Estimation of internal reliability

January 2013

Malcolm Hayes
Jeremy Pritchard

Ofqual/13/5258

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

Acknowledgements

We would like to thank: the members of our supervisory group – Kath Thomas, Rose Clesham, Michael Young and Roger Murphy – for their input and advice; Ofqual for feedback on the draft version of this report; Edexcel for allowing access to data from their processing systems and databases; and the Technical Support Team in Edexcel for supplying assessment details and answering queries.

Contents

1. Executive summary

This report considers a range of measures of internal consistency for over 300 different assessments. These measures are referred to as reliability coefficients but it is important to recognise that reliability in this context is simply an estimate of the reproducibility of outcomes, that is, the extent to which candidates could be expected to achieve the same results on a different occasion. The values relate to the way that students interact with the assessment items and not to operational accuracy.

The focus of the report is on empirical rather than theoretical studies and seeks to provide evidence related to a wide range of assessments with varying structures and lengths. The choice of assessments has been restricted to level 2 and level 3 assessments where candidates are required to answer all questions. This restriction to assessments with no optional parts was to avoid unnecessary complications in processing the data. It should be noted that in most cases these assessments formed part of an award and would not therefore have cut-scores set in isolation. For this reason, it is not considered appropriate to investigate classification accuracy or consistency.

The main findings of this report are that MacDonald's ω_t, a coefficient derived from a factor analysis of test items, provides a closer estimate of true reliability than the more conventional Cronbach's α. However, calculation of this coefficient is technically more demanding, likely to be time-consuming and may require investment on the part of the awarding bodies.

Several factors are identified as impacting on values of reliability including the spread of scores, the mean number of marks per item and the distribution of marks per item.

Estimates of confidence intervals for values of reliability have been made and shown to be related to the number of candidates taking the assessment and the value of the coefficient itself.

The findings of this report may contribute to decisions made by Ofqual with respect to documentation of assessment reliability that awarding bodies might be required to furnish as part of their operational processes. The recommendations therefore lean towards practicality, usefulness and applicability to a wide range of assessments.

2. Introduction

This report forms part of Ofqual's ongoing research into the reliability of assessments. While previous research on the Reliability Programme explored mainly the use of Cronbach's alpha as a measure of internal reliability, this report discusses a range of internal reliability indices for a selection of assessments and identifies factors that influence the values of the various indices.

Ofqual's specification for this research required that in the context of this report, reliability should be taken to refer to the consistency of outcomes that would be observed from an assessment process were it to be repeated. High reliability is taken to mean that broadly the same outcomes would arise. A range of factors that exist in the assessment process can introduce unreliability into assessment results. However, this report is concerned with the aspect of reliability that relates to internal consistency.

Amongst other things, these measures of reliability reflect the extent to which the assessments each measure a single construct and although several 'rules of thumb' exist for values of reliability, giving ranges for *excellent* down to *unacceptable,* these 'rules' fail to take into account the nature of the assessment in question. To set raising the level of reliability to the highest possible value as the goal would be to drive assessment towards a homogeneity that would impoverish the whole system.

It is certainly possible to construct highly reliable tests by asking what amounts to the same question many times, thereby reducing measurement error variance but such a test would be of limited value. Nevertheless, in assessments where the questions tend to access a well-defined, coherent set of skills or knowledge, correlations between items would be expected to be high and this would generate high reliability coefficients. On the other hand, assessments covering a broad range of topics that access different skills and knowledge would be expected to show lower inter-item correlations and consequently, lower estimates of reliability.

This tension between coherence and breadth in an assessment inevitably means a reliability value that can be achieved in one subject may not be possible in another without destroying the validity and authenticity of the latter.

Typically, quantitative subjects exhibit higher reliability values than qualitative subjects and it is inevitable that any measure intended to estimate the likelihood of attaining the same result on a different occasion will produce different ranges of values depending on both the content domain and the nature of the assessment. The strength of reliability measures lies in their capacity to provide quality control information across series rather than a quality metric for comparing assessments from different subjects.

3. Aims and objectives of research

The research aims were:

- To select a range of assessments where it can reasonably be expected that test-related unreliability represents the major source of measurement error. For example, tests and examinations that are composed of multiple choice questions (MCQs) and/or short-answer constructed responses that can be objectively marked were to be investigated;
- To produce estimates for a range of internal reliability indices for the selected assessments;
- To analyse, interpret, compare and report on the reliability evidence generated;
- To assess the practical applications of the specific estimation techniques used in the research.

The assessments considered in this report are all drawn from GCSE or GCE qualifications and while the focus was on assessments composed of short questions that could be marked objectively, other assessment structures have been included for comparison purposes.

GCSE and GCE qualification grades are generally determined by the aggregation of performances across several assessments. For unitised qualifications, it has become the norm to report the grades for these individual assessments. For example, over the last few years, unit grades for GCE qualifications have been reported to UCAS. This may imply that they have some currency but they do not have any direct impact on qualification grades. Overall qualification grades are determined from the aggregation of the *marks* achieved on each assessment.

For linear qualifications, that is qualifications in which all of the assessments must be entered at the same time, individual assessment grades are not reported. The reason for this is that the full set of grade boundaries is only determined for the overall qualification, hence, the full set of individual assessment grade boundaries does not exist. As a consequence, any grades quoted for individual assessments would be, at best, an approximate summary of performance.

Given the processes in place, an analysis of ‘grade misclassification’ at unit level would be misleading and is not pursued in this report. Similarly, the measures suggested by Bramley and Dhawan (2010), such as the average grade band width:SEM ratio are not considered appropriate measures of quality in this context.

3.1. Reliability in this context

The measures computed in this research are all estimates of internal consistency reliability. They all use the test data to infer what might happen if candidates took another version of the test on another occasion and while it may be true that the calculated values could be impacted by operational issues such as who actually marked the assessment, clerical errors and even methods of administration, the measures do not provide evidence of the magnitude of any such effects.

Test scores x are assumed to consist of unobserved true scores τ and unobserved measurement errors ε, so that $x = \tau + \varepsilon$. The usual assumption is that the errors have zero mean and are uncorrelated with true scores.

More formally, E(x)=0 and Cov(x, τ)=0 and it follows that

$$\text{var}(x) = \sigma_x^2 = \sigma_\tau^2 + \sigma_\varepsilon^2 \qquad (1).$$

Reliability is defined as the ratio of true score variance to observed score variance but can also be shown to be the squared correlation of true score and observed score.

Since

$$\text{cov}(x,\tau) = \text{cov}(\tau+\varepsilon,\tau) = \text{cov}(\tau,\tau) = \text{var}(\tau) = \sigma_\tau^2 \qquad (2)$$

Then

$$\rho_{x\tau}^2 = \frac{[\text{cov}(x,\tau)]^2}{\text{var}(x)\text{var}(\tau)} = \frac{(\sigma_\tau^2)^2}{\sigma_x^2\sigma_\tau^2} = \frac{\sigma_\tau^2}{\sigma_x^2} \qquad (3)$$

Clearly, in order to calculate an estimate of reliability, either true score variance or error variance must be estimated. If the assumptions made in formulating these estimates are violated, then the results may be quite inaccurate.

Different approaches to estimating error or true score variance make different assumptions and as a result, the estimates can vary widely. However, most estimators are claimed to be lower bounds for 'true reliability' and therefore the greatest value of any set of estimators might be considered to be closer to the true reliability than any of the others.

If the same test were to be administered to the same sample of candidates on different occasions, making a series of practically implausible assumptions such as no learning taking place and aggregate levels of motivation being constant, a correlation between the test scores could be computed and this test-retest correlation taken as a measure of reliability. However, even under experimental conditions this is often impractical, much more so for 'live' assessments such as GCSE and GCE. In order to make an estimate of reliability from a single administration, various split-half methods have been devised. Broadly, the idea is to split the test at random into two equal parts and compute the correlation between scores on the two halves. This can then form the basis of an estimate of the test-retest reliability, though a different split would certainly generate a different value.

Cronbach's α is the most commonly-used measure of internal consistency reliability and is relatively simple to compute. Cronbach (1951) showed that α is equivalent to the mean of all possible split-half reliabilities and that *it is not a direct estimate of the reliability coefficient but rather an estimate of the lower bound of that coefficient* (Crocker and Algina, 1986).

Several researchers have shown that better lower bounds exist but these alternatives are invariably more complex to compute, a major consideration when computers were less readily available but of little consequence now.

4. Selection of assessments

Suitable assessments were identified from three consecutive Summer series of examinations from one awarding body. Each assessment formed part of an award for the particular subject and does not therefore provide direct evidence of the reliability of the whole award. Assessments were first selected from the last of the three series and then equivalent assessments from the previous two series were retrieved. Thus all the assessments used were based on specifications that were in operation in the latest of the three series. Changes in specifications meant that not all of the selected assessments had equivalents in the earlier series.

Assessments were included on the basis that candidates were required to answer all questions on the paper, since reliability analyses can only be carried out on complete datasets and optional questions would require partitioning of the datasets by the particular set of questions answered.

Some assessments had very limited numbers of candidates and in these circumstances, estimated confidence intervals for reliability coefficients can be very wide. As an additional constraint, results are only reported for assessments with over 100 candidates. A total of 165 suitable assessments were identified from the latest series of which 120 had been in operation one year earlier and 66 two years earlier, giving a total of 351 assessments. Details of the number of assessments at each level and within each series are given in table 4.1.

Table 4.1: Numbers of assessments

	Series 1	Series 2	Series 3	Total
GCSE	6	48	88	142
GCE	60	72	77	209
Total	66	120	165	351

5. Methods

5.1. Extraction of data

Data were stored in flat-file format with one record per candidate per item. Data for each subject and series were extracted to provide the following fields:

- centre number
- unique candidate number
- item number
- item mark

The files were converted to SPSS format and then restructured to give one record per candidate and one field per item. The data were merged with descriptive information pertaining to assessment design.

Concurrent with the computation of indices, assessment designers reviewed the marking rubrics for each assessment to determine which of the items in the assessment required expert judgement and which could be marked objectively.

The assessments were categorised according to the marking requirement and values of reliability measures were correlated with rankings of assessments based on the extent to which the assessment can be marked objectively. Factors such as mark allocations and length of assessment were also investigated.

5.2. Selection of indices to be computed

Revelle & Zinbarg (2009) discuss a wide range of reliability measures, pointing out that each one is an estimate of the lower bound for true reliability. Thus, the best estimate of true reliability is likely to be the greatest of these lower bound estimates. They show that for a selection of (rather short) tests, this is often McDonald's ω_t. However, they point out that the variability of the measures for any one data set shows that it would be a mistake to assume that any particular index gave the definitive lower bound for reliability.

Sijtsma (2009a) discusses limitations of Cronbach's α (1951), pointing out that

it is difficult to defend convincingly using one of the smallest lower bounds, α, given the availability of many greater lower bounds….

5.3. Classical test indices

The following indices can be computed directly from the SPSS reliability routine:

- Split-half
- Cronbach's α
- Guttman λ_1 to λ_6
- Parallel
- Strict parallel

The split-half model splits the scale into two parts and examines the correlation between the parts. The value provided is only one of many possible split-half values and depends on the order in which the variables are entered.

Cronbach's α is a model of internal consistency, based on the average inter-item correlation and is a generalisation of the Kuder-Richardson 20 (KR-20) formula.

The six versions of Guttman's λ each provide measures of reliability that all give lower bounds for the true reliability of the assessment. Although all six can be computed using SPSS, λ_6 requires that the determinant of the covariance matrix is non-zero, a condition that is likely to be breached in a significant number of the type of assessments considered here. The following descriptions are taken from the SPSS help file:

λ_1 is a simple estimate that is the basis for computing some of the other lower bounds.

