nd ed.). Chicago: MESA Press.

Macintosh, H. (2000). Review of 'How exams really work'. *Assessment in Education: Principles, Policy & Practice, 7*(2), 280-281.

Massey, A. & Foulkes, J. (1994). Audit of the 1993 KS3 Science national test pilot and the concept of quasi-reconciliation. *Evaluation and Research in Education*, *8*, 119-132.

Massey, A.J. & Raikes, N. (2006). *Item level examiner agreement.* Paper presented at the Annual Conference of the British Educational Research Association, 6-9 September 2006, University of Warwick, UK. http://www.cambridgeassessment.org.uk/ca/digitalAssets/171646_BERA06_Massey_and_Raikes.pdf Accessed 22/3/10.

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability.* Manchester: AQA.

Murphy, R.J.L. (1978). Reliability of marking in eight GCE examinations. *British Journal of Educational Psychology, 48*, 196-200.

Murphy, R.J.L. (1979). Removing the marks from examination scripts before re-marking them: does it make any difference? *British Journal of Educational Psychology, 49,* 73-78.

Murphy, R.J.L. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology, 52,* 58-63.

Myford, C.M., & Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: part I. *Journal of Applied Measurement, 4 (4)*, 386-422.

Myford, C.M., & Wolfe, E.W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: part II. *Journal of Applied Measurement, 5(2)*, 189-227.

Newton, P.E. (1996). The reliability of marking of General Certificate of Secondary Education scripts: Mathematics and English. *British Educational Research Journal*, *22*, 405-420.

Newton, P.E. (2005a). The public understanding of measurement inaccuracy. *British Educational Research Journal, 31(4)*, 419-442.

Newton, P.E. (2005b). Threats to the professional understanding of assessment error. *Journal of Education Policy, 20(4)*, 457-483.

Newton, P.E. (2009). The reliability of results from national curriculum testing in England. *Educational Research, 51(2)*, 181-212.

Ofqual (2009). GCSE, GCE and AEA code of practice, April 2009. http://www.ofqual.gov.uk/files/2009-04-14-code-of-practice.pdf Accessed 08/01/10.

Pinot de Moira, A., Massey, C., Baird, J.-A., & Morrissy, M. (2002). Marking consistency over time. *Research in Education, 67*, 79-87.

Raikes, N. Fidler, J. & Gill, T. (2009). *Must examiners meet in order to standardise their marking? An experiment with new and experienced examiners of GCE AS Psychology.* Paper

presented at the Annual Conference of the British Educational Research Association, September 2009, Manchester, UK.

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes, 8*, 13-17. http://www.cambridgeesol.org/rs_notes/rs_nts8.pdf Accessed 22/3/10.

Suto, W.M.I. and Nadas, R. (2008). What determines GCSE marking accuracy? An exploration of expertise among maths and physics markers. *Research Papers in Education, 23(4)*, 477-497.

Suto, W.M.I., & Nadas, R. (2009). Why are some GCSE questions harder to mark accurately than others? Using Kelly's Repertory Grid technique to identify relevant question features. *Research Papers in Education, 24(3)*, 335-377.

Suto, W.M.I., Nadas, R. and Bell, J. F. (in press). Who should mark what? A study of factors affecting marking accuracy in a biology examination. *Research Papers in Education*.

Tomlinson, M. (2002). *Inquiry into A level standards. Final report.* London: Department for Education and Skills. http://image.guardian.co.uk/sys-files/Education/documents/2002/12/03/alevelinquiry.pdf Accessed 22/3/10.

Vidal Rodeiro, C. (2007). Agreement between outcomes from different double-marking models. *Research Matters: A Cambridge Assessment Publication*, *4*, 28-34.

Section 3 – Grading-related variability

3.1 Introduction

The grade boundaries on each unit/component of a GCSE or GCE assessment are set by a panel of experts, following a procedure laid down by the regulator in the Code of Practice (Ofqual, 2009). Not all boundaries are set by the panel – usually only certain specified 'key boundaries' are set and the others are interpolated (or extrapolated) between them arithmetically, with rounding rules specified in the Code of Practice. In general, for GCSE units/components, the key boundaries are A, C and F; and for GCE units/components, the key boundaries are A and E.

There are several sources of information taken into account by the panel of experts, including:

- 'archive' scripts at the key grade boundary marks from previous sessions;
- information about the size and composition (e.g. type of school attended) of the cohort of examinees;
- teachers' forecast grades;
- the distribution of scores (mean, SD, cumulative % of examinees at each mark);
- at GCE, 'putative' grade distributions (grade distributions generated by matching examinees with their GCSE results and taking account of changes in the 'ability' of the cohort of examinees from a previous[26] session, as indicated by changes in the distribution of mean GCSE scores;
- experts' judgments about the quality of work evident in a small sample of scripts covering a range of consecutive marks (total scores) around where the boundary under consideration is expected to be found;
- experts' judgments about the difficulty of the question paper;
- other external evidence suggesting that the particular unit/component (or assessment as a whole) had previously been severely or leniently graded and needs to be 'brought into line' – for example with other examination boards, or with other similar subjects or specifications within the same board.

The grade boundary setting process is essentially a 'standard-maintaining' process (except in the case of completely new assessments) where the aim is to 'carry forward' the standard set on the equivalent unit/component in previous examination sessions[27]. Comparability of examination standards is a topic which is the subject of perennial debate (for a comprehensive recent review see Newton, Baird, Goldstein, Patrick & Tymms, 2007). It is probably reasonable to say that standard maintaining from one examination session to the next in the same unit/component is the least controversial context in which issues of comparability arise, and so this report will not consider any of the issues that arise in the 'standards debate'.

In what sense are reliability issues relevant to grade boundary setting? If the same approach is applied as with test-related and marker-related reliability, then the question is to what extent the outcomes of a grade boundary setting meeting, i.e. the decisions on the grade boundary marks, would be replicated in different, but appropriately similar, conditions. Examples of the kind of thing that could in principle vary are:

- the members of the expert panel making the judgments, particularly the Chair of Examiners, with whom most of the responsibility rests;
- the particular scripts scrutinised by the panel;
- the mark range of scripts scrutinised by the panel;
- the type and content of the 'external' information available to the panel.

[26] Usually this is the previous session with a cohort believed to be most similar to the current session's cohort. E.g. for a June 2009 unit, the June 2008 session might be used rather than the January 2009 session.

[27] Some tensions can arise in this process because grade boundaries are set at unit/component level, but arguably the 'standard' resides at the level of the whole assessment.

Variations in the first three of these should arguably have little or no effect on the decisions made, and this could (and perhaps should) be verified experimentally. However, creating an authentic experiment that successfully replicated the 'real time' pressures and demands of an awarding meeting would be a difficult challenge. As far as we are aware, no-one has yet attempted to do this. There is some research (e.g. Novakovic & Suto, 2010) that has investigated the replicability of the judgmental aspect of the awarding (the evaluations of the grade-worthiness of individual scripts), but this is arguably of little relevance to the question of what outcomes would be observed if all the conditions of the award meeting were replicated. Variation in the fourth factor – type and content of 'external' information – presumably ought to affect the outcome in a predictable way[28]. Again, this could be investigated experimentally, but as far as we are aware has not been.

It is clear that with the variety of sources of (potentially conflicting) evidence to be integrated, the setting of grade boundaries is not an exact science. There has been a trend in recent years (at GCE) to place most weight on the 'putative' grade boundaries, to the extent that as the new GCE specifications are examined for the first time, examination boards are required to report to the regulator any incidences where the cumulative grade distributions (at assessment level) deviate by more than a specified amount from notional 'target' values created by statistical methods.

Given that it is not possible to determine exactly what the grade boundaries 'should' be, it is of interest to investigate what the impact of slightly different decisions at unit/component level would be on the grade distributions at whole assessment level. In particular, it seems likely that the evidence for any particular grade boundary decision could support two possible boundary marks, and perhaps more. Therefore in this section we investigate the effect on assessment grade boundaries of varying the (judgmentally set) key grade boundaries on the units/components by ±1 mark. This is not really an investigation of reliability, as such – it is probably more appropriately characterised as a 'sensitivity analysis'.

3.2 A tiered, linear GCSE examination

The aggregation of unit/component grade boundaries into grade boundaries at assessment level is considerably simpler for linear assessments than for unitised assessments. The example given below is one of the most straightforward scenarios possible.

Foundation Tier examinees took Paper 1, Paper 2 and coursework. Higher Tier examinees took Paper 3, Paper 4 and coursework. The coursework was the same for both tiers and had the same grade boundaries.

In linear assessments, there are two ways of deriving the aggregate grade boundary from the component grade boundaries:
'Indicator 1' is the simple aggregate of the component grade boundaries, taking account of the weight of each component in the aggregate total. In this GCSE the two written papers each carried 40% weight and the coursework 20%, and their paper totals were in these proportions, which meant that indicator 1 could simply be obtained by adding up the grade boundary marks on the three components.
'Indicator 2' is the mark on the aggregate distribution of marks (the distribution obtained by adding together each examinee's mark on each component) where the cumulative percentage of examinees obtaining that mark corresponds most closely to the percentage obtained by taking a weighted average of the cumulative percentage of examinees at that particular boundary on each of the components.

[28] This is perhaps a controversial statement. Those who believe that grade boundary judgments should be purely 'criterion-referenced' might argue that if the judged quality of work does not change, and the judged demand of the question paper does not change, then neither should the grade boundaries – even if there is other evidence that suggests the cohort of examinees was 'better' or 'worse' by some external criterion.

The calculation of the two indicators is illustrated in Table 3.1 below.

Table 3.1: Linear GCSE calculation of aggregate indicators.

		Total	Grade C	% examinees	Ind. 1	%	Ind. 2	%
Foundation Tier	Paper 1	80	33	37.77				
	Paper 2	80	34	36.62	90	39.47	88	42.50
	Coursework	40	23	61.66				
		Total	Grade A	% examinees	Ind. 1		Ind. 2	
Higher Tier	Paper 3	80	67	17.71				
	Paper 4	80	63	18.38	164	16.55	158	23.13
	Coursework	40	34	44.03				

For the 'C' boundary at Foundation Tier, Indicator 1 is simply 33 + 34 + 23 = 90. To calculate Indicator 2, the weighted average cumulative percentage at or above the C boundary is first calculated as (0.4×37.77 + 0.4×36.62 + 0.2×61.66) = 42.09. Then the closest matching mark on the aggregate frequency distribution is found to be 88 (see Table 3.2).