λ_3 is a better estimate than λ_1, in the sense that it is larger, and is equivalent to Cronbach's α.

λ_2 is better than both λ_1 and λ_3 but is more complex.

λ_4 is, in fact, the Guttman split-half coefficient. Moreover, it is a lower bound for the true reliability for any split of the test. Therefore, Guttman suggests finding the split that maximizes λ_4, comparing it to the other lower bounds, and choosing the largest.

λ_5 is better than λ_2 when there is one item that has a high covariance with the other items, which in turn do not have high covariances with each other. Such a situation may occur on a test that has items that each pertain to one of several different fields of knowledge, plus one question that can be answered with knowledge of any of those fields.

λ_6 is better than λ_2 when the inter-item correlations are low compared to the squared multiple correlation of each item when regressed on the remaining items. For example, consider a test that covers many different fields of knowledge and each item covers some small subset of those fields. Most item pairs will not have overlapping fields, but the fields of a single item should be well represented given all the remaining items on the test.

The parallel model assumes that all items have equal variances and equal error variances across replications. The strict parallel model makes the assumptions of the parallel model and also assumes equal means across items.

The conditions for the parallel and strict parallel indices are unlikely to be met for tests with a range of item tariffs, as is the case for the assessments analysed here. Since Cronbach's α is equivalent to the mean of all possible split-half reliabilities, there is nothing to be gained in considering a single split-half as an alternative. Thus split-half, parallel, strict parallel and the KR-20 measures of reliability have not been considered further and only the Guttman λ and Cronbach's α coefficients were computed directly from the reliability routine. McDonald's ω_t was calculated from a factor analysis of each test.

Although algebraically, λ_3 is equivalent to α, SPSS uses different methods by default. The calculation for the Guttman coefficients requires computing a covariance matrix of the variables. It is slower and requires more space than the alternative. However, it can process all models, statistics, and options.

The default method for α does not involve computing a covariance matrix. It is faster than the first method and, for large datasets, requires much less workspace, but it is more limited in the options available.

The two methods of calculation differ in one other important respect. The second method will continue processing a scale containing variables with zero variance and leave them in the scale. The first will delete variables with zero variance and continue processing if at least two variables remain in the scale. This can produce significant differences between λ_3 and α but only if the data contains zero variance items. Where no zero-variance items appeared in the data, the differences, if they occurred, were of the order of 10^{-14}, that is, negligibly small.

Formulae for each of the coefficients are given in appendix 1.

5.4. Rasch-based reliability

The software program WINSTEPS (Linacre, 2011) was used to perform Rasch analyses of each dataset. WINSTEPS reports two types of reliability measures, *person separation reliability* and *item separation reliability*.

Item separation reliability indicates whether the sample is sufficient to precisely locate the items on the latent variable and has no classical test theory equivalent. This is not an area of interest for the current study and further references to the Rasch separation indices will relate to the *person separation*.

Person separation reliability gives an indication of the assessment's ability to discriminate the sample into different levels. Linacre claims:

- Values of 0.9 or more suggest 3 or 4 levels;
- Values between 0.8 and 0.9 suggest 2 or 3 levels;
- Values between 0.5 and 0.8 suggest 1 or 2 levels.

Presumably values below 0.5 do not support division into levels.

Person reliability is intended to be equivalent to the traditional test reliability. Values are sensitive to the spread of person measures and the number of items. Test targeting is also thought to impact on reliability values.

Two estimates of person separation reliability are given. The *model* person reliability (R_M) is an upper bound and the *real* person reliability (R_R) is a lower bound to this value.

Although these separation indices can be thought of as measures of reliability, they do not correlate very highly with the internal consistency measures. Linacre (2011) asserts that person separation indicates the reproducibility of relative measure location and is the equivalent of test reliability. High person reliability means that there is a high probability that persons estimated with high measures actually do have higher measures than persons estimated with low measures. Linacre (2011) also states that person separation is independent of sample size and is largely uninfluenced by model fit.

5.5. Confidence intervals for reliability coefficients

Confidence intervals for values of reliability were computed where it was possible to do so in the time available. Although analytical methods exist for computing confidence intervals, a bootstrap method provided a more practical solution. Using an adaptation of a macro provided in the SPSS help files and Output Management System (OMS) commands it was possible to select samples, run the reliability routines and collect the outcomes over large numbers of replications without the need to generate separate data files for each replication. Confidence intervals were then computed from the distributions of indices.

The sampling method works on the basis of assigning integer weights to each case such that the sum of the weights is the sample size. The macro does this by computing a binomial random variable for each case such that:

N= the number of cases still to be selected

$$p=\frac{1}{\text{number of cases to be allocated a weight}}$$

Details of the macro are given in appendix 2.

The method was appropriate for the Guttman λ coefficients and for Cronbach's α but other coefficients could not be dealt with in the same way. However, correlations between particular coefficients may allow for further inferences to be made.

6. Analysis and results

6.1. Sample sizes

The entry sizes for these assessments varied from 104 to 175,000 so to illustrate the distribution the data were grouped by order of magnitude. Approximately half the entry sizes were between 1000 and 10000. Datasets were available for assessments with entry sizes below 100 but calculation of reliability coefficients proved to be very unstable. It would be inadvisable to attach any weight to findings based on such small samples and therefore these assessments were excluded from the results presented in this report. The distribution of entry size is given in table 6.1.1.

Table 6.1.1: Entry sizes

Entry size	Number of assessments
100-999	84
1000-9999	188
10000-99999	67
Over 100000	12

6.2. Assessment formats

The assessments reviewed varied widely in terms of the number of marks, the number of items and the number of marks per item. Charts 6.2.1 to 6.2.3 illustrate the differences between the two levels. Half of the GCSE assessments have test totals of 40 or 50 marks whereas almost all of the GCE assessments are in the range 70–90 marks. In contrast, 80% of the GCSE assessments comprised more than 20 items while almost 80% of GCE assessments comprised 20 or less items. The mean number of marks per item tends to be much lower for GCSE than for GCE. Tables of the distributions of test maxima and number of items are given in appendix 3.

Assessments with questions attracting ten or more marks generally require human marking and therefore introduce subjectivity into the process. While it might be expected that such assessments would exhibit lower levels of reliability, the effect size is not known.

Chart 6.2.1: Distribution of test maxima

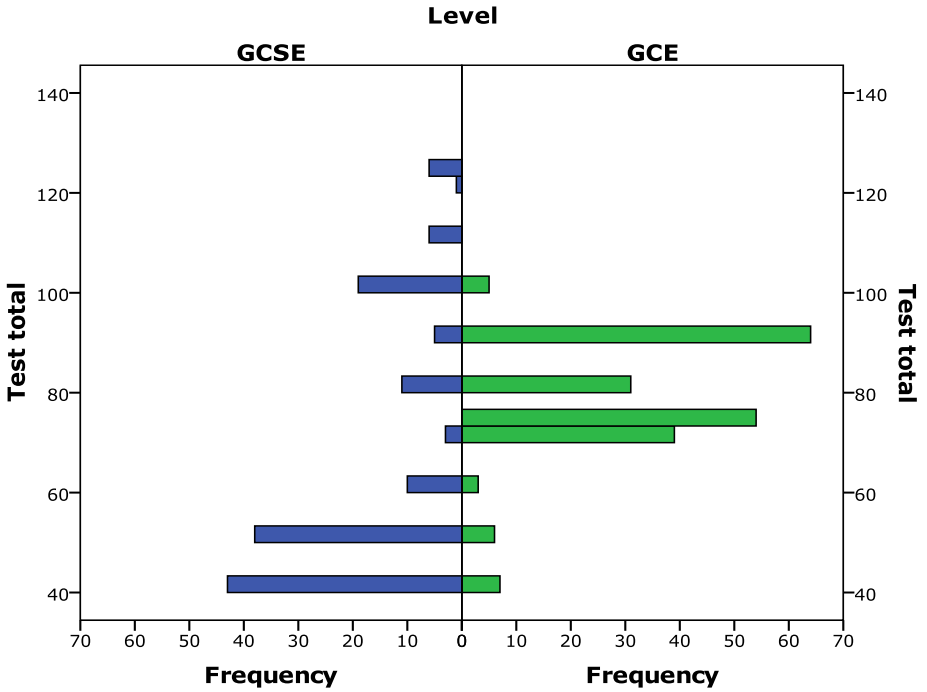


Chart 6.2.2: Distribution of number of items

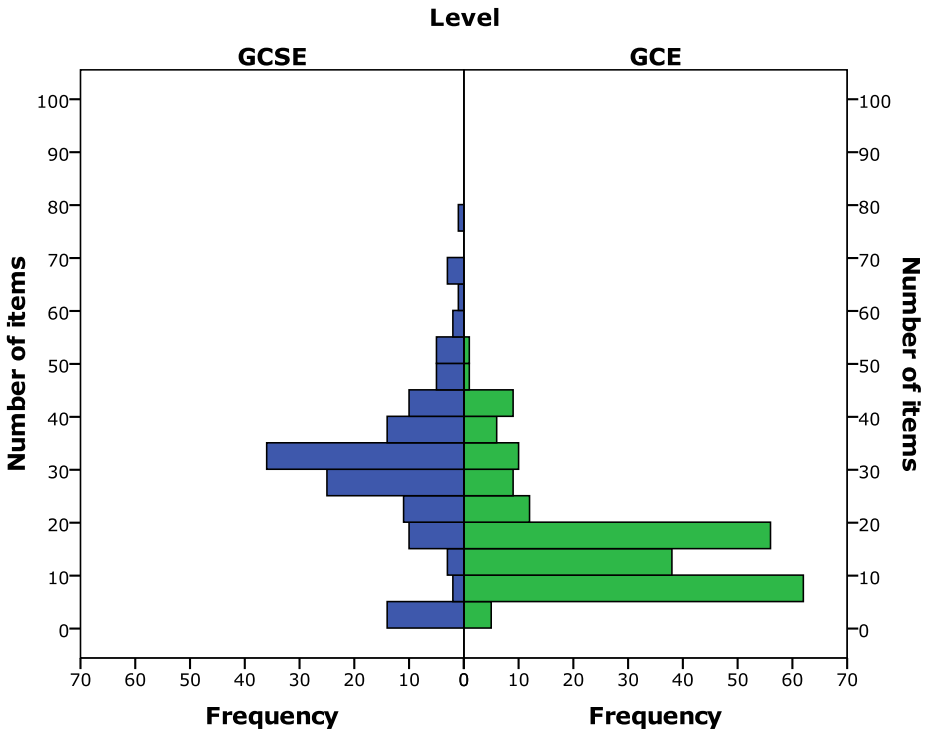
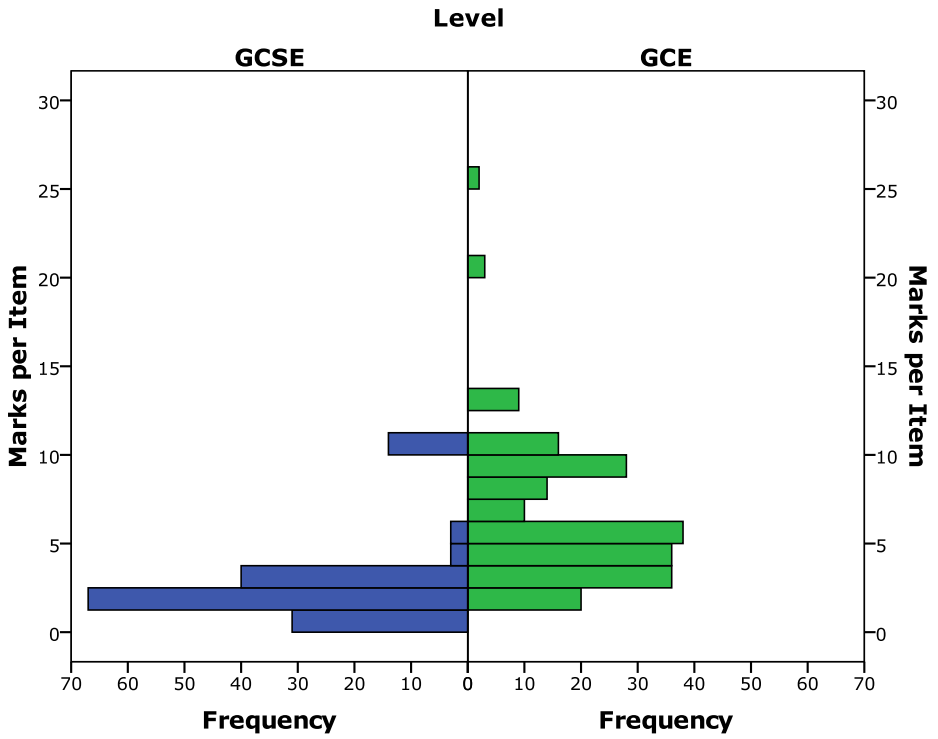


Chart 6.2.3: Distribution of mean marks per item



GCSE assessments tend to have shorter questions than GCE assessments. 60% had less than 5 marks per item and 80% had less than 10 marks per item.

Given the differences in structure of the assessments analysed, differences in levels of internal consistency reliability are to be expected.

6.3. Descriptions of the data sets

Values of reliability coefficients may depend on:

- the length of the test;
- the number of marks per item;
- the dispersion of student scores;
- test targeting;
- The amount of subjectivity in the marking.

While the first two factors are fixed for each of the assessments to be investigated, it was necessary to review the data to ensure that the last three factors provide sufficient scope for test reliability to be estimated effectively.

For a given level of error variance, lower dispersion of observed scores would be expected to result in a lower ratio of true score and observed score variance and hence lower reliability estimates.

Test targeting is the interaction between the test and the population of candidates taking the test. A well-targeted test is one that presents an appropriate level of difficulty for the population as a whole and successfully discriminates between candidates operating at different levels with respect to the content domain. Questions that are too easy or too hard effectively shorten the portion of the test that gives information about candidates' ability and might therefore be expected to reduce reliability.

The impact of the amount of subjectivity in the marking may be more to do with the nature of the questions than the mode of marking but the factors are inseparable in that responses to open questions tend to require judgement in marking.

Chart 6.3.1 shows the distribution of mean percentage scores on the assessments. The chart shows that test targeting varied considerably across the assessments. However, table 6.3.1 shows that the correlations between mean score and the coefficients were low, ranging from 0.02 for λ_4 to 0.28 for ω_t. The Rasch measures of separation reliability showed negative correlation with mean percentage scores, somewhat different behaviour to the other measures. These results indicate that test targeting does not have a large impact on the value of reliability coefficients.

Chart 6.3.1: Distribution of mean percentage scores

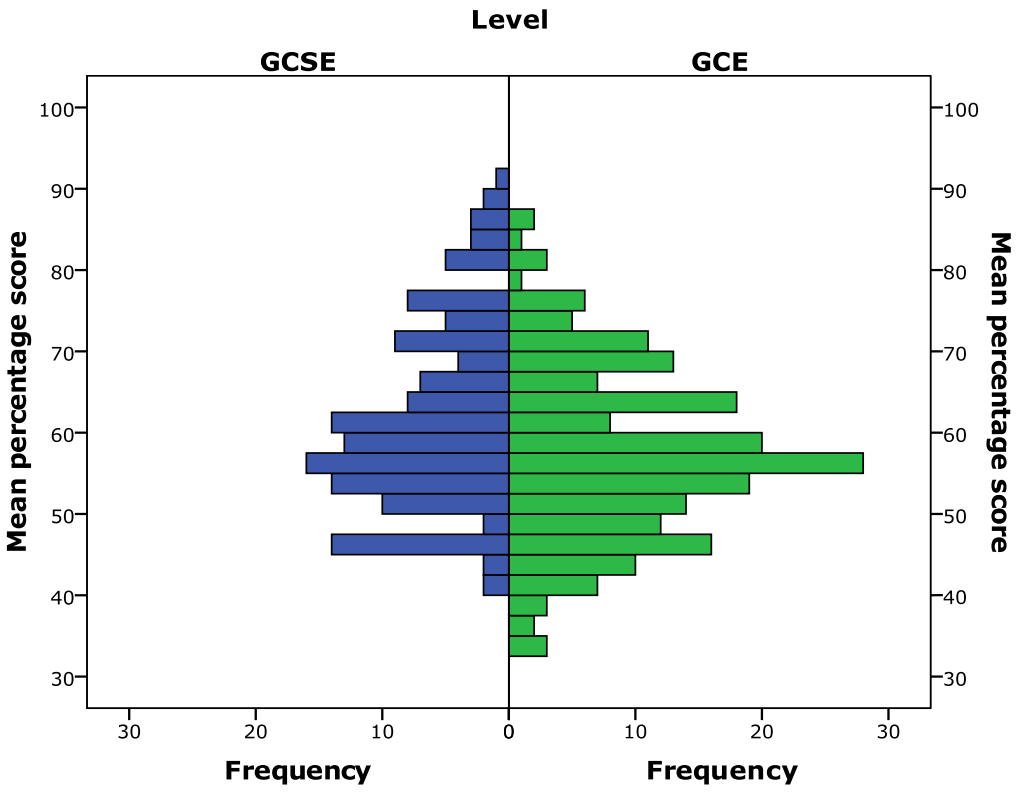


Table 6.3.1: **Correlations of reliability coefficients with mean percentage score**

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	ω_t	R_R	R_M
0.05	0.16	0.09	0.02	0.20	0.25	0.28	-0.12	-0.15

To compare the spread of scores across all the assessments, the standard deviations were expressed as percentages of the total score. Chart 6.3.2 shows the distribution of percentage standard deviation of scores on the assessments. Table 6.3.2 shows that correlations between percentage standard deviation of score and the coefficients were higher than the corresponding values for correlations between mean percentage score and the coefficients. This illustrates that spread of marks is likely to influence reliability statistics more than test targeting.

Chart 6.3.2: Distribution of standard deviations of scores

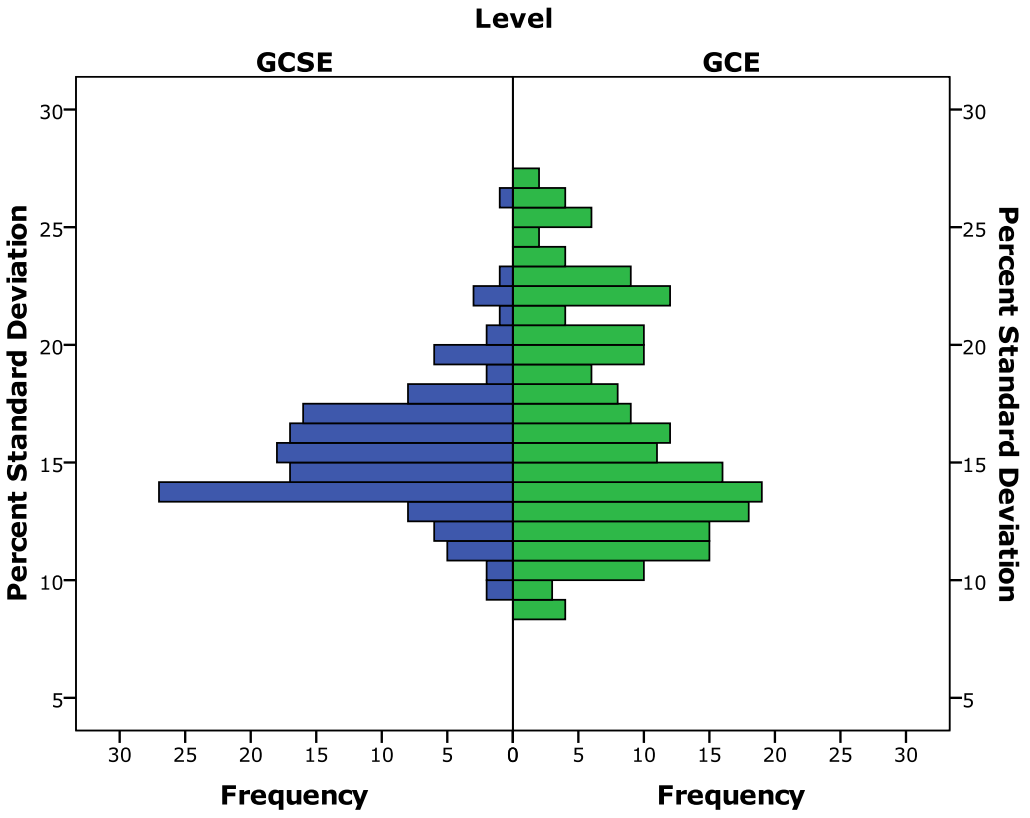


Table 6.3.2: **Correlations of reliability coefficients with percentage standard deviation of score**

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	ω_t	R_R	R_M
0.34	0.57	0.53	0.40	0.60	0.63	0.42	0.50	0.49

6.4. Comparison of coefficients

The extent to which the various coefficients measure the same construct can be seen from the correlations between the measures, given in table 6.4.1. High correlations can be taken to indicate that different coefficients are measuring a similar construct.

Table 6.4.1: Correlations between coefficients

	λ_2	$\lambda_3(\alpha)$	λ_4	λ_5	λ_6	ω_t	R_M	R_R
λ_1	0.93	0.95	0.61	0.89	0.92	0.90	0.65	0.71
λ_2		0.97	0.61	0.98	0.98	0.95	0.69	0.74
$\lambda_3(\alpha)$			0.73	0.92	0.94	0.90	0.65	0.71
λ_4				0.52	0.52	0.54	0.34	0.41
λ_5					0.95	0.95	0.71	0.75
λ_6						0.98	0.77	0.70
ω_t							0.66	0.72
R_M								0.99

Correlations between λ_1, λ_2, λ_3, λ_5, λ_6 and ω_t are all fairly good, as might be expected from the formulae used to compute them. However, λ_4 does not correlate well with any of the other coefficients. Values for α are identical to those for λ_3.

The correlation between the Rasch model separation and the Rasch real separation is very high, as is expected given that they are upper and lower bounds respectively for the same quantity. However, these coefficients do not measure the same construct as the λ or ω coefficients.

Tables 6.4.2 and 6.4.3 give summary statistics for each of the coefficients for GCSE and GCE assessments respectively.