For the 'A' boundary at Higher Tier, Indicator 1 is 67 + 63 + 34 = 164. The weighted average cumulative percentage at or above the A boundary is (0.4×17.71 + 0.4×18.38 + 0.2×44.03) = 23.24. Then the closest matching mark on the aggregate frequency distribution is found to be 158 (see Table 3.2).

Table 3.2: Extracts from aggregate frequency distributions at Foundation and Higher tier.

Foundation Tier			Higher Tier	
Mark	Cumulative %		Mark	Cumulative %
91	37.74		164	16.55
90	39.47		163	17.30
89	41.04		162	18.47
88	42.50		161	19.70
87	43.96		160	20.93
			159	22.41
			158	23.13

Indicator 2 is usually lower than Indicator 1 at the top end of the mark range, and higher than Indicator 1 at the bottom of the mark range. It is sometimes described as allowing for 'regression to the mean', in the sense that it is (nearly always) closer to the mean mark than Indicator 1. The difference between the indicators at a particular boundary is likely to be greater when i) the correlations between the scores on the components are lower; ii) when there are more components (because this is likely to lower the correlations); iii) when there is a greater disparity among the components in cumulative % at the boundary; and iv) when the component boundary is further from the mean.

The Code of Practice (Ofqual 2009) allows the awarding panel to choose any mark between (and including) the two indicators as the final aggregate boundary mark. The 'default' position is to take the lower of the two indicators[29]. Other evidence may suggest moving from this default position. In this case, the Foundation Tier 'C' boundary was taken at 90, and the Higher Tier 'A'

[29] The Code of Practice states "Whenever the two indicators do not coincide, the grade boundary should normally be set at the lower of the two indicator marks, unless, in the awarders' judgement, there is good reason, as a result of a review of the statistical and technical evidence, to choose a higher mark within the range spanned by the indicators." Ofqual (2009) p53.

boundary was taken at 159[30]. It is clear from Table 3.2 that at the Higher Tier, the range of choices available to the panel, *even after setting the component boundaries*, could have allowed considerable variation in the percentage of examinees achieving a grade A.

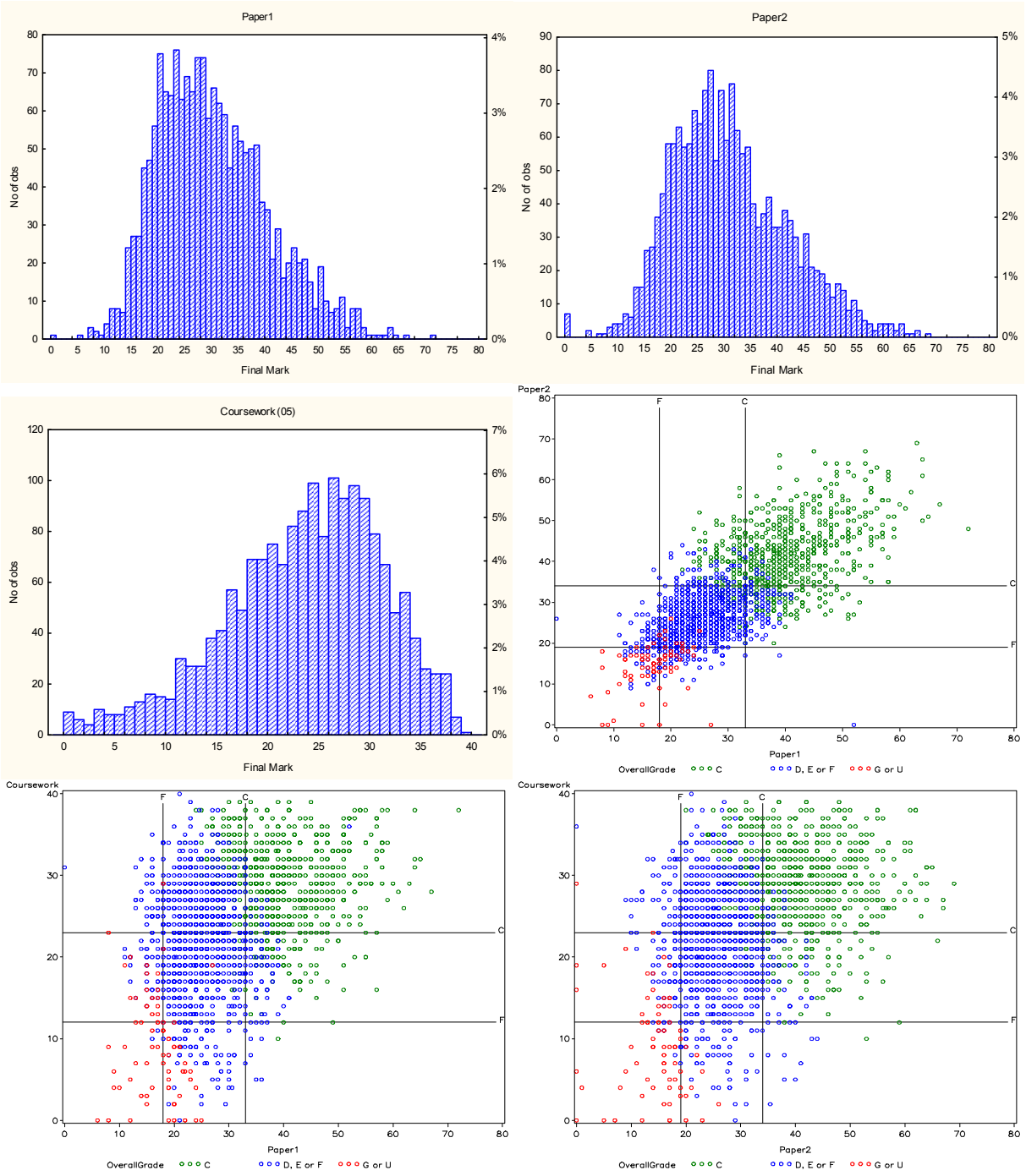

Figure 3.1: Linear GCSE Foundation tier – component score distributions and relationships between the distributions.

[30] The data used for the calculations was that available at the time of writing, which differs slightly from the data available at the time the grade boundaries were actually set 'live' in June 2009. Therefore the numerical results of the indicator calculations reported here may differ slightly from those used at the time.

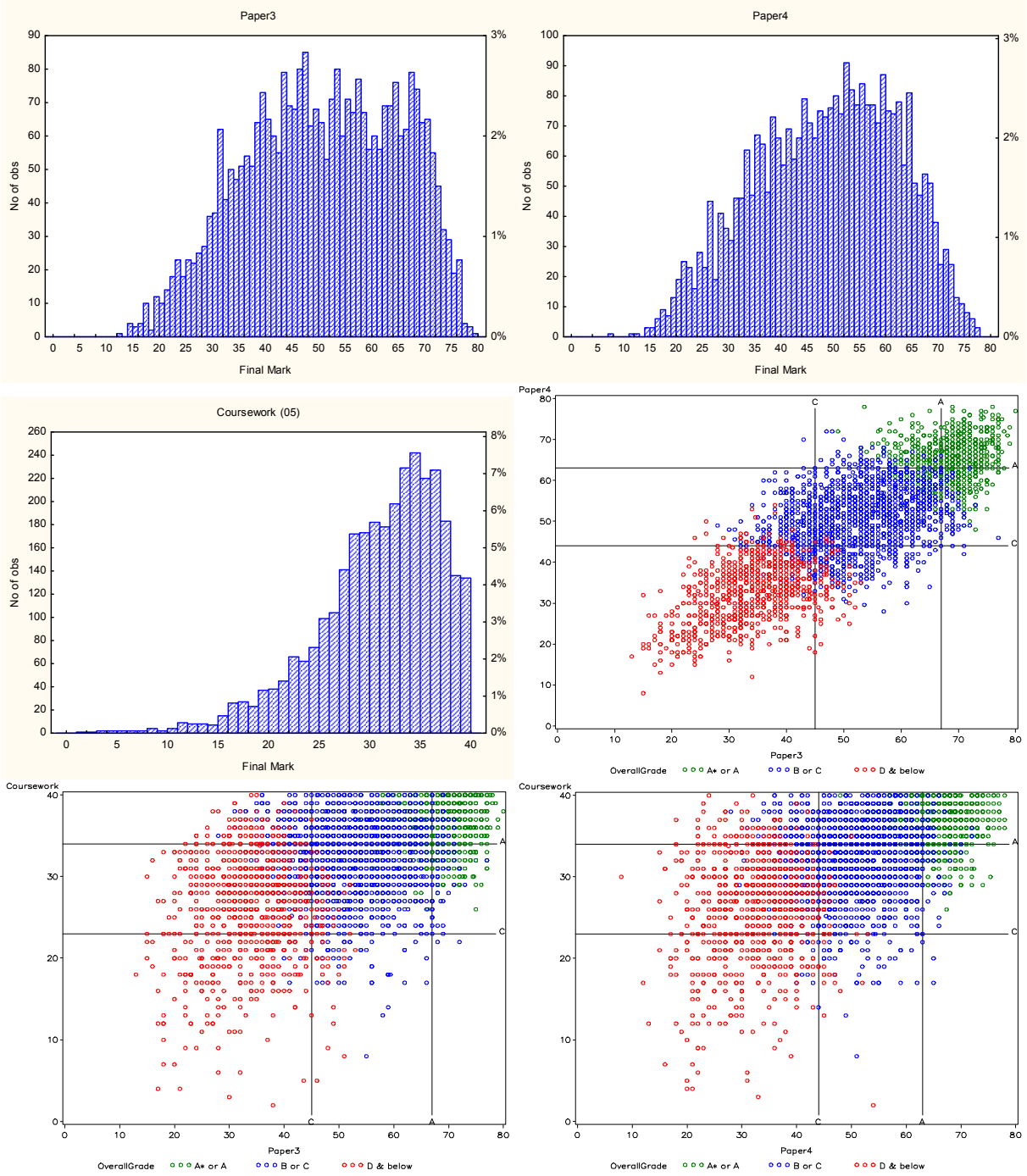

Figure 3.2: Linear GCSE Higher tier – component score distributions and relationships between the distributions.

In the scatter plots in Figures 3.1 and 3.2, the 'key' boundaries set by the panel are shown with vertical and horizontal reference lines. These are C and F on Foundation Tier, and A and C on Higher Tier. The colour coding indicates *aggregate* grade, i.e. those above or below the key boundaries on the aggregate grade distribution.