Table 6.4.2: Coefficients for GCSE assessments

	N	Mean	Median	Minimum	Maximum	Range	Std. Dev
λ_1	142	0.79	0.81	0.47	0.93	0.46	0.11
λ_2	142	0.84	0.85	0.57	0.95	0.38	0.08
λ_3	142	0.83	0.84	0.53	0.95	0.41	0.08
λ_4	142	0.74	0.78	0.32	0.93	0.61	0.13
λ_5	142	0.84	0.85	0.59	0.95	0.36	0.07
λ_6	52	0.80	0.84	0.54	0.96	0.41	0.12
ω_t	142	0.89	0.90	0.64	0.97	0.33	0.06
R_R	142	0.82	0.83	0.64	0.95	0.31	0.08

Table 6.4.3: Coefficients for GCE assessments

	N	Mean	Median	Minimum	Maximum	Range	Std. Dev
λ_1	209	0.71	0.72	0.41	0.90	0.50	0.10
λ_2	209	0.80	0.81	0.53	0.94	0.41	0.09
λ_3	209	0.77	0.79	0.46	0.92	0.47	0.10
λ_4	209	0.68	0.73	0.13	0.91	0.77	0.17
λ_5	209	0.81	0.82	0.55	0.92	0.37	0.08
λ_6	169	0.80	0.81	0.49	0.93	0.43	0.09
ω_t	209	0.85	0.86	0.61	0.95	0.34	0.06
R_R	209	0.79	0.79	0.44	0.94	0.50	0.08

Some of the minimum values appear to be very low. However, minimum values are vulnerable to outliers, as are ranges and the low values tend to occur in coefficients that suffer from other disadvantages. Most of the coefficients show similar levels of variation but these statistics indicate that ω_t is consistently higher than the others and shows less variation.

It can be shown that $\lambda_1 < \lambda_3 < \lambda_2$, so λ_2 will always be a 'better' lower bound than either λ_1 or λ_3. Since the most common measure in use is Cronbach's α (equivalent to λ_3), there is little point in considering λ_1 but it is of interest to consider the size of the difference between α and λ_2 before concluding that the latter is preferable as a lower bound of reliability.

There are some suggestions that different coefficients may be more appropriate in particular circumstances, for example, the SPSS help file states that λ_5 *is better than* λ_2 *when there is one item that has a high covariance with the other items, which in turn do not have high covariances with each other.*

However, since each measure is claimed to be a lower bound for true reliability and the greatest of any set of lower bounds is always a better lower bound, it follows that a more rational approach might be to calculate a range of estimates and pick the highest value, rather than try to choose a method to fit the assessment. Such an approach would be time-consuming and perhaps unnecessary if one coefficient could be shown to be consistently higher than the others. Comparisons can be made by considering plots of one coefficient against another. In charts 6.4.1 to 6.4.6, the reference line y=x is plotted to show where one coefficient is greater than the other.

Chart 6.4.1 illustrates the comparison of λ_2 and α, showing that $\lambda_2 > \alpha$ for all the assessments analysed. For values of α below 0.8 values of λ_2 can be up to 0.16 higher. To put this another way, α may underestimate true reliability by more than 0.16 in some cases.

Chart 6.4.1: Comparison of λ_2 and α

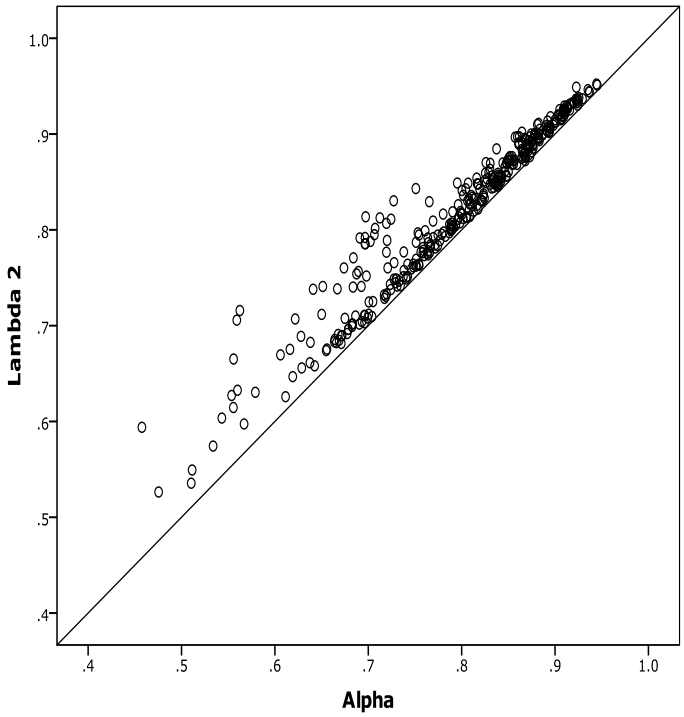


Chart 6.4.2 shows a similar comparison of λ_2 and λ_5 is less clear. Although in general, λ_2 is close to λ_5, $\lambda_5 > \lambda_2$ for about 25% of cases.

Chart 6.4.2: Comparison of λ_2 and λ_5

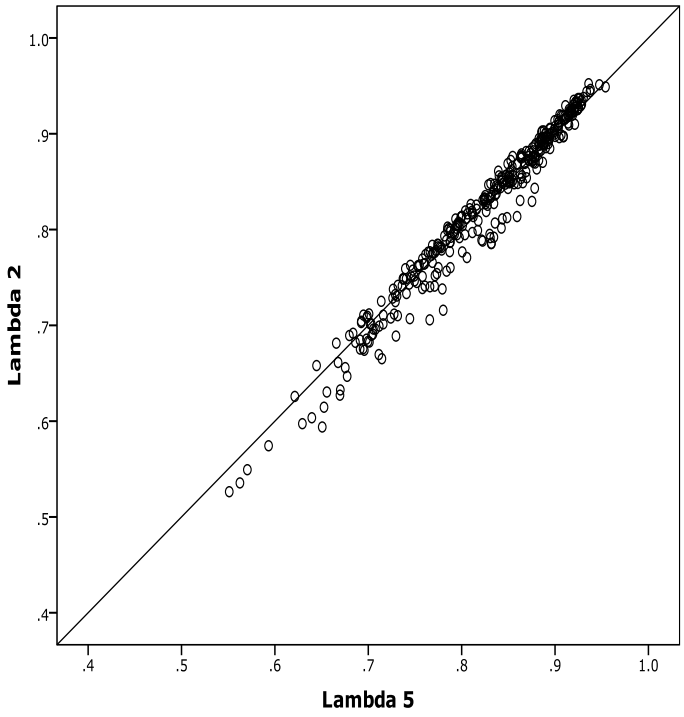


Chart 6.4.3: Comparison of λ_2 and λ_6

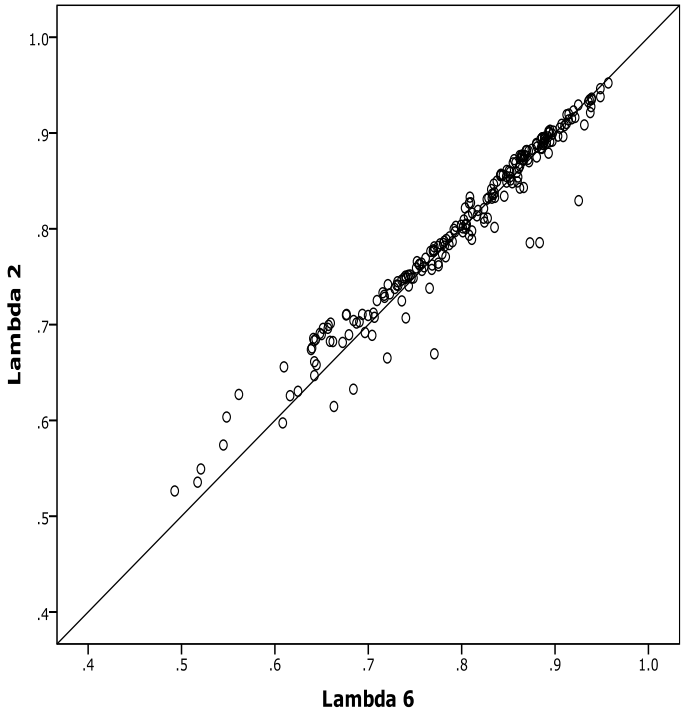
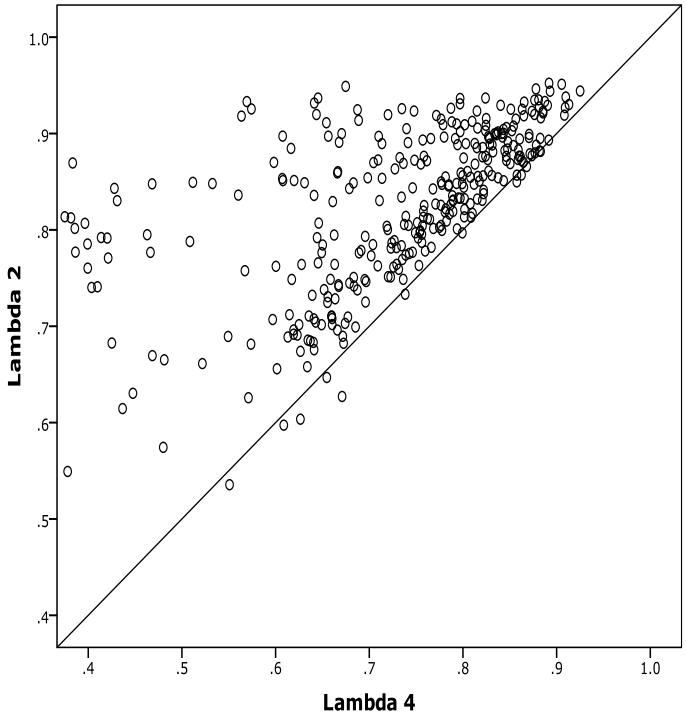


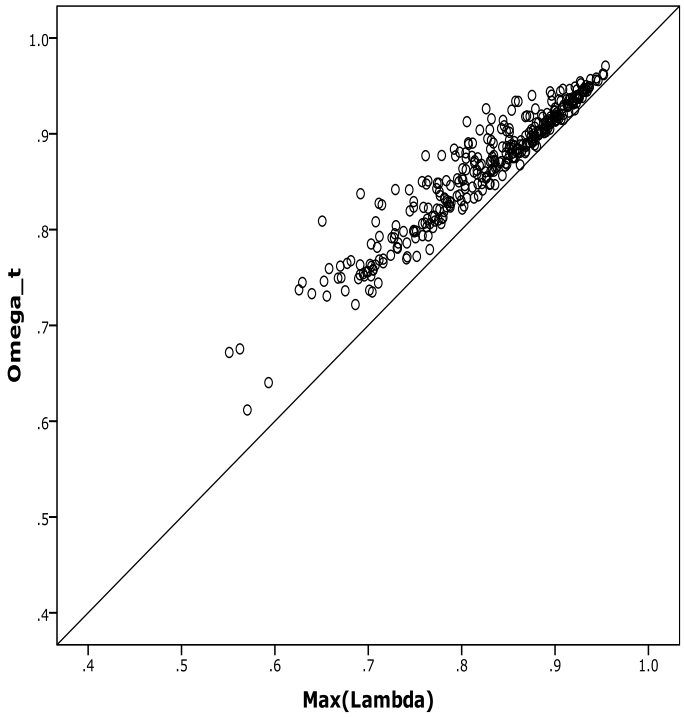
Chart 6.4.4: Comparison of λ_2 and λ_4



Charts 6.4.2 to 6.4.4 show that the maximum of the λ coefficients may be any of λ_2, λ_4, λ_5 or λ_6. Since each coefficient is an estimate of the lower bound for true reliability, it would be reasonable to take the maximum of these values as the best estimate of reliability. The poor correlation between λ_4 and the other coefficients is due in part to the way in which λ_4 is calculated by SPSS. It is merely one possible split-half value and varies according to the order in which the items are entered. This is a severe limitation on its utility as a reliable measure of what it sets out to estimate.