At both tiers, it is not surprising to note that performance on the written papers was more closely associated with overall grade than performance on the coursework (given that the written papers contributed 80% of the overall total). Comparing the scatter plot of scores on the two written papers in the Foundation Tier (Figure 3.1: Paper 1 against Paper 2) with the equivalent plot for

the Higher Tier (Figure 3.12: Paper 3 against Paper 4) shows the effect of the greater difference between the two indicators on the Higher Tier. On the Higher Tier there were relatively more examinees obtaining grades A and A* overall who did not exceed the A boundary on both written papers than was the case for grade C on Foundation Tier.

If Indicator 2 had not been invented[31], then the effect of changing the grade boundaries by ±1 mark on each of the three components would be easy to derive – the possible aggregate boundaries would range ±3 marks round the actual grade boundary. These possibilities are shown below in Tables 3.3 and 3.4.

Table 3.3: Foundation Tier - possible aggregate grade boundaries (Indicator 1 only).

Aggregate boundary	# combinations	Cumulative % of examinees at Grade C
93	1	34.50
92	3	35.96
91	6	37.74
90	7	39.47
89	6	41.04
88	3	42.50
87	1	43.96

Table 3.4: Higher Tier - possible aggregate grade boundaries (Indicator 1 only).

Aggregate boundary	# combinations	Cumulative % of examinees at Grade A
167	1	13.71
166	3	14.47
165	6	15.41
164	7	16.55
163	6	17.30
162	3	18.47
161	1	19.70

The column showing the number of combinations shows that there is only one way to get an aggregate boundary 3 marks lower (1 mark lower on each component), but three ways to get an aggregate boundary 2 marks lower (1 mark lower on each of the three possible pairs of two components) etc.

Tables 3.3 and 3.4 show that the possible fluctuations around the actual aggregate grade boundary could have led to fluctuations covering a range of up to 9.5 percentage points in the pass rate at grade C on the Foundation Tier, and up to 6 percentage points in the pass rate at grade A on the Higher Tier – and this without taking Indicator 2 into account.

Bringing Indicator 2 into play complicates the situation, because changing the boundary by ±1 mark will have a different effect on the weighted average cumulative percentage depending on which component's boundary is changed, and hence on the mark on the aggregate grade distribution corresponding most closely to this weighted average.

Table 3.5: Indicator 1 and Indicator 2 for the 27 possible grade C boundary combinations (Foundation Tier).

[31] Indicator 2 was used by some examining boards in the 1980s, and was enshrined in the first mandatory Code of Practice (SCAA, 1994). It will be discontinued from June 2010, when the majority of GCSEs will be unitised. Unitised assessments do not use Indicator 2, because the logical basis of the calculation requires the same examinees to have done all components (units).

Paper 1	Paper 2	Coursework	Indicator 1	Indicator 2
34	35	24	93	91
34	35	23	92	90
34	34	24	92	90
33	35	24	92	90
34	34	22	90	89
33	35	22	90	89
34	35	22	91	89
34	34	23	91	89
33	35	23	91	89
33	34	24	91	89
34	33	24	91	89
32	35	24	91	89
33	34	22	89	88
34	33	22	89	88
33	34	23	90	88
34	33	23	90	88
32	35	23	90	88
32	34	24	90	88
33	33	24	90	88
32	34	22	88	87
33	33	22	88	87
32	34	23	89	87
33	33	23	89	87
32	33	24	89	87
32	35	22	89	87
32	33	22	87	86
32	33	23	88	86

One way to summarise the information in Table 3.5 is to count the number of times each aggregate boundary mark was a possibility for the actual boundary mark, across the 27 combinations created by varying the three component boundaries by ±1 mark. This is shown in Table 3.6 below.

Table 3.6: Number of times each aggregate mark was a possible choice for the Foundation Tier grade C boundary across the 27 combinations (actual boundary in bold).

Aggregate boundary mark	Number of times a possible choice	Cumulative % (N=1913)
93	1	34.50
92	4	35.96
91	10	37.74
90	**16**	**39.47**
89	19	41.04
88	14	42.50
87	8	43.96
86	2	45.74

Table 3.7 shows the equivalent set of values for the Grade A boundary on the Higher Tier.

Table 3.7: Number of times each aggregate mark was a possible choice for the Higher Tier grade A boundary across the 27 combinations (actual boundary in bold).

Aggregate boundary mark	Number of times a possible choice	Cumulative % (N=3173)
167	1	13.71
166	4	14.47
165	10	15.41
164	17	16.55
163	23	17.30
162	26	18.47
161	27	19.70
160	26	20.93
159	**22**	**22.41**
158	15	23.13
157	9	24.30
156	4	25.50
155	1	26.41

Tables 3.6 and 3.7 show that taking Indicator 2 into account when looking at the effect of changes of ±1 mark to the component grade boundaries increases the potential variability of the aggregate grade boundary. It is interesting to note that Tables 3.6 and 3.7 raise another possibility for deriving the aggregate boundary – namely the boundary that appears the most times as an aggregate possibility, given the component boundaries, and an assumption that these might fluctuate by ±1 mark. This would have led to a mark of 89 as the Foundation Tier C boundary, and a mark of 161 as the Higher Tier A boundary.

Table 3.8: Number of times each aggregate mark was a possible choice for the Higher Tier grade C boundary across the 27 combinations (actual boundary in bold).

Aggregate boundary mark	Number of times a possible choice	Cumulative % (N=3173)
115	3	68.89
114	9	69.75
113	14	70.60
112	**13**	**71.45**
111	9	72.14
110	4	72.90
109	1	73.78

Table 3.8 shows that the variability in outcomes was lower for the C boundary on the Higher Tier. This is partly because Indicator 1 and Indicator 2 were much closer together, and partly because there were fewer examinees at this part of the aggregate distribution.

Finally, we consider the overall GCSE outcomes – i.e. the results of combining the Foundation and Higher Tiers. This is what is reported in tables of statistical outcomes (the tier of entry does not even appear on the examinee's certificate). Grade A is only available on the Higher Tier, so the possible cumulative percentage outcomes for the different possible boundaries in Table 3.9 only differ from Table 3.7 in the denominator (total number of examinees) used to calculate the cumulative percentage. However, grade C is available on both the Foundation and Higher Tiers. If we assume for the sake of illustration that the grading decisions on the two tiers are made independently[32] then the 8 possible boundaries in Table 3.6 can be combined with the 7 possible boundaries in Table 3.8 to give 56 possible overall outcomes at grade C. Instead of presenting all 56 outcomes in a similar table to Table 3.9, we treated the relative frequencies that each

[32] These decisions are not made independently - the Code of Practice specifies that the C boundary on the Foundation Tier is set first, followed by the C boundary on the Higher Tier, and the overall outcome is likely to inform both decisions.

boundary occurred within a Tier as a 'weight' and re-normalised so that the weights summed to 100, in order to generate a 'distribution' of possible overall cumulative percentage outcomes at grade C that could be shown graphically, as in Figure 3.3.

Table 3.9 Linear GCSE overall cumulative percentage outcomes with different possible aggregate grade A boundaries (actual boundary in bold).

Aggregate boundary mark	Number of times a possible choice	Cumulative % (N=5086)
167	1	8.55
166	4	9.02
165	10	9.61
164	17	10.32
163	23	10.79
162	26	11.52
161	27	12.29
160	26	13.06
159	**22**	**13.98**
158	15	14.43
157	9	15.16
156	4	15.91
155	1	16.48

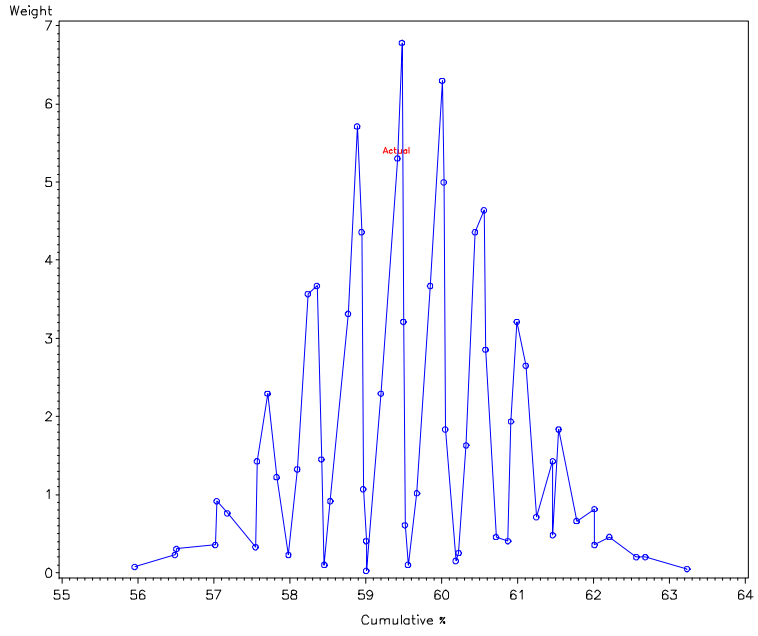

Figure 3.3: Linear GCSE: distribution of overall possible grade C outcomes.

The distribution in Figure 3.3 is very 'spiky' because some of the possible aggregate boundaries had a very low relative weight (chance of occurring, given the method used here for arriving at this value). The range of possible outcomes (percentage of examinees at or above grade C) covered 7 percentage points from 56% to 63%, with the most likely outcomes in a narrower range from about 58% to about 61%.

3.3 A 2-unit GCE AS level

This was a 'new' 2/4 unit GCE (see section 1). In June 2009 the full A level was not yet available, so examinees could only choose to 'aggregate' their unit results towards the AS qualification. The two units in question were available in both January and June 2009, so aggregating examinees could have taken their two units in any of the four possible combinations (both in January, Unit 1 in January / Unit 2 in June, Unit 2 in January / Unit 1 in June, or both in June). It is important to note for the purpose of the 'grading sensitivity analysis' carried out here that the grading decisions taken in June 2009 have no effect on examinees' results in units taken in January 2009. This means that the impact of changes to the boundaries in June 2009 depends to a certain extent on the proportion of aggregating examinees who took their units in June. Table 3.10 below shows the number of aggregating examinees taking units in June 2009.

Table 3.10: A 2-unit AS level - number of aggregating examinees taking each unit in June 2009 (total N=4792).

Unit 1 in June 2009	Unit 2 in June 2009	N	%
✗	✗	0	0
✗	✓	2013	42.01
✓	✗	0	0
✓	✓	2779	57.99

Table 3.10 shows that all the aggregating examinees took Unit 2 in June 2009, and that over half of them also took Unit 1 in June 2009. In this particular case where the specification was new, January 2009 was the only other possibility, but in the case of a long-running unitised specification, the units not taken in June 2009 could come from several previous sessions, as will be clear in sections 3.4 and 3.5.

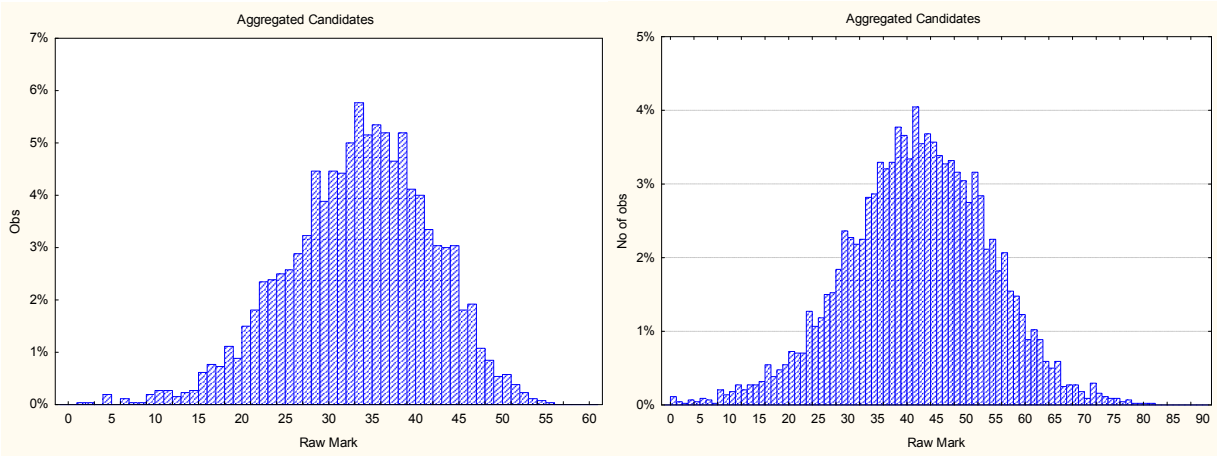

Figure 3.4: Distribution of raw scores on Unit 1 and Unit 2 in June 2009, including only those examinees requesting aggregation in June 2009.

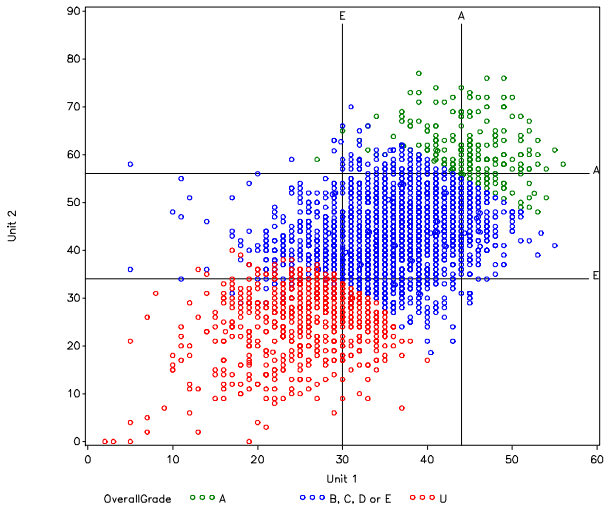

Figure 3.5: Scatter plot of raw scores on Unit 1 and Unit 2 in June 2009, including only those examinees requesting aggregation in June 2009 who had taken both units in June 2009.

Figures 3.4 and 3.5 show the separate and joint distribution of raw scores on Unit 1 and Unit 2. Figure 3.5 shows that many more examinees passed Unit 1 but did not pass Unit 2 than vice versa. The numbers getting grade A on only one of the two units were much more evenly distributed.

The Uniform Mark Scale (UMS) used to derive grades on unitised assessments was discussed in Section 1 and that discussion will not be repeated here. Operationally, the aggregation process is complex, because of the flexibility offered to examinees in when they take units, and which units they take. Essentially, when aggregating a unitised assessment to arrive at a final grade for the examinee, the following steps are taken:

- the examinees requesting aggregation are identified;
- the UMS scores obtained by each examinee on each unit of the assessment are obtained;
- the best total UMS score for each examinee is determined after adding up the UMS scores on the units for each valid combination of units they have taken (some examinees may have re-taken units, or taken more than the minimum number required);
- an overall aggregate grade is determined by comparing the best aggregate UMS score to the UMS grade boundary mark.

For this AS level, the process is fairly straightforward. Unit 1 had a maximum of 80 UMS points, with the grade A boundary on the raw mark scale mapping to 64 UMS points (80% of the maximum) and the grade E boundary on the raw mark scale mapping to 32 UMS points (40% of the maximum). Unit 2 had a maximum of 120 UMS points with the A and E boundaries mapping to 96 and 48 UMS points respectively. The aggregate grade A boundary on the UMS scale was therefore 64+96=160, and the aggregate grade E boundary was 32+48=80. The maximum UMS points available for each unit usually remains fixed for the lifetime of the assessment. Note that unitised assessments are still 'compensatory' like linear assessments, in that it is not necessary to get a grade A on both units to get a grade A overall. An examinee obtaining 60 UMS points on Unit 1 and 100 UMS points on Unit 2 would have the necessary 160 UMS points for an overall grade A.

Table 3.11 below shows the effect of varying the grade A boundaries on the raw mark scales of Unit 1 and Unit 2 in June 2009 by ±1 mark on the overall percentage of examinees with a grade A[33].

[33] For these calculations we used the data that was on the system at the time the grade boundaries were set in June 2009. Subsequent changes to the data (e.g. from data arriving late, or changes made as a result of enquiries or appeals) might mean that the figures reported here do not exactly match the final outcomes.

Table 3.11: A 2-unit AS level – effect of varying June 2009 unit grade A boundaries on overall % of examinees at grade A.

Unit 1 June 2009	Unit 2 June 2009	Cumulative % aggregate grade A (N=4792)
-1	-1	12.69
0	-1	12.19
+1	-1	11.60
-1	0	11.19
0	**0**	**10.66**
+1	0	10.14
-1	+1	9.89
0	+1	9.35
+1	+1	9.02

It is clear from Table 3.11 that, as expected, changing the boundaries on Unit 2 was more influential than changing them on Unit 1, because all of the aggregating examinees had taken Unit 2 in June 2009 whereas only 42% of them had taken Unit 1 (see Table 3.10). On the other hand, this would have been mitigated by the fact that Unit 1 had a shorter raw mark scale with a greater percentage of examinees on each mark, implying that changes of ±1 mark to the raw grade boundaries would affect more examinees on this unit. The resultant variability in cumulative percentage at grade A was approximately 3.6 percentage points.

Table 3.12: A 2-unit AS level – effect of varying June 2009 unit grade E boundaries on overall % of examinees at grade E.

Unit 1 June 2009	Unit 2 June 2009	Cumulative % aggregate grade E (N=4792)
-1	-1	83.99
-1	0	83.99
-1	+1	83.99
0	-1	83.99
0	**0**	**83.99**
0	+1	83.99
+1	-1	83.91
+1	0	83.91
+1	+1	83.91

In contrast, Table 3.12 shows that at grade E there was very little effect of changing the boundaries by ±1 mark – only 4 examinees were affected, and this only by a rise in boundary on Unit 1.

3.4 A 3-unit GCE AS level

This specification was one of the 'old' 3/6 unit GCEs which had been running for several years in June 2009. The set of examinees requesting aggregation for the AS qualification contained those with units taken as far back as June 2007.

Table 3.13: A 3-unit AS level – number of aggregating examinees taking each unit in June 2009 (total N=2898).

Unit 1 in June 2009	Unit 2 in June 2009	Unit 3 in June 2009	N	%
×	×	×	1339	46.20
×	×	✓	256	8.83
×	✓	×	327	11.28
×	✓	✓	139	4.80
✓	×	×	318	10.97
✓	×	✓	178	6.14
✓	✓	×	139	4.80
✓	✓	✓	202	6.97

It is interesting to note that nearly half of the examinees requesting aggregation had not taken any of the AS units in June 2009. This could have been for a number of reasons. For example, some of them may previously have done badly on A2 units and decided just to aggregate for the AS qualification.

The numbers in Table 3.13 should make it clear that changing the grade boundary marks by ±1 on each of the June 2009 units would only be likely to have a relatively small effect on the overall grade distribution for the examinees aggregating in June 2009, for the simple reason that most of those examinees had taken some or all of their units in other sessions. However, for the sake of illustration, we made these changes and carried out the aggregations. To keep the number of combinations to a manageable size, we only considered the grade A boundary, we treated each of the three different options within Unit 3 as a single option, and took no account of the fact that each of the three options in that unit was made up of two components. In other words, we varied the boundaries on the within-option aggregate mark scales by ±1 mark as a single block. This gave 27 different combinations to consider. The effect on the distribution of overall grades for the aggregating examinees is shown in Table 3.14.

Table 3.14: A 3-unit AS level – effect of varying June 2009 unit grade A boundaries on overall % of examinees at grade A (actual outcome in bold).

Unit 1 June 2009	Unit 2 June 2009	Unit 3 June 2009	Cumulative % aggregate grade A (N=2898)
-1	-1	-1	33.61
-1	-1	0	33.51
-1	-1	+1	33.13
-1	0	-1	33.40
-1	0	0	33.23
-1	0	+1	32.82
0	-1	-1	33.33
0	-1	0	33.16
0	-1	+1	32.71
0	0	-1	33.16
+1	-1	-1	33.13
0	**0**	**0**	**32.92**

Unit 1 June 2009	Unit 2 June 2009	Unit 3 June 2009	Cumulative % aggregate grade A (N=2898)
+1	-1	0	32.92
0	0	+1	32.51
+1	-1	+1	32.51
-1	+1	-1	33.02
-1	+1	0	32.78
-1	+1	+1	32.30
+1	0	-1	32.85
+1	0	0	32.64
+1	0	+1	32.23
0	+1	-1	32.82
0	+1	0	32.51
0	+1	+1	31.95
+1	+1	-1	32.51
+1	+1	0	32.23
+1	+1	+1	31.64

Table 3.14 shows that the variability of aggregation outcomes was only ≈2 percentage points. No one unit seemed to be noticeably more influential than the others in terms of the effect on the aggregation outcome of changes to its grade boundary.

3.5 A 6-unit GCE A-level

The full A level consisted of the three AS units featured in the previous section, plus three A2 units. As with the AS, the set of examinees aggregating in June 2009 contained those with units taken in sessions as far back as June 2007.

Table 3.15: A 6-unit A-level – number of aggregating examinees taking each unit in June 2009 (total N=11,603). Only combinations with more than 100 examinees are shown.

Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	N	%
×	×	×	×	✓	✓	2929	25.24
×	×	×	✓	×	✓	288	2.48
×	×	×	✓	✓	✓	2348	20.24
×	×	✓	×	✓	✓	657	5.66
×	×	✓	✓	✓	✓	500	4.31
×	✓	×	×	✓	✓	239	2.06
×	✓	×	✓	✓	✓	736	6.34
×	✓	✓	✓	✓	✓	327	2.82
✓	×	×	×	✓	✓	738	6.36
✓	×	×	✓	✓	✓	476	4.10
✓	×	✓	×	✓	✓	405	3.49
✓	×	✓	✓	✓	✓	283	2.44
✓	✓	×	✓	✓	✓	317	2.73
✓	✓	✓	✓	✓	✓	352	3.03

Table 3.15 makes it clear that, as might be expected, the examinees aggregating in June 2009 had mostly taken A2 units in June 2009. Nonetheless, there were significant numbers also taking at least one of the AS units in June 2009. As before, to keep the number of combinations manageable when considering the effect of changes of ±1 mark to the unit grade boundaries, we only considered the grade A boundary, and we treated the within-unit options as a single block, ignoring the two-component structure of these options. The number of possible combinations for aggregation compared with the AS (section 3.4) rose from 27 (3^3) to 729 (3^6). It was not feasible to derive the outcome for all these combinations, so a relevant selection is shown below.

Table 3.16: A 6-unit A level – effect of varying June 2009 unit grade A boundaries on overall % of examinees at grade A (actual outcome in bold).

Unit 1 June 2009	Unit 2 June 2009	Unit 3 June 2009	Unit 4 June 2009	Unit 5 June 2009	Unit 6 June 2009	Cumulative % aggregate grade A (N=11,603)
-1	-1	-1	-1	-1	-1	33.41
0	0	0	-1	-1	-1	32.94
0	0	0	0	0	-1	32.35
-1	-1	-1	0	0	0	32.33
0	0	0	0	-1	0	32.31
0	0	0	-1	0	0	32.07
0	**0**	**0**	**0**	**0**	**0**	**31.84**
0	0	0	+1	0	0	31.60
0	0	0	0	0	+1	31.39
+1	+1	+1	0	0	0	31.36
0	0	0	0	+1	0	31.27
0	0	0	+1	+1	+1	30.79
+1	+1	+1	+1	+1	+1	30.35

Table 3.16 shows that the variability of aggregation outcomes at grade A was ≈3 percentage points when the grade A boundaries on the 6 units were moved by ±1 mark. Changing all the AS units simultaneously only affected the outcome by about 0.5 percentage points. Not surprisingly, given the entry patterns shown in Table 3.15, changes to the A2 units had more impact – changing the boundary on either Unit 4, Unit 5 or Unit 6 had as much impact as changing the boundary on all three AS units. Unit 5 and Unit 6 on the A2 appeared to be more influential than Unit 4, but given that more of the aggregating examinees had taken Unit 5 and Unit 6 in June 2009 this is not surprising.

3.6 Discussion

A great deal of information contributes to the decision on whereabouts on the raw mark scale to locate each key grade boundary. This information can point in different directions, because the different sources of information are related to different definitions of what it means to maintain a standard. For example, in simple cohort referencing the cohort is assumed to be unchanging and the grade boundary is set at the point that gives as similar as possible a cumulative percentage of examinees in the grade. In more sophisticated cohort referencing, the cohort is not assumed to be unchanging, and external information (e.g. on prior attainment) is used to allow for changes in the cohort to derive a 'putative' value for the cumulative percentage of examinees in the grade. Neither of these two sources of evidence takes any account of the quality of work produced by the examinees. Expert judgment does do this – but there is no reason why the expert judgment should necessarily agree with the cohort referencing.

In our view, the only good reason to change a unit/component grade boundary from one session of an examination to another is if there is evidence that the overall difficulty of the questions has changed![34] Unfortunately, it is very difficult to disentangle question difficulty from examinee ability. As seen in Section 1, in IRT models difficulty and ability are conceived conjointly, as 'two sides of the same coin'. Some examination / testing systems allow for pre-testing of questions and a statistical calibration of question difficulty via an IRT model. In principle this allows grade boundaries to be set before the examination is taken. GCE / GCSE examinations are (generally) regarded as too 'high stakes' for any pre-testing with its consequent risk to question security. Pre-testing is also, of course, extremely expensive. In the absence of pre-testing, the difficulty of the questions is judged by how well the examinees have scored on them, which introduces an unwelcome circularity into the process, as shown below (see Bramley, 2010 for further discussion of this point).

Ideal (example) chain of reasoning:
1. The questions are slightly easier than they were last session;
2. Therefore we should raise the grade boundary by x marks to compensate.

Actual (example) chain of reasoning (using the sophisticated cohort-referencing approach):
1. The cohort of examinees is of slightly lower ability than it was at the last equivalent session, according to our information about prior attainment;
2. Therefore we expect a slightly lower cumulative percentage of examinees to achieve the grade;
3. If we raise the grade boundary by x marks the grade will be achieved by a slightly lower cumulative percentage of examinees than at the last equivalent session, in line with the 'putative' prediction;
4. Therefore the questions must have been slightly easier than they were at the last equivalent session.

[34] Assuming that the structure of the unit and the assessment of which it is a part have not changed.

The difficulty with the 'ideal' scenario above is in establishing evidence for point 1. If there was some way to estimate the relative difficulty of the questions independently of how examinees scored on them, this would presumably be a good thing. If experts could accurately judge relative overall difficulty of examination questions, the grade boundaries could be set before the examination is taken. Unfortunately the available evidence suggests that experts are not very good at judging question difficulty, either in absolute terms or relative terms (e.g. Curcin, Black & Bramley, 2009). This means that examination boards (and the regulator) tend to have more confidence in the sophisticated cohort referencing approach, despite the fact that it does not involve consideration of either the difficulty of the questions or the quality of work produced by the examinees.

It should be clear that the setting of grade boundaries is not a problem with a clear-cut answer. Therefore it is perhaps of interest to consider how the outcomes might have been different if different decisions had been taken[35]. The analyses presented here give some indication of what such reporting might look like. Two potentially useful ways of quantifying the potential variability in aggregate outcome are:

- to determine the range of possible aggregate outcomes that could have arisen if all relevant key grade boundary decisions at unit/component level had been 1 mark lower or 1 mark higher;
- to discover the largest change to the aggregate outcome that could have arisen from a 1-mark change in the boundary on a single unit/component.

The most obvious factors affecting the sensitivity of the aggregate outcome to decisions on the individual units/components are: i) the number of units/component to be aggregated – the greater the number the less the effect of changes on any one unit/component; and ii) the percentage of examinees on each mark point at the part of the distribution where the grade boundary lies (on each unit in unitised schemes, but on the aggregate distribution in linear schemes[36]). Units with longer raw mark scales, all things being equal, might be expected to have a lower percentage of examinees on each mark point. The correlation of scores among the units can also be expected to have an effect, with changes to grade boundaries on more highly correlated units/components affecting the aggregate more.

A more subtle point relating to unitised assessments is the effect of potential grade boundary changes to the 'conversion rate' of raw marks to uniform marks. Changes that reduce the distance between the A and the E boundary (i.e. lowering the A boundary and/or raising the E boundary) increase the rate of exchange; and vice versa. So whereas on a linear assessment a change to a component boundary changes the aggregate boundary but does not affect the aggregate totals of any examinees, in a unitised assessment a change to a unit boundary does not affect the aggregate UMS boundary but does affect the unit (and hence the aggregate) UMS total of most of the examinees who took that unit. So on a linear assessment (for example a higher tier GCSE) a change to a component grade A boundary could not affect the cumulative percentage of examinees obtaining aggregate grade C, but on a unitised assessment a change to a unit grade A boundary could conceivably affect the cumulative percentage of examinees obtaining aggregate grade E. Admittedly this effect is likely to be very small for the ±1 mark changes we are talking about. In the case of the 6-unit A level reported here, lowering the grade A boundary by 1 mark on all six units would have resulted in an extra 3 examinees (out of 11,603) obtaining an aggregate grade E. Lowering the A boundary and the E boundary by 1 mark on all six units would have resulted in an extra 30 examinees obtaining an aggregate grade E. Interestingly, lowering the E boundary by 1 mark and raising the A boundary by 1 mark on all six units would have resulted in an extra 38 examinees obtaining aggregate grade E! This

[35] Of course, examination boards do consider the aggregate effect of the decisions made at unit level at the time when those decisions are taken, in 'modelling' exercises. We are suggesting here that the range of fluctuation could be reported more systematically.

[36] For linear schemes that use indicator 1 only. If indicator 2 is used then the number of examinees at mark points around the boundary on the individual components is also relevant.

illustrates the point that the UMS conversion can have some slightly counter-intuitive effects – but supports the claim that the proportion of examinees affected is likely to be very small.

In unitised assessments it is very difficult to gauge or control the impact of changes at unit level because of the large number of different valid combinations of units, from different examination sessions, that can be aggregated to achieve an overall result at assessment level. Decisions made in a particular examination session cannot have any effect on the UMS scores on units from previous sessions. For the new unitised GCSEs, 'terminal rules' specify that a certain proportion of the units must be taken in the same session that aggregation will take place, which will presumably mitigate this problem to some extent.

For the new 2/4 unit GCE AS and A levels, however, there are no new 'terminal' requirements. Table 3.17 below shows some of the combinations seen in the aggregated data for the 6-unit GCE A-level in June 2009. This involved units from five previous examination sessions, going back to June 2007.

Table 3.17: Frequency of combinations of 'best six' modules of examinees aggregating in June 2009 (N=11,603). Selection of rows in descending order of frequency.

Combination	Unit 1	Unit 2	Unit 3	Unit 4	Unit 5	Unit 6	N
1	Jan 08	June 08	June 08	Jan 09	June 09	June 09	1952
2	June 08	June 08	June 08	June 09	June 09	June 09	792
3	June 08	June 08	June 08	Jan 09	June 09	June 09	589
4	Jan 08	Jan 09	June 08	Jan 09	June 09	June 09	575
…							
203	Jan 09	Jan 09	June 07	Jan 09	June 09	June 09	2
204	June 09	June 08	Jan 08	June 09	June 09	June 09	1
…							
461	June 07	June 07	June 07	June 09	June 09	June 08	1

Table 3.17 shows that there were 461 different combinations of 'best six' units observed in the data set. The most common combination only involved ≈17% of the cohort, and more than half of the observed combinations only involved a single examinee.