Chart 6.4.5 compares ω_t with Max(λ_i) and it can be seen that ω_t > Max(λ_i) for all of the assessments in this study.

Chart 6.4.5: Comparison of ω_t and Max(λ_i)



The value of ω_t is always greater than the maximum of the λ coefficients with differences ranging from zero to 0.16. It is noticeable that the difference is greater for lower values of Max(λ_i). This makes it a better lower bound for true reliability than any of the λ coefficients or Cronbach's α.

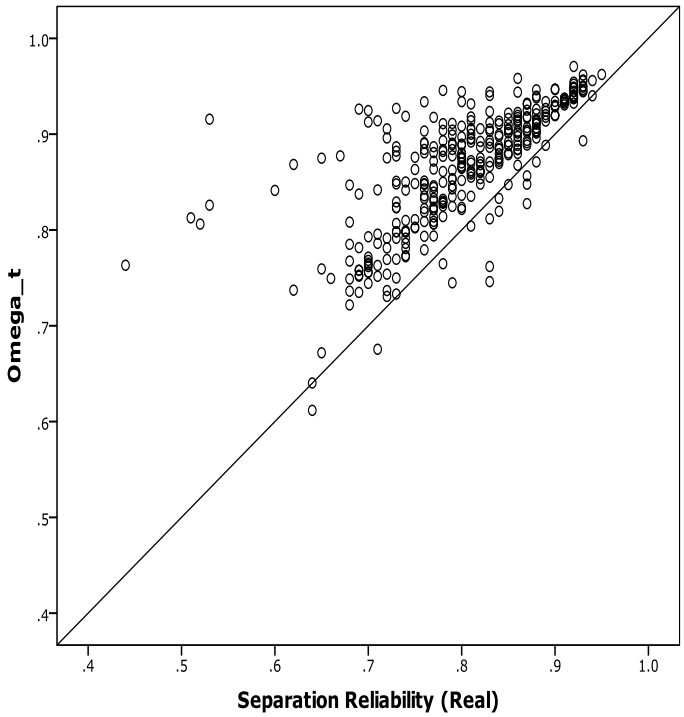
A disadvantage ω_t is that it is more difficult to derive, in that it is not obtainable directly from the SPSS reliability routine. If the reliability routine is to be used, a reasonable estimate could be obtained as Max(λ_i), though this can be expected to underestimate true reliability.

This disadvantage may not be a consideration if an awarding body were to program the analysis rather than use proprietary software such as SPSS.

There can be little doubt that Cronbach's α would be the easiest coefficient to obtain, whatever the analysis platform being used. Max(λ_i) requires the computation of the covariance matrix and ω_t requires a factor analysis of the non-zero variance items in the assessment. Both of these would incur significant development time if they were to be programmed from scratch.

Chart 6.4.6 compares ω_t and Rasch Separation Reliability R_R. The Rasch Reliability (separation index) measures the reproducibility of relative measure location (Linacre 2011), so high reliability means that there is a high probability that persons estimated with high measures actually do have higher measures than persons estimated with low measures. The model reliability is an upper bound and the real reliability is a lower bound to this value, so true reliability would lie somewhere between the two.

Chart 6.4.6: Comparison of ω_t and Rasch Separation Reliability (real)



The Rasch separation reliability measures show relatively poor correlation with the other coefficients and are generally lower than ω_t despite the claims that ω_t is a lower bound and R_M is an upper bound. This can only be reconciled by viewing the coefficients as measuring different constructs, or at least, measuring them on a different scale.

While there may be much to recommend the psychometric analysis of item data to inform test development, neither R_M nor R_R would be a suitable alternative to the other coefficients discussed in this report.

6.5. Consistency over time

Of the 165 assessment types that were analysed, a total of 110 appeared in two or more series, some with multiple versions. For each of these cases, the range of values of ω_t was calculated and tabulated (Table 6.5.1). In approximately 65% of cases, the difference over series was less than 0.03.

Table 6.5.1: Range of values of ω_t

Range	Number of assessments
<0.01	23
0.01 -0.02	18
0.02 -0.03	30
0.03 -0.04	13
0.04 -0.05	11
0.05 -0.06	5
0.06 -0.07	4
>0.07	6

where Range = maximum difference between values of ω_t across series.

Table 6.5.2 groups the assessments by subject and level. There were 21 subjects at each of GCSE and GCE with assessment numbers ranging from 2 to 24. This encompasses different papers within the same subject, irrespective of the structure of the papers. A wider range is therefore to be expected.

Table 6.5.2: Distribution of ranges of ω_t

	Number of assessments	
Range	GCE	GCSE
<0.01	0	1
0.01 -0.02	3	2
0.02 -0.03	1	5
0.03 -0.04	2	0
0.04 -0.05	2	4
0.05 -0.10	6	6
0.10 -0.15	5	2
0.15 -0.20	1	0
0.20 -0.25	1	1

6.6. Factors affecting the value of the coefficients

It is well-known that reliability estimates are affected by test length. The Spearman-Brown prediction formula (see appendix 1) provides a way of estimating the reliability of a test after changing the test length, assuming the added items are comparable in properties to the original test items. However, there are several other factors that contribute to the value of each coefficient. Charts 6.6.1 to 6.6.3 show the distributions of values for α, λ_2 and ω_t for GCSE and GCE. In each case reliabilities for GCSE assessments tend to be higher than those for GCE and to exhibit less spread. For all three coefficients, most values are in the range 0.7 – 0.9. It should be noted that there were very few cases for GCSE in series 1 but there were more cases each year. There was no other evidence of a series effect.

In the box-and-whisker diagrams presented in charts 6.6.1 – 6.6.4, the boxes indicate the positions of the upper and lower quartiles, the horizontal line within the box indicates the median and the whiskers indicate the most extreme value or a distance of 1.5 times the interquartile range from the quartile. Values outside the latter bound are shown as outliers. Thus in chart 6.6.1 a total of six values are plotted as outliers.