It is also of interest to see how many times units are re-taken. Table 3.18 shows how often each of the units in the 6-unit A level had been taken by examinees aggregating in June 2009.

Table 3.18: A 6-unit A-level – number of examinees aggregating in June 2009 (N=11,603) re-sitting each unit.

	0 re-sits	1 re-sit	2 re-sits	3 re-sits	4 re-sits
Unit 1	6,026	4,258	1,135	169	15
Unit 2	4,904	5,038	1,552	98	11
Unit 3	7,844	3,284	444	28	3
Unit 4	7,961	3,475	120	47	0
Unit 5	10,762	763	70	8	0
Unit 6	11,291	291	20	1	0

The AS units are taught first and thus have more opportunity to be re-taken over a 2-year course. It is clear from Table 3.18 that that opportunity is taken by many examinees!

The purpose of Tables 3.17 and 3.18 is to re-iterate the point made in Section 1 when considering 'composite reliability' – that the extreme flexibility of the unitised assessment structure in most GCE AS and A levels (and soon in GCSEs too) makes some traditional conceptions of reliability inappropriate. What we would argue is appropriate, in terms of grading

reliability, is to consider the range of possible outcomes (grade distributions) that could have been obtained if grade boundary decisions taken in a particular session had been slightly different. We have chosen to define 'slightly different' as 'varying by ±1 mark', i.e. the smallest difference possible. There would be some justification for taking a wider range, given that the 'zone of uncertainty[37]' in expert judgment of script quality usually spans a range wider than ±1 mark. The results in sections 3.2 to 3.5 could then be seen as lower bounds.

To put the kinds of variability we have found into context, Table 3.19 below shows the cumulative percentage of examinees obtaining grade A from June 2006 to June 2009 in each of the four assessments we looked at. This table uses the 'final' data on the system, rather than the data available at the time of awarding used in the analyses in sections 3.2 to 3.5, so the numbers of examinees do not exactly match.

Table 3.19: Grade A cumulative percentages and number of examinees, 2006-2009.

Qualification		2006	2007	2008	2009 old	2009 new	2009 combined
A linear GCSE	%	13.6	12.2	12.5	14.6		
	N	3323	3977	4764	5244		
A 2-unit AS level (new in 2009)	%	10.2	10.7	12.2	19.6	10.7	12.5
	N	6617	6785	6834	1253	4896	6149
A 3-unit AS level (new in 2009)	%	20.3	20.7	20.0	32.9	17.6	19.9
	N	14192	14836	15166	2936	16234	19170
A 6-unit A level	%	28.6	29.9	30.9	31.7		
	N	10290	11113	11472	11874		

Two of the assessments we considered had a new and an old version available in June 2009. We have assumed that combining the outcomes for the old and the new gives the most appropriate comparison with previous years, an assumption which seems borne out given the similarity of the combined percentage at grade A with previous years.

It is very striking how similar the cumulative percentages gaining grade A were from year to year in the period 2006-2009, given that the examinees were different and the size of the entry varied somewhat. In no case was the largest difference between any pair of years more than 3.1 percentage points, and most adjacent pairs of years differed by less than 1 percentage point. On the other hand, the analysis in sections 3.2 to 3.5 showed that the possible range of variation in percentage at grade A with *exactly the same examinees* could be from around 2 to 4 percentage points, if boundary marks on all units/components were changed by ± 1 mark. This suggests that the current statistically driven grade-boundary setting procedures could be 'overfitting' and producing a year-on-year grade distribution that does not fluctuate enough, given all the conceptual conflicts and practical limitations of the standard maintaining process described above. Of course, given public expectations about 'standards' it might be difficult to explain that a more fluctuating grade distribution is perfectly acceptable. On the other hand, it would help to avoid the pattern that is sometimes seen of steady year-on-year small incremental rises in pass rates that lead to accusations of 'grade drift' (see for example Oates, 2009 and its coverage in Paton, 2010).

The findings reported here also suggest that the aggregate outcomes on the new unitised GCSEs should be monitored carefully, in terms of investigating their sensitivity to changes to unit grade boundaries. Although the 'terminal rules' reduce the chance for flexibility and variety in the combinations of units making up the aggregates, this is more than offset by the increase in complexity arising from the possibility of tiering, particularly given that examinees do not have to

[37] The term formerly given to the range of marks over which there was no consensus among a panel of experts that the quality of scripts was definitely worth the higher or lower of two adjacent grades. Nowadays this range is referred to simply as the 'zone' – presumably so as not to give the impression that there is any uncertainty in the process!

take every unit in an aggregate at the same tier. Furthermore, when compared with GCE AS and A levels, GCSE units tend to be both shorter (have a lower paper total), and yet contain more grade categories (even when tiered). This suggests that the aggregate outcome is likely to be more sensitive to changes in unit grade boundaries than was the case for the GCE subjects investigated in this report. This is because: i) with shorter mark scales the unit score distributions are likely to be more bunched and hence a greater proportion of examinees will be affected by changes to unit grade boundaries; ii) the tiering leads to a wider variety of possible combinations for grades that are available on both tiers; and iii) with the shorter mark scales, the effect of changes to unit grade boundaries on the raw score to UMS conversion rate is likely to be more significant.

3.7 References for Section 3

Bramley, T. (2010). *Locating objects on a latent trait using Rasch analysis of experts' judgments.* Paper presented at the conference "Probabilistic models for measurement in education, psychology, social science and health", Copenhagen, Denmark, June 2010.

Curcin, M., Black, B. & Bramley, T. (2009). *Standard maintaining by expert judgment on multiple-choice tests: a new use of the rank-ordering method.* Paper presented at the British Educational Research Association annual conference, University of Manchester, September 2009.

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

Novakovic, N., & Suto, I. (2010). The reliabilities of three potential methods of capturing expert judgement in determining grade boundaries. *Research Matters: A Cambridge Assessment Publication, 9*, 19-24.

Paton, G. (2010). GCSE and A-level results being 'inflated'. *Daily Telegraph.* http://www.telegraph.co.uk/education/educationnews/7528383/GCSE-and-A-level-results-being-inflated.html Accessed 17/5/10.

Oates, T. (2009). 'Standards are up this year' – what does this mean? The question of standards in public examinations. http://cambridgeassessment.files.wordpress.com/2010/01/the-question-of-standards-in-public-examinations-by-tim-oates1.pdf Accessed 17/5/10.

OCR (2009). Entry Codes 14-19 Qualifications. January 2009, March 2009, June 2009. http://www.ocr.org.uk/administration/documents/general.html Accessed 01/11/09.

Ofqual (2009). GCSE, GCE and AEA code of practice, April 2009. http://www.ofqual.gov.uk/files/2009-04-14-code-of-practice.pdf Accessed 08/01/10.

SCAA (1994). GCSE mandatory code of practice, 1994. London: School Curriculum and Assessment Authority.

Summary

Summary of Section 1 - test-related reliability

Classical Test Theory (CTT)

The average (median) value of Cronbach's Alpha was 0.83 across the 287 units/components for which were able to obtain this index. It was also 0.83 for the 97 A level and 190 GCSE units/components considered separately. In terms of CTT, this suggests that, on average, around 80% of the variability in test scores on these units/components was true score variance and 20% was random error variance. It should be noted that these units/components were all part of larger assessments, and hence that these figures do not represent the reliability of complete GCSEs or A levels.

Plots of Cronbach's Alpha against total number of marks, number of questions, and unit/component weighting in overall assessment all showed the expected increasing relationship. Graphs like this can be used to identify units/components that have unexpectedly low values for Alpha, given their total mark or number of questions, and hence could be a useful tool for prioritising reliability investigations.

If a single index of reliability is to be used for comparing different units/components, it should attempt to take account of relevant differences among the units/components. We suggested that a 'grade bandwidth : SEM' ratio might be a suitable index. With this index, a unit/component with a grade bandwidth of 10 marks and a SEM of 5 marks would be of equivalent reliability to a unit/component with a grade bandwidth of 6 marks and a SEM of 3 marks because an examinee with a true score in the middle of a grade band would have the same likelihood of being misclassified (receiving a different grade from their 'true' grade) in both cases. We showed that this index succeeded to some extent in 'controlling' for differences among units/components in total number of marks.

Item Response Theory (IRT)

In IRT models, reliability is conceptualised in terms of information. The more information about a parameter, the smaller its standard error. In IRT models the standard error of an examinee's estimated ability varies across the score scale. Information is greater, and hence standard errors are smaller, in the middle of the score scale than at the extremes. It is possible to derive an overall index of reliability that is analogous to Cronbach's Alpha – known as the Person Separation Reliability, R_β.

We analysed 12 of the units/components with a Rasch model (the simplest IRT model). The resulting values for R_β were slightly higher than the corresponding values for Cronbach's Alpha, but the relative ordering of the two sets of results was very similar.

Composite reliability

Since nearly all GCSE and A level assessments are made up of a number of units/components, it is important to consider the overall reliability of the composite. The educational measurement literature provides several formulas for calculating 'composite reliability' from the reliability of individual elements. However, these cannot be applied directly to GCSEs and A levels because of the complexity of the assessment structures, particularly for modular or 'unitised' assessments. The amount of choice available to examinees in which units they take and when they take them makes it difficult to define what 'the' composite might be. No information was available on the reliability of units/components that had not been externally assessed, such as coursework or practical examinations, meaning that the reliabilities of these had to be estimated. The non-linear weightings introduced by use of the Uniform Mark Scale and the possibility of re-sits further complicate the picture.

Nonetheless, in the spirit of exploration, we calculated CTT indices of composite reliability ('composite Alpha') for four assessments. In all cases the value for the composite was higher than that of the elements comprising it, and this was also true for the grade bandwidth : SEM index.

Deriving an IRT index of composite reliability creates yet further problems which we argued would stretch credibility beyond breaking point for GCSEs and A levels. For interest, we derived such an index by two different methods for one assessment. The results were higher than the reliabilities of the individual elements, and only differed by one percentage point from each other.

Classification consistency

Given that the grade rather than the score is the focus of reporting, both for the examinee and for the examination system as a whole, arguably it is more appropriate to evaluate reliability at this level too. The relevant question becomes 'To what extent would the grade outcomes be the same if the test or assessment were to be replicated?' – a question about 'classification consistency'.

The question can be framed in different ways:

- in terms of the examinee – the probability that a given individual would get the same grade;
- in terms of all the examinees – the proportion that would get the same grade;
- in terms of the examinees with a given grade – the proportion that would get the same grade.