Chart 6.6.1: Distribution of values of α

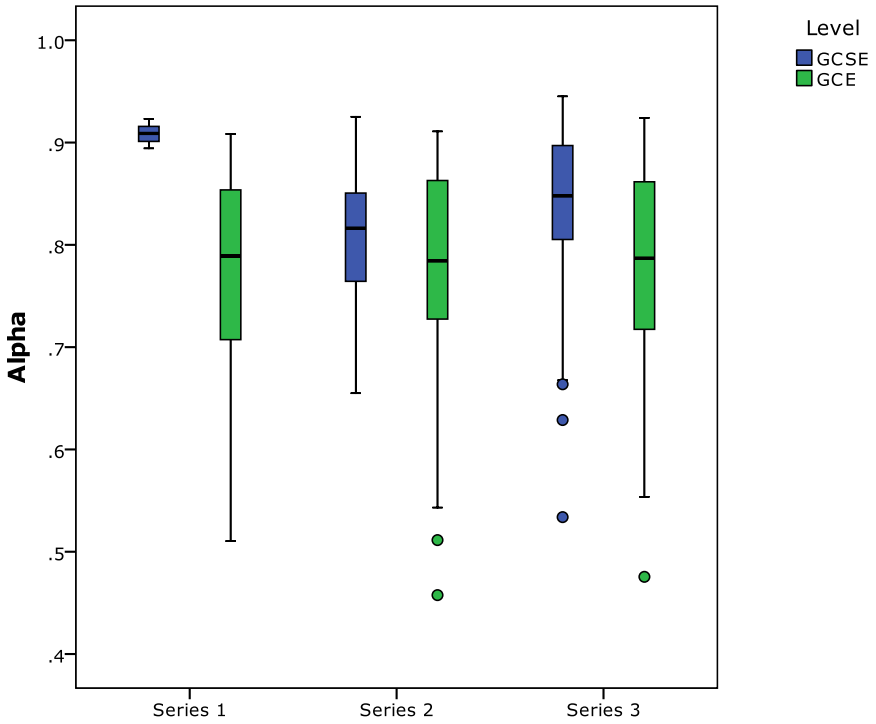


Chart 6.6.2: Distribution of values of λ_2

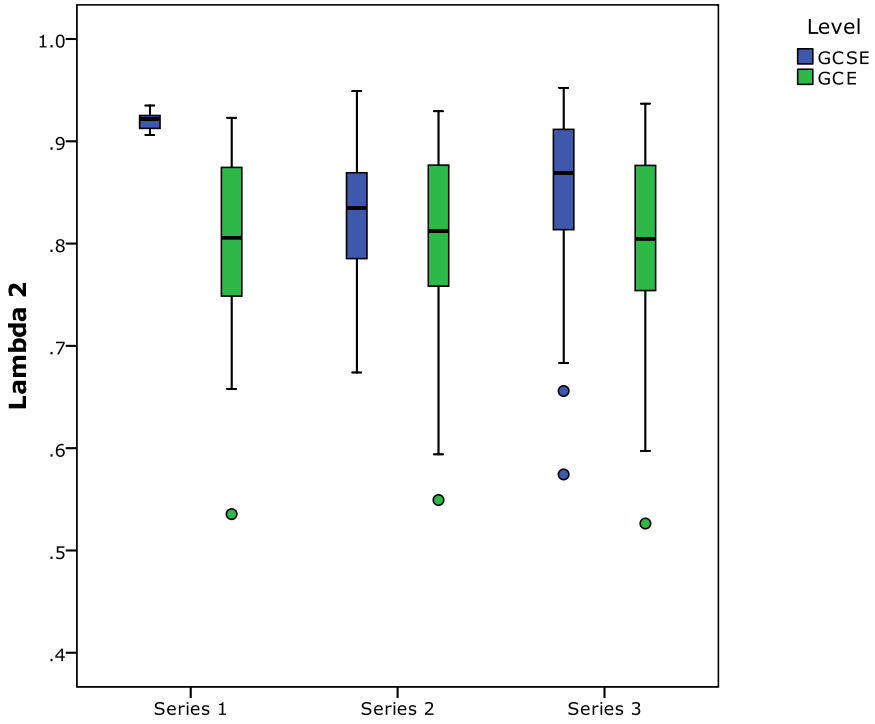


Chart 6.6.3: Distribution of values of ω_t

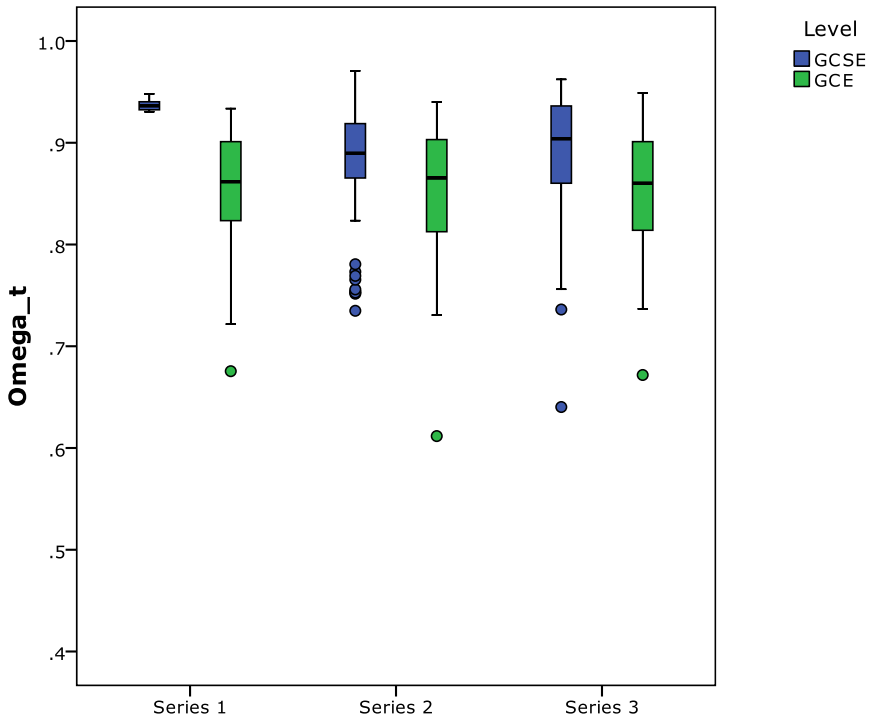
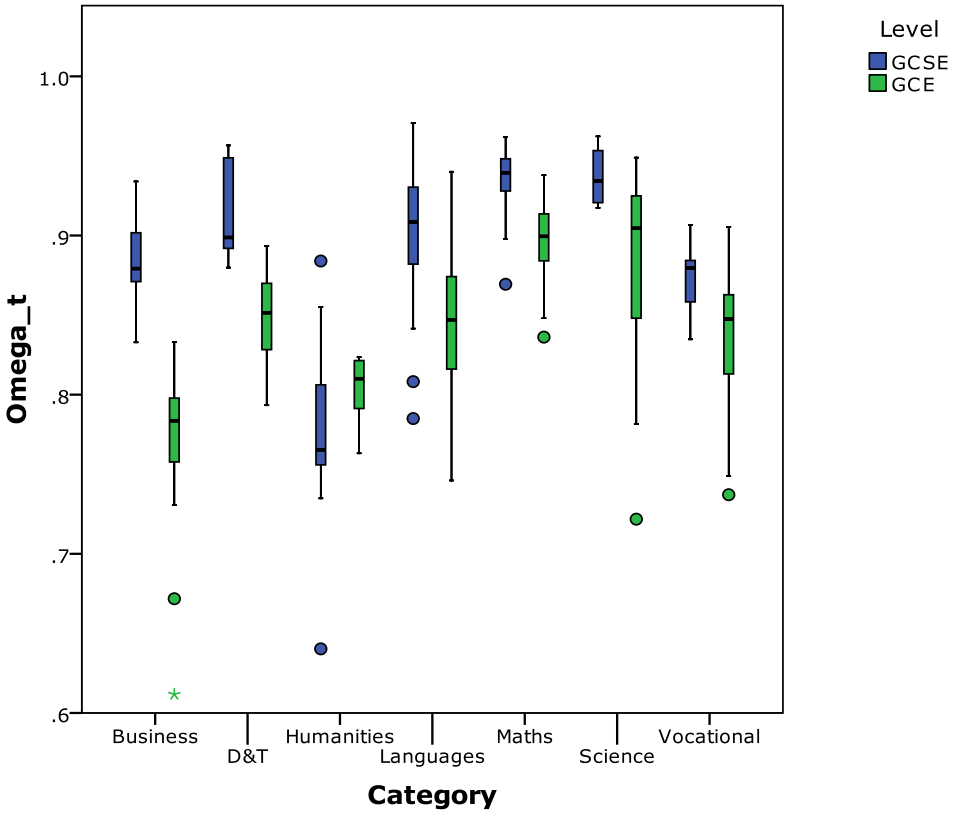


Chart 6.6.4 groups the assessments according to subject area. It can be seen that in all categories other than humanities, reliabilities tend to be higher in GCSE than in GCE. The difference in humanities is due to the particular subjects included at the two levels.

The differences observed are to be expected, given that GCE assessments tend to have fewer items, more marks per item and require more judgement in the awarding of marks.

Chart 6.6.4: Distribution of values of ω_t by subject area



One aspect of reliability relates to the extent to which the assessment measures a single construct. It would be possible to construct a highly reliable test by asking what amounts to the same question many times, although such a test would be of limited value. Nevertheless, in assessments where the questions tend to access a well-defined, coherent set of skills or knowledge, correlations between items would be expected to be high and this would generate high reliability coefficients. On the other hand, assessments that cover a broad range of topics that access different skills and knowledge would be expected to show lower inter-item correlations and consequently, lower estimates of reliability.

This tension between coherence and breadth in an assessment inevitably means that a reliability value that can be achieved in one subject may not be possible in another without destroying the validity and authenticity of the latter. This limits the value of comparisons of reliability across different subjects and different assessment designs.

Quantitative subjects tend to exhibit higher reliability values than qualitative subjects, undoubtedly because, for example, the solution to an equation is usually the same no matter how many times you solve it, but your response to an essay prompt may be different every time. It is therefore inevitable that any measure intended to estimate the likelihood of attaining the same result on a different occasion will produce different ranges of values depending on the level of subjectivity required in the student's response.

Chart 6.6.4 illustrates that it would be a mistake to set parameters for expected values of a reliability coefficient without taking into account the content of the assessment.

Other factors can be categorised as related to test targeting, test structure and item type. To investigate these effects, several statistics were computed and regression analysis used to identify significant variables.

The following statistics were considered as explanatory variables for the coefficients:

Test targeting:

- Test Mean
- Test mean as a percentage
- Test Standard Deviation
- Test Standard Deviation as a percentage

Test Structure

- Number of items
- Total marks available
- Statistics on number of marks per item
  - Mean
  - Maximum
  - Minimum
  - Range
  - Standard deviation
  - Skewness
  - Kurtosis

Item type

- Proportion of marks associated with open/closed responses
- Number of marks associated with open/closed responses

Correlations of these variables with the coefficients can give an indication of which variables may have a significant impact on the value of the reliability coefficients. The correlations differ significantly between the GCSE and GCE subjects, so the correlations are presented separately. Tables 6.6.1 and 6.6.2 are in descending order of the correlation with ω_t. Non-significant values have been shaded.

Table 6.6.1: GCSE Correlations

Variable	ω_t	α	λ_2
Number of items	0.77	0.75	0.74
Mean score	0.56	0.62	0.63
Std. Deviation of scores	0.52	0.60	0.61
Maximum marks available	0.46	0.53	0.54
Number of Objective Marks	0.45	0.36	0.34
Std. Deviation of score as percent	0.43	0.53	0.53
Percent of Objective marks	0.26	0.16	0.13
Skewness of number of marks per item	0.17	0.03	0.04
Number of Subjective Marks	0.14	0.24	0.26
Kurtosis of number of marks per item	0.13	0.02	0.02
Mean percentage score	0.06	-0.02	-0.03
Range of number of marks per item	-0.42	-0.41	-0.35
Maximum number of marks per item	-0.59	-0.56	-0.52
St. Deviation of number of marks per item	-0.66	-0.62	-0.58
Minimum number of marks per item	-0.69	-0.63	-0.66
Mean number of marks per Item	-0.70	-0.62	-0.63

Table 6.6.2: GCE Correlations

Variable	ω_t	α	λ_2
Std. Deviation of score as percent	0.65	0.62	0.68
Std. Deviation of scores	0.60	0.53	0.57
Mean percentage score	0.32	0.12	0.28
Number of items	0.25	0.19	0.23
Number of Objective Marks	0.19	-0.03	0.12
Mean score	0.19	-0.04	0.08
Percent of Objective marks	0.18	-0.04	0.12
Minimum number of marks per item	0.00	0.10	0.03
Kurtosis of number of marks per item	-0.03	-0.33	-0.13
Maximum marks available	-0.13	-0.19	-0.23
Skewness of number of marks per item	-0.13	-0.40	-0.24
Mean number of marks per Item	-0.19	-0.17	-0.22
Number of Subjective Marks	-0.22	-0.15	-0.27
Maximum number of marks per item	-0.33	-0.64	-0.49
Range of number of marks per item	-0.34	-0.69	-0.51
St. Deviation of number of marks per item	-0.42	-0.68	-0.57

At both levels, the standard deviation of scores was positively correlated with reliability as were the number of items in the test. Expressing the standard deviation as a percentage made little difference. The mean score also correlates with α, λ_2 and ω_t for GCSE and the mean percentage score with λ_2 and ω_t for GCE. Factors that appear to have a negative impact on reliability are the mean number of marks per item and the spread of item tariffs.

Regression analyses were carried out for each of the two levels to examine whether information available before an assessment went live could be used to estimate reliability. Tables 6.6.3 and 6.6.4 show the regression output for ω_t based on a set of six measures. Two observations stand out:

- The coefficients are all very small, indicating that though the factors are significant they might be considered unimportant.
- The factors with the largest influence are those that are only available post-hoc.

The model for GCSE predicts 79% of the variance in ω_t but the model for GCE only predicts 32%. Furthermore, the coefficients for some variables differ in sign from the correlations, suggesting that the models may be fitting the data without providing any predictive validity.

This does not appear to be a fruitful line of enquiry and is not pursued further in this report.

Table 6.6.3: Regression model for GCSE

	Unstandardised Coefficients		Standardised Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	0.838	0.017		49.19	0.00
Number of Items	-0.002	0.001	-0.38	-2.53	0.01
Maximum marks available	0.002	0.000	0.92	7.33	0.00
St. Deviation of marks per item	-0.051	0.007	-0.84	-7.85	0.00
Mean marks per Item	-0.004	0.002	-0.14	-1.67	0.10
Skewness of marks per item	0.049	0.011	1.10	4.31	0.00
Kurtosis of marks per item	-0.006	0.001	-0.74	-3.67	0.00
Number of Objective Marks	-0.001	0.000	-0.17	-1.64	0.10

Table 6.6.4: Regression model for GCE

	Unstandardised Coefficients		Standardised Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	0.864	0.027		32.51	0.00
Number of Items	0.001	0.001	0.23	1.30	0.20
Maximum marks available	-0.001	0.000	-0.11	-1.47	0.14
St. Deviation of marks per item	-0.011	0.002	-0.63	-5.49	0.00
Mean marks per Item	0.008	0.002	0.51	3.88	0.00
Skewness of marks per item	-0.019	0.007	-0.37	-2.56	0.01
Kurtosis of marks per item	0.007	0.002	0.45	2.99	0.00
Number of Objective Marks	0.002	0.001	0.26	1.68	0.10

6.7. Effect of choice of coefficient on standard error of measurement

The standard error of measurement (SEM) is an estimate of the error in observed scores. Although SEM is of limited value for component parts of awards where separate grades are not awarded, it may be useful to review the impact of the choice of reliability estimate on this measure. SEM is calculated as

$$\text{SEM} = \sigma\sqrt{(1-r)} \qquad (4)$$

Where σ is the standard deviation of test scores and *r* is the reliability coefficient

The size of the estimate of SEM depends on the choice of reliability coefficient and this in turn will produce different results for measures such as classification accuracy and classification consistency.

The usual approach is to use α as the estimate of reliability but since it has been shown that α underestimates true reliability, this will lead to inflated estimates of SEM. Even if the intention is to make a conservative estimate of SEM, that is, an upper bound for the value, this would still be achieved by taking a better lower bound for reliability.

Table 6.7.1 illustrates that a consequence of taking an unnecessarily low estimate of reliability would be to underestimate classification accuracy for the assessment

The effect of using ω_t in place of α can be considerable, reducing the size of the estimate by between 7% and 49% with the average reduction being 19% for these assessments.

Table 6.7.1: Summary statistics for SEM

	Level	N	Mean	Minimum	Maximum
SEM based on α	GCSE	142	3.84	1.76	6.75
	GCE	209	5.65	2.97	11.43
	All	351	4.92	1.76	11.43
SEM based on ω_t	GCSE	142	3.10	1.15	5.57
	GCE	209	4.56	2.55	7.52
	All	351	3.97	1.15	7.52
Percentage change in SEM	GCSE	142	20	7	40
	GCE	209	19	7	49
	All	351	19	7	49
Difference in SEM	GCSE	142	0.74	0.17	1.68
	GCE	209	1.09	0.28	5.40
	Total	351	0.95	0.17	5.40

Chart 6.7.1 shows the distribution of values of SEM based on Cronbach's α for the 351 assessments. Chart 6.7.2 shows the corresponding distribution if the estimates are based on ω_t.