An IRT approach is more suitable than a CTT approach for answering these questions because the SEM varies across the score range and thus should give a more accurate answer. We used a relatively simple method based on calculating each examinee's distance from each grade boundary in terms of the IRT scale and the standard error of their ability estimate.

The results were presented as two graphs and a table. One graph showed the probability of being consistently classified within each grade for each score on the test. The other showed the score distribution and the percentage of examinees estimated to be classified consistently at each part of the distribution. The table showed the number and percentage of examinees in each grade category and the number and percentage estimated to be consistently classified. We showed how equivalent graphs and a table could be produced using a crude approximation from a CTT analysis.

Conclusions

Given that GCSE and A level units/components are not currently designed to fit any particular IRT model, and that information about reliability will need to be generated routinely in 'batch jobs', a CTT approach is likely to be the only feasible one in the short to mid-term. Given the problems with interpreting Cronbach's Alpha and the grade bandwidth : SEM ratio in an absolute sense, we recommend using them comparatively to identify units/components that seem to have lower values for reliability than other similar units/components. Further investigations could then aim to discover whether there was a legitimate reason for such lower values.

The location of the grade boundaries, and the score distribution are interesting and relevant features of any test administration, showing how well the test was 'targeted' at its cohort of examinees. Graphical displays of this are informative in their own right. We have shown how information about classification consistency can be included in these displays. If the assumptions underlying the calculation of the classification consistency statistics can be supported, this would seem to be a good visual way of presenting information about test-related reliability.

It is not easy to conceptualise test-related reliability, or to interpret the various indices of it. Hence we would recommend that test-related reliability information is not presented in a way that invites (possibly unjustified) inferences at the level of the individual examinee.

Suggestions for further research

The current structure of GCSEs and A levels means that calculations of composite reliability are complex, and the interpretation of the resulting indices is unclear. However, it is arguably at the level of the whole assessment that errors of classification have the most impact on the life chances of individual examinees. We therefore recommend that more work is carried out to investigate the best way to conceptualise and report test-related reliability at whole assessment level.

We found that the four different indices of test-related reliability that we considered (Cronbach's Alpha, the Rasch separation reliability, the bandwidth:SEM index and the classification consistency) were positively correlated. It would be interesting to investigate further the relationship between classification consistency at both individual grade and overall unit levels and the grade bandwidth:SEM ratio. Although high levels of classification consistency within individual grades may be associated with higher grade bandwidth:SEM ratios, the overall classification consistency is a function of a range of factors including boundary locations, mark distributions and the shape and size of the error distribution.

We considered the relationship between indices of test-related reliability and readily available, easily quantifiable features of the test such as paper total mark, number of items and weighting. Further investigation could explore the relationship between indices of reliability and more qualitative features such as the content and format of the questions and mark schemes. Such investigations would yield a deeper understanding of the factors affecting reliability.

IRT-based estimates of classification accuracy and consistency could be investigated further, perhaps by simulation studies. There is certainly scope for more research on composite reliability in an IRT framework.

Reporting reliability in terms of grade classification consistency is very appealing, but this requires accurate standard errors to be estimated for examinees at all parts of the score scale. The most natural way to do this is with an IRT approach, but it may not be feasible to run IRT analyses routinely in batch jobs across large numbers of units/components. It is therefore worth investigating further how estimates of classification consistency derived from approximate methods compare with those derived from theoretically preferable methods, and whether better approximations can be found than the crude one suggested here based on the average SEM from CTT.

Summary of Section 2 – Marker-related reliability

We summarised existing publicly available research on marker reliability in GCSEs and A levels, noting that much of it has not had the calculation of a reliability coefficient for a particular examination as its primary purpose. Instead, the focus has been on discovering factors affecting levels of marker agreement with a view to understanding how marking reliability can best be monitored and improved. The general finding in published research has been, not surprisingly, that marker reliability is higher on exams containing structured, analytically marked questions than on exams containing essays, and in general the less subjective the mark scheme, the higher the marker reliability.

We noted an unfortunate tendency of much of the published research to use either correlations or similar indices to describe marker reliability, which can sometimes be misleading or less informative than other indices.

In assessing the reliability of GCSE and A levels we were interested in discovering the extent to which the differences between the markers' marks and the correct or 'definitive' marks had a mean and SD of zero. By analogy with Cronbach's Alpha in Section 1, it was also of interest to quantify the proportion of variance in marks that could be attributed to differences among the markers. We used data arising from the live examination process, rather than from specially constructed research exercises.

Paper-based marking system

The 'traditional' system of monitoring markers in paper-based (as opposed to online) marking is hierarchical – a Team Leader (TL) monitors the marking of the Assistant Examiners (AEs) in their team. This monitoring is achieved by the TL re-marking a sample of each of their team's allocation of scripts, at one or more points in the marking process. This second marking is *non-blind* – i.e. the TL can see the original marks awarded by the AE – which as might be expected (and as research has shown) leads to higher levels of agreement than blind double-marking.

We reported here some information about the agreement between TL and AE in this process, using data from 22 OCR units/components covering both GCSE and A level taken in the June 2006 examination session. The distribution of differences between AE and TL was presented, along with the correlations.

The AE/TL correlations were all very high, ranging from 0.999 to 0.964, and compared favourably with comparable research outcomes from non-blind marking studies. In all of the units/components sampled, more than 50% of AE marks were within ± 1 mark of the TL's mark. In all but two cases more than 90% of the AE marks were within ±4 of the TL's mark.

On average, the differences were close to zero – in other words there did not seem to be any systematic bias across the different subjects for AEs to be more severe or lenient than the TLs. In all except five units/components, the absolute value of the mean difference was less than 0.5 marks. In four of the five exceptions to this, the mean difference was negative, suggesting that where there was a bias, AEs tended to be more severe than the TL.

On-screen marking system

The system for monitoring on-screen marking is very different from that for monitoring marking in the paper-based system. 'Seed scripts', for which the 'definitive' mark on each item has been established by a panel of senior examiners, are regularly inserted into each AE's marking allocation.

From the point of view of comparing marker agreement statistics in the on-screen system with the paper-based system, there are several important differences to be aware of:

- the second-marking is blind, because the AEs are not aware of the definitive marks;
- the same seed scripts are marked by all markers;
- the seed scripts are marked at consistent intervals throughout the marking process rather than at specific points in the process;
- the definitive mark is established by the principal examiner (PE) and senior examiners, not the TL.

Also, as we mentioned in the summary of Section 1, and repeatedly stressed throughout the full report, the units/components currently marked on-screen tend not to include long answer or essay questions where there might be expected to be less marker agreement. Statistics on marker agreement from these units/components are therefore not representative of the system overall.

We presented four different ways that data on marker agreement from the seed scripts could be presented for a single unit/component:

- a histogram showing the distribution of differences between markers' mark and definitive mark;
- a dotplot showing the marks awarded to each seed script, including the definitive mark;
- a dotplot showing the difference between awarded and definitive mark for each seed script (including mean difference for each seed script);
- a dotplot showing the difference between awarded and definitive mark for each marker (including mean difference for each marker).

Each of these plots gives a clear, easy-to-interpret, picture of the level of marker agreement, and together allow identification of overall level of agreement, overall tendency towards severity or lenience, individual seed scripts that were particularly difficult to mark, and individual markers whose marks were consistently severe or lenient (or just discrepant in either direction).

Across all units/components, not only was the median of the median difference between awarded and definitive mark equal to zero, the lower and upper quartiles were too. This is good evidence that, for the type of examination currently marked on screen, systematic severity or lenience of all the markers relative to the definitive mark is relatively rare.

Interestingly, the units/components with the least marker agreement had also been among those with the lowest values for Cronbach's Alpha in Section 1, implying that low values of Cronbach's Alpha can (sometimes) be attributed to unreliable marking.

In order to make comparisons with the results in Section 1, we treated the standard deviation of the differences as an approximate SEM attributable to marker variability, and compared it with the test-related SEM derived via Cronbach's Alpha in Section 1. We showed that, for the kind of unit/component currently marked on-screen, different markers contribute much *less* to score unreliability than different questions.

To investigate whether the differences between awarded mark and definitive mark arose mainly because markers differed systematically in their levels of severity, or because seed scripts differed systematically in how severely or leniently they were marked, we carried out some variance components analysis. Although there were many caveats around the interpretation of the results, they seemed to support the following two generalisations:

- most of the variance in differences between awarded and definitive mark is not attributable to systematic differences among the markers or the seed scripts;
- systematic differences among seed scripts are relatively larger than systematic differences among markers.

Although the focus of the report was on marker agreement at the whole script level, we reported some overall agreement statistics at item level. Defining a 'marking event' as an instance of each item (part-question) on a seed script being marked by a marker, and using exact agreement between marker's mark and definitive mark as the index of reliability, out of nearly 3 million marking events at the item level, over 90% had exact agreement between marker and definitive mark. The percentage was lower for GCE units/components than for GCSEs, presumably because they had relatively fewer low-tariff objective questions.

Conclusions

Data routinely collected from seed scripts in on-screen marking is a rich source of evidence about marker agreement. The seed scripts admittedly are a very small proportion of the total number of scripts in each examination, but the agreement statistics can be calculated in batch jobs across the full range of examinations marked on screen, thus giving a useful snapshot of the whole screen-marked system.

We recommend that analysis of seed script agreement data should be based on the distribution of differences between awarded and definitive mark. Graphical displays of these distributions

can be very informative. Correlations are not very informative and should not be presented on their own.

Suggestions for further research

Obviously it would be of great interest to extend the analysis of seed scripts to other kinds of units/components with longer written answers. These are probably the kinds of assessment where there is most concern over the reliability of marking, and the most requests for re-marks. It is possible that this information will become available routinely as on-screen marking becomes more widespread.

It would also be of great interest to investigate the reliability of marking (or other assessment judgments) in the kinds of units/components that are not externally assessed (coursework, practicals, orals, etc.).

A more ambitious project would be to attempt to investigate test-related and marker-related variability within a single conceptual framework. Generalizability Theory is a more sophisticated conceptual development of CTT. In this framework

> "*The error component can be broken down into several different subcomponents, the contributions to error of those separated components quantified, and their effects combined in a single comprehensive reliability coefficient*" (Johnson & Johnson, 2009, p21).

Generalizability theory would also allow other potential sources of systematic variability in test scores (for example attributable to differences among schools) to be investigated too. There are also IRT approaches that might be worth investigating (e.g. Linacre, 1994). These kinds of study need to have a rigorous design that ensures that sufficient data will be collected to estimate the relevant parameters, rather than opportunistically making use of existing 'operational' data sets as was the case in this report.

Summary of Section 3 – Grading-related variability

We reviewed the process by which grade boundaries are set on GCSE and A level units/components, noting that evidence from a variety of sources needs to be integrated. Given that this evidence can potentially conflict, the setting of grade boundaries is not an exact science. There has been a trend in recent years (at GCE in particular) to place most weight on the 'putative' grade boundaries derived from statistical information on changes in the ability of the cohort as indicated by mean GCSE grade.

Given that it is not possible to determine exactly what the grade boundaries 'should' be, it is of interest to investigate what the impact of slightly different decisions at unit/component level would be on the grade distributions at whole assessment level. In particular, it seems likely that the evidence for any particular grade boundary decision could support two possible boundary marks, and perhaps more. We investigated the effect on assessment grade boundaries of varying the (judgmentally set) key grade boundaries on the units/components by ±1 mark.

A linear GCSE assessment

Our first example was a 'linear' specification where two written papers and a coursework component were aggregated. Based on a straightforward aggregation of the three component grade boundaries, raising or lowering each component boundary by 1 mark created 7 possible aggregate grade boundaries (from 3 marks lower to 3 marks higher). The corresponding range in the cumulative percentage pass rate was 9.5 percentage points at grade C on the Foundation Tier, and up to 6 percentage points at grade A on the Higher Tier.

Combining the results from both tiers, and weighting each possible outcome by its chance of occurring (in a somewhat arbitrary but fairly reasonable way) gave a range for the cumulative

percentage pass rate at grade C covering 7 percentage points from 56% to 63%, with the most likely outcomes in a narrower range from about 58% to about 61%.

Unitised (modular) assessment

The effect of grade boundary changes in unitised (modular) assessments is much more difficult to derive, because of the large number of possible routes to the final assessment, and the possibility of taking different units in different examination sessions. We only considered the possibility of making changes to units in the June 2009 examination session – such changes could of course not affect the results on units taken in previous examination sessions.

At the time of writing the report, the majority of A levels were in transition from the old structure (3-unit AS levels and 6-unit A levels) to a new structure with 2-unit AS levels and 4-unit A levels. The first (large) cohort of examinees was aggregating for the 2-unit AS levels. There were not yet any examinees aggregating for the new 4-unit A levels.

In a 2-unit AS level, the difference in aggregate outcome from lowering the grade boundaries on the two June 2009 units by one mark to that from raising them by one mark was approximately 3.6 percentage points at grade A. At grade E virtually no examinees were affected. The situation was simplified by there having been only two examination sessions to consider (January 2009 and June 2009). Half of the examinees aggregating in June 2009 had taken one of the units in the January session and were thus unaffected by changes to the boundaries of that unit.

In a 3-unit AS level, the situation was more complex because the unit had been running for several years, and the set of examinees requesting aggregation for the AS qualification contained those with units taken as far back as June 2007. Nearly half of the examinees requesting aggregation had not taken any of the AS units in June 2009. The effect on the pass rate at grade A of lowering boundaries by one mark on all three June 2009 units compared with raising them all by one mark was only around 2 percentage points.

In the 6-unit A level corresponding to the AS level the examinees aggregating in June 2009 had mostly taken their A2 units in June 2009, and AS units in previous sessions. The effect on the pass rate at grade A of lowering boundaries by one mark on all six June 2009 units compared with raising them all by one mark was around 3 percentage points. Not surprisingly, changes to the A2 units had much more effect than changes to the AS units.

There were 461 different combinations of the six A level units in our data set, more than half of which involved only a single examinee. For the AS units, around half of the examinees had re-taken them at least once. This emphasises the point made in Section 1 that it is practically meaningless to calculate a value for 'the' composite reliability of a unitised A level.

To contextualise the variability we obtained from making changes of ±1 mark to the grade boundaries in a single session, we compared the resulting fluctuations in cumulative percentage pass rate at grade A with the fluctuations found over the time period 2006-9 in the same subjects. It was very striking how similar the cumulative percentages gaining grade A were from year to year in the period 2006-2009, given that the examinees were different and the size of the entry varied somewhat. In no case was the largest difference between any pair of years more than 3.1 percentage points, and most adjacent pairs of years differed by less than 1 percentage point.

On the other hand, our analysis showed that the possible range of variation in percentage at grade A with *exactly the same examinees* could be from around 2 to 4 percentage points, if boundary marks on all units/components were changed by only ± 1 mark. This suggests that the current statistically driven grade-boundary setting procedures could be 'overfitting' and

producing a year-on-year grade distribution that does not fluctuate enough, given all the conceptual conflicts and practical limitations of the standard maintaining process.

Of course, given public expectations about 'standards' it might be difficult to explain that a more fluctuating grade distribution is perfectly acceptable. On the other hand, it would help to avoid the pattern that is sometimes seen of steady year-on-year small incremental rises in pass rates that lead to accusations of 'grade drift'.

Conclusions

The setting of grade boundaries is not a problem with a clear-cut answer. Therefore it is perhaps of interest to consider how the outcomes might have been different if different decisions had been taken. The analyses presented here gave some indication of what such reporting might look like. Two potentially useful ways of quantifying the potential variability in aggregate outcome are:

- to determine the range of possible aggregate outcomes that could have arisen if all relevant key grade boundary decisions at unit/component level had been 1 mark lower or 1 mark higher;
- to discover the largest change to the aggregate outcome that could have arisen from a 1-mark change in the boundary on a single unit/component.

The most obvious factors affecting the sensitivity of the aggregate outcome to decisions on the individual units/components are: i) the number of units/component to be aggregated – the greater the number the less the effect of changes on any one unit/component; and ii) the percentage of examinees on each mark point at the part of the distribution where the grade boundary lies. Units with longer raw mark scales, all things being equal, might be expected to have a lower percentage of examinees on each mark point.

Suggestions for further research

Because the exam boards already have the wherewithal to determine the grade distribution on any unit/component or aggregate assessment for a given set of grade boundaries, it would be interesting to see the effect of changing the 'key' grade boundaries by ±1 mark on the assessment outcomes in a much wider range of assessments than it was possible to consider here. It might even be possible for exam boards to produce this information in a batch job in a similar way to that in which the statistics for Cronbach's Alpha are produced. With a much wider range of information to look at, it would be possible to identify those assessments where the final outcome (in terms of the percentage of examinees in each grade band) was particularly sensitive to the particular set of grade boundaries that happened to be chosen.

Appendix

Assessment terminology used in examinations in England

As with any complex process, a specialist terminology has developed to describe various entities and aspects of the examination system in England. To avoid cluttering the report with more footnotes or other explanations in the text, some of the terminology used in this report is explained below in Table A1.

Table A1: Glossary of assessment terminology used in this report.

Term	Description
Assessment	The entire set of units/components comprising a particular qualification.
Qualification	GCSE, AS and A level are the only qualifications referred to in this report.
GCSE	General Certificate of Secondary Education. Usually taken by 16 year olds.
GCE	General Certificate of Education. Consists of AS and A level.
A level	Assessment consisting of AS units/components and A2 units/components.
AS	Advanced Subsidiary. AS units/components are usually taken by 17 year olds in the first year of a two-year A level course. The standard is slightly below that of an A level.
A2	A2 units/components are usually taken by 18 year olds in the second year of a two-year A level course. The standard is slightly above that of an A level. A2 units do not form a qualification on their own, unlike AS units.
Specification	Formerly 'syllabus' – the document describing what will be assessed and how it will be assessed.
Linear assessment	An assessment where all the components are examined at the same time at the end of the course. The components of linear assessments are known as 'components'. At the time of writing this report, most GCSE assessments were linear but in the process of transition to unitised assessments. A typical example of a linear assessment might be one consisting of two written papers and a coursework component.
Unitised or modular assessment	An assessment that is broken down into discrete 'units' or 'modules' that can be taken in any examination session where that unit is available, subject to the rules in the scheme of assessment laid out in the specification for that particular assessment. Some units can contain two or more components.
Unit/component	The term used in this report to refer generically to either a unit of a unitised assessment, or a component of a linear assessment, or a component of a unit of a unitised assessment. Usually this distinction is of little relevance, but where it is it will be clarified in the text.
Written paper	A traditional examination unit/component where the candidate writes their answers to the questions in 'exam conditions' (as opposed to a unit/component of coursework, practical, portfolio, performance, or oral examination).

Term	Description
Examinee	The person taking the examination. Sometimes referred as the ‘candidate’ or ‘test-taker’.
Centre	The examination centre that the examinee is registered with. Usually but not necessarily a school.
Script	The physical paper or digital image containing an examinee’s answers to the questions on a written paper.
Mark scheme	Written document specifying how many marks (score points) are available for each question (or part-question) in the examination, and explaining how to allocate marks to examinee responses.
Raw score	The score obtained by adding up the marks obtained by the examinee on the questions in the unit/component.
Weighted score	The score obtained on a unit/component after multiplying the raw score by the weighting factor necessary to give the unit/component the weight prescribed in the specification for the assessment when it is aggregated with the other units/components in that assessment.
Grade boundary	The lowest mark on the raw score scale corresponding to a particular grade classification. (i.e. one mark less would have obtained the grade below).
Grade scale	The letter classifications labelling achievement in the unit or assessment. Different qualifications have different grades available. See Table 1.5.
UMS	Uniform Mark Scale – a more fine-grained numerical form of the grade scale with fixed boundaries corresponding to the different grades. Raw scores are converted to UMS scores in unitised assessments in order to aggregate the units. The number of UMS points available for a particular unit reflects the weight of that unit in the overall assessment, as set out in the specification.
Tiered assessment	A tiered assessment contains units/components where the range of grades available is a continuous but restricted subset of the full range of grades available for the qualification. GCSEs, but not AS or A levels, contain tiered units/components. The most usual situation is to have ‘foundation tier’ units/components targeting the lower grades, and ‘higher tier’ units/components targeting the higher grades, with two or three ‘overlapping’ grades – i.e. grades available on both foundation and higher tiers.
Options	In a linear assessment or within a unit of a unitised assessment, any valid combination of components that can be chosen by the examinee (or in some cases by their centre). The term does not usually apply to the choice of units in a unitised assessment.

We wish to make our publications widely accessible. Please contact us if you have any specific accessibility requirements.

First published by the Office of Qualifications and Examinations Regulation in 2011

Office of Qualifications and Examinations Regulation
Spring Place
Coventry Business Park
Herald Avenue
Coventry CV5 6UB

2nd Floor
Glendinning House
6 Murray Street
Belfast BT1 6DN

Telephone 0300 303 3344
Textphone 0300 303 3345
Helpline 0300 303 3346

www.ofqual.gov.uk