Chart 6.7.1: Distribution of SEM values based on Cronbach's α

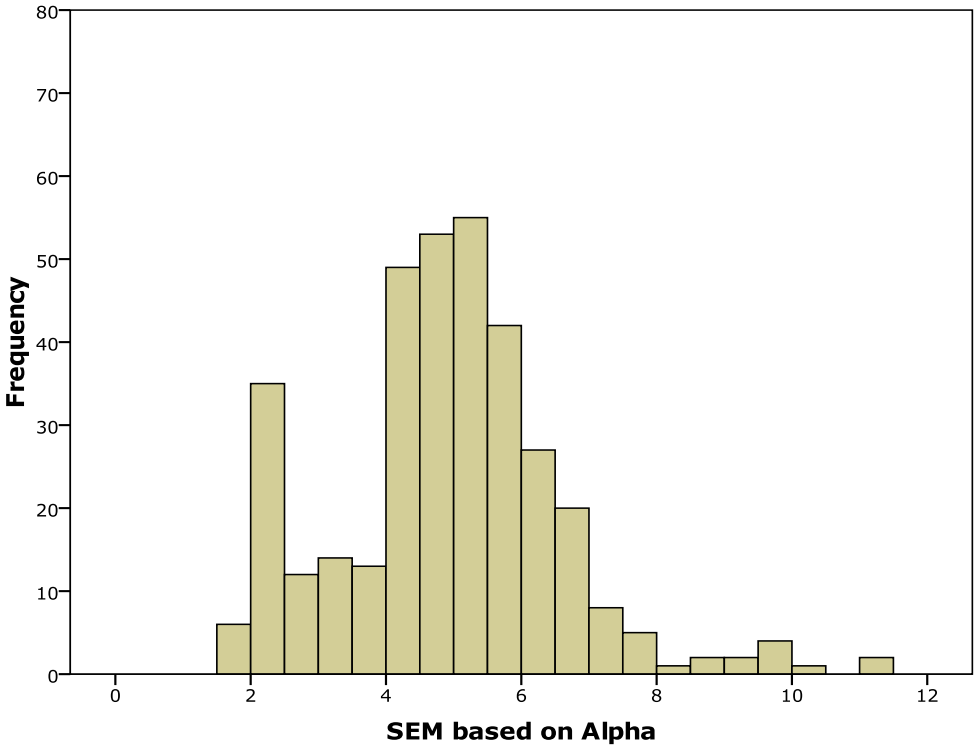
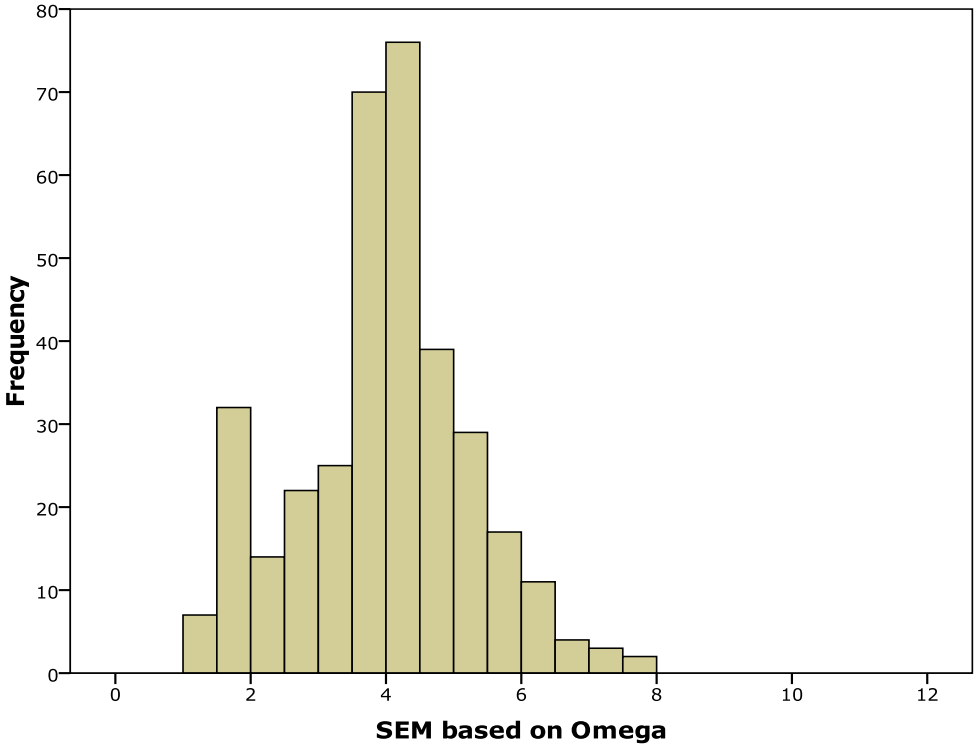


Chart 6.7.2: Distribution of SEM values based on ω_t.

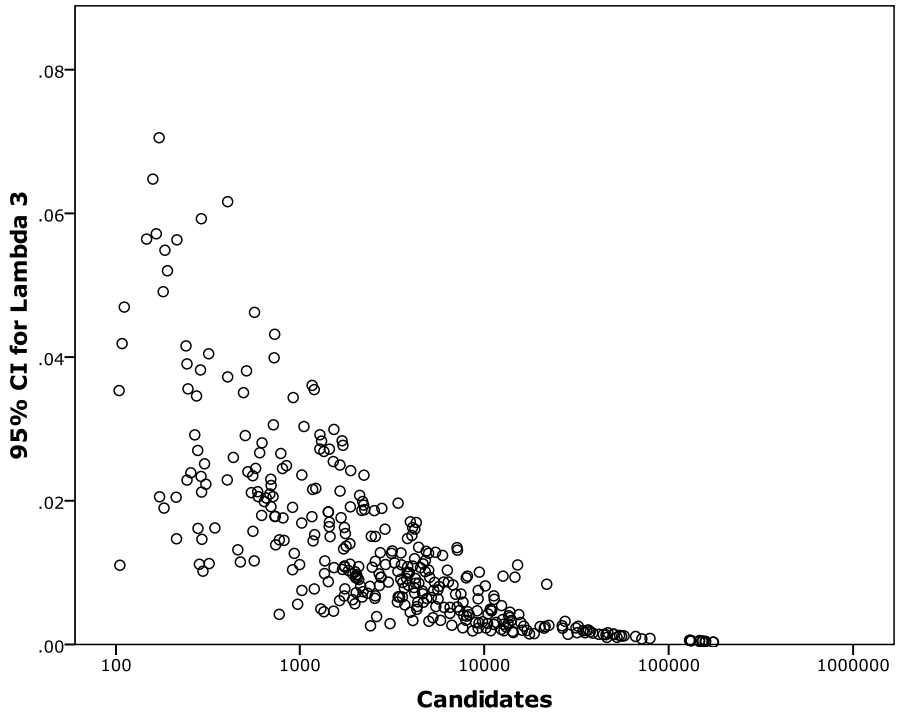


7. Confidence intervals

Confidence intervals were computed for the Guttman λ coefficients using the bootstrap method. Since α is equivalent to λ_3, estimates for λ_3 also provide estimates for α.

Chart 7.1 illustrates the relationship between the 95% confidence interval width and the number of candidates for λ_3. The horizontal axis is a logarithmic scale to accommodate the large range of entry numbers. For low entry numbers, there is considerable variation in the widths. Similar patterns appear for the other coefficients.

Chart 7.1: Confidence interval widths for λ_3 (α)



As would be expected, confidence interval widths vary with $\frac{1}{\sqrt{N}}$ where N = the number of candidates. They also tend to decrease as the value of the coefficient increases.

A regression analysis revealed that approximately 65% of the variance in confidence interval widths can be explained by the number of candidates and the value of the coefficient.

In tables 7.1 and 7.2 assessments are grouped by order of magnitude of the entry numbers. Computing the maximum value of the confidence interval width for each group gives conservative estimates for confidence intervals.

Table 7.1: Approximate 95% Confidence interval widths for GCSE assessments

	Number of Candidates			
GCSE	100-999	1000-9999	10000-99999	Over 100000
λ_1	0.069	0.027	0.010	0.001
λ_2	0.064	0.028	0.012	0.001
λ_3 (α)	0.071	0.028	0.011	0.001
λ_4	0.100	0.038	0.015	0.001
λ_5	0.069	0.029	0.013	0.001
λ_6	0.062	0.031	0.013	0.001

These are maximum values for the groups and therefore represent upper bounds for the confidence interval widths.

- For sample sizes of 1000 or more the 95% confidence intervals for these coefficients are likely to be less than ±0.03.
- For sample sizes of 10000 or more the 95% confidence intervals for these coefficients are likely to be less than ±0.01.
- For sample sizes of 100000 or more the 95% confidence intervals for these coefficients are likely to be less than ±0.001, that is, there is no need to calculate them.

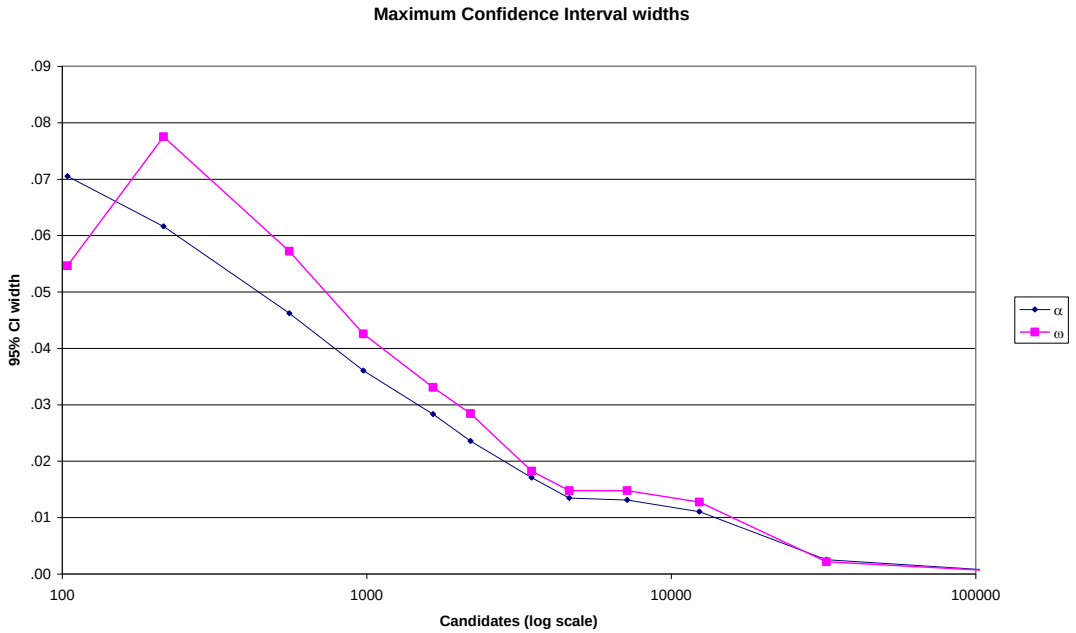
Table 7.2: Approximate 95% Confidence interval widths for GCE assessments

	Number of Candidates			
GCE	100-999	1000-9999	10000-99999	Over 100000
λ_1	0.061	0.032	0.007	-
λ_2	0.072	0.041	0.008	-
λ_3 (α)	0.065	0.036	0.007	-
λ_4	0.116	0.048	0.010	-
λ_5	0.078	0.045	0.008	-
λ_6	0.078	0.043	0.005	-

- For sample sizes of 1000 or more the 95% confidence intervals for these coefficients are likely to be less than ±0.05.
- For sample sizes of 10000 or more the 95% confidence intervals for these coefficients are likely to be less than ±0.01.
- None of the GCE assessments had sample sizes of 100000 or more.

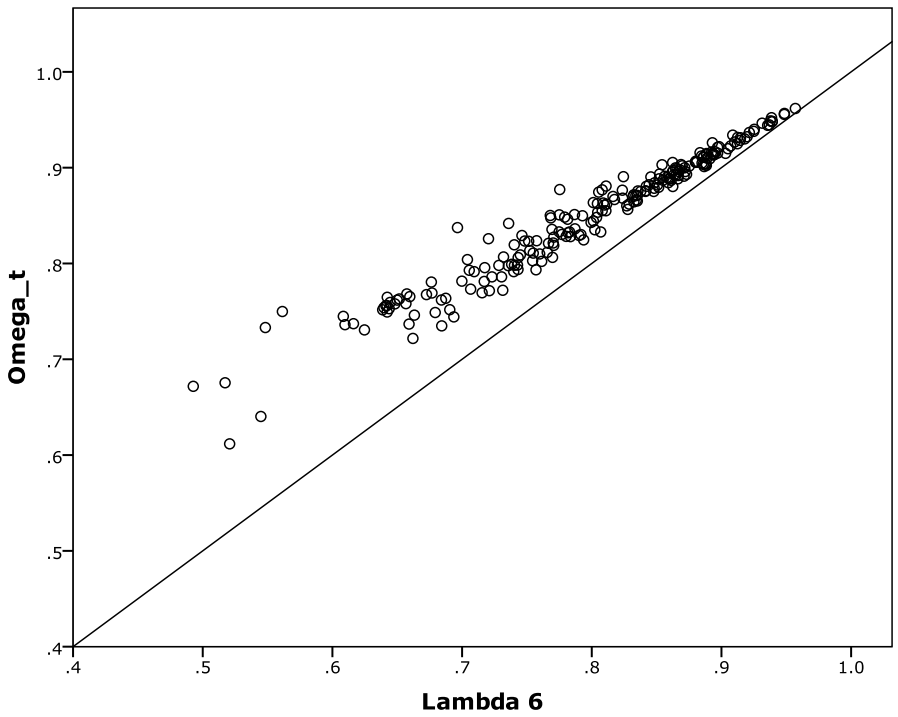
Since the widths were similar for both levels, they were aggregated across the levels and Chart 7.2 illustrates the relationship between maximum confidence interval widths and entry sizes aggregated across the two levels.

Chart 7.2: Estimated Confidence interval widths



Computation of bootstrap estimates of confidence intervals for ω_t proved to be computationally intensive and may not be particularly practicable. The high correlation (0.98) between ω_t and λ_6 suggests that confidence intervals for ω_t might be expected to be very similar to those for λ_6. However, the plot of ω_t against λ_6 in chart 7.3 illustrates that the former shows less spread than the latter with the ration of standard deviations being of the order of 2:3. Confidence intervals for λ_6 could therefore be taken as upper bounds for the confidence intervals for ω_t.

Chart 7.3: Comparison of ω_t and λ_6



8. Conclusions

The ubiquity of Cronbach's α and its ease of calculation puts it in the position of being the benchmark against which other coefficients can be measured, but there is clear evidence that there are indices that give better estimates of true reliability.

There are many factors that combine to influence the expected level of any given reliability coefficient and comparisons between subjects are unhelpful. Where comparisons should be made is across series and estimated confidence intervals could be used to identify where assessments ostensibly built to the same specification differed significantly over time.

There is little to be gained from using a coefficient that significantly underestimates reliability as this may lead to conclusions that the assessment is unsatisfactory when in fact it is functioning within acceptable parameters. The requirement should therefore be to use the coefficient with minimum variance that provides the best lower bound for reliability.

λ_1 is of little value other than as a step to calculating the other coefficients. In particular, since it is always lower than α, there is no point in considering it as an alternative.

λ_4 is calculated as one of the possible splits of the test and varies according to the order in which the items are entered. According to Revelle & Zinbarg (2009) the maximum value of λ_4 may exceed ω_t but it is impractical for awarding bodies to have to compute the value many times in order to find the maximum in all of the assessments reviewed. For the particular split computed by SPSS λ_4 never exceeded ω_t. Practical considerations exclude this coefficient as an improvement on α.

λ_6 could not be calculated for over a third of the assessments reviewed. This alone makes it unsuitable as a straight replacement for α.

λ_2 was always greater than α. It could be taken as an improvement on α. However, λ_5 was better than λ_1 on around 25% of the assessments reviewed. This leads to the recommendation that α could be replaced by Max(λ_i) though the evidence suggests that this would still underestimate true reliability.

ω_t was more difficult to derive than Max(λ_i) in that it was not obtainable directly from the reliability routine, requiring the extraction of communalities from a factor analysis on the non-zero variance items, together with item and test variances. Nevertheless, it is clearly a better lower bound for true reliability than α or Max(λ_i). Together with the improvement in estimates of SEM that would be gained by adopting ω_t as the default measure of reliability may justify the additional work. Furthermore, awarding bodies may wish to program the analysis on their own system rather than use proprietary software such as SPSS, removing much of the advantage of the λ coefficients.

No evidence was found to support the idea that different coefficients could be identified as more appropriate for different assessments. Since each measure is claimed to be a lower bound for true reliability and the greatest of any set of lower bounds is always a better lower bound, it follows that a more rational approach would be to calculate a range of estimates and pick the highest value, rather than try to choose a method to fit the assessment. However, the empirical evidence shows that ω_t will be the highest of all the coefficients.

The complex interactions between factors makes building a predictive model from information available before an assessment takes place unsatisfactory. The models tended to fit the data without providing predictive validity and were unlikely to be applicable to datasets from other awarding bodies. For this reason this approach has not been pursued in this report.

Significant influences on test reliability were found to relate to test targeting and, in particular, the spread of scores obtained by the candidates. Reliability tends to be higher where item tariffs are less disparate, however this effect can be overtaken if there are sufficient low-tariff items.

9. Recommendations

1. The findings of this report point clearly to McDonald's ω_t being the coefficient of choice for all the forms of assessment considered.

2. If awarding bodies are to routinely undertake reliability analyses of all assessments and track values for each specification over time, further investigation of the practical implications of applying the coefficient will be required.

3. Comparisons should be made between parallel assessments of the same specification across series and significant differences should trigger an investigation of the reasons for the differences.

4. Where differences have been identified, two lines of enquiry should be followed:

    a. Comparisons of test-taking populations.
    b. Comparison of test structures, content and range of difficulty of questions

10. References

Bramley, T. & Dhawan, V. (2010). Estimates of reliability of qualifications. Cambridge Assessment/Ofqual http://www.ofqual.gov.uk/files/reliability/11-03-16-Estimates-of-Reliabilityof-qualifications.pdf.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, Winston.

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

Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika, 10 (4), 255-282.

Kuder, G. F., & Richardson, M. W. (1937). *The theory of the estimation of test reliability*. Psychometrika, 2, 151-160.

Linacre, J. M. (2011). WINSTEPS® *Rasch measurement computer program*. Beaverton, Oregon: Winsteps.com.

McDonald, R. P. (1970). *The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis*. British Journal of Mathematical and Statistical Psychology, 38, 1–21.

Revelle, W., & Zinbarg, R. E. (2009). *Coefficients alpha, beta, omega, and the GLB: Comments on Sijtsma*. Psychometrika, 74, 145–154.

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

11. Bibliography

Bentler, P. M. (2009). *Alpha, distribution-free, and model-based internal consistency reliability*. Psychometrika, 74, 137–143.

Cronbach, L. J., & Shavelson, R. J. (2004). *My current thoughts on coefficient alpha and successor procedures*. Educational and Psychological Measurement, 64, 391–418.

Efron, B. (1987). *Better bootstrap confidence intervals*. Journal of American Statistical Association, 82, 171–185.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman & Hall/CRC.

Feldt, L. S. & Brennan, R. L. (1989). *Reliability*. In R. L. Linn (Ed.) *Educational measurement* (3rd ed., p. 105-146). New York: American Council on Education and Macmillan.

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

Green, S. B., & Yang, Y. (2009b). *Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha*. Psychometrika, 74, 155–167.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.

Guttman, L. (1945). A basis for analyzing test-retest reliability. Psychometrika, 10, 255–282.

Haertel, E. H. (2006). *Reliability*. In R. L. Brennan (Ed.) *Educational measurement* (4th ed., p. 65-110). New York: American Council on Education and Praeger Publishers.

Kelley, K. (2005). *The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrapping as an alternative to parametric confidence intervals*. Educational and Psychological Measurement, 65, 51–69.

Komaroff, E. (1997). *Effect of simultaneous violations of essential τ-equivalence and uncorrelated error on coefficient α*. Applied Psychological Measurement, 21, 337–348.

Li, H. (1997). *A unifying expression for the maximal reliability of a linear composite*. Psychometrika, 62, 245–249.

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

Novick, M. R., & Lewis, C. (1967). *Coefficient alpha and the reliability of composite measurements*. Psychometrika, 32, 1–13.

Oehlert, G. W. (1992). *A note on the delta method*. The American Statistician, 46, 27–29.

Raykov, T. (1997). *Estimation of composite reliability for congeneric measures*. Applied Psychological Measurement, 21, 173–184.

Raykov, T. (2002). *Analytic estimation of standard error and confidence interval for scale reliability*. Multivariate Behavioral Research, 37, 89–103.

Revelle, W. (1979). *Hierarchical cluster-analysis and the internal structure of tests.* Multivariate Behavioral Research, 14, 57–74.

Sijtsma, K. (2009b). *Reliability beyond theory and into practice*. Psychometrika, 74, 169–173.

Ten Berge, J. M. F. (2004). *The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality*. Psychometrika, 69, 613–625.

Thompson B. (Ed.), *Score reliability: Contemporary thinking on reliability* issues (p. 3–23). Thousand Oaks, CA: Sage Publications.

Yuan, K.-H., & Bentler, P. M. (2002). On *robustness of the normal-theory based asymptotic distributions of three reliability coefficient estimates*. Psychometrika, 67, 251–259.

Yuan, K.-H., Guarnaccia, C. A., & B. Hayslip, J. (2003). *A study of the distribution of sample coefficient alpha with the hopkins symptom checklist: Bootstrap versus asymptotics*. Educational and Psychological Measurement, 63, 5–23.

Zimmerman, D. W., Zumbo, B. D., & Lalonde, C. (1993). *Coefficient alpha as an estimate of test reliability under violations of two assumptions*. Educational and Psychological Measurement, 53, 33-49.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). *Cronbach's α, Revelle's β, and McDonald's ω_h: Their relations with each other and two alternative conceptualizations of reliability*. Psychometrika, 70, 123–133.

Appendix 1: Formulae

The following formulae are as specified in the SPSS help files for the case where N persons take a test consisting of k items. For a detailed description of these formulae, see Revelle & Zinbarg. (2009)

S_i^2 and S_P^2 are unbiased estimates of item and person variances respectively.

v_{ij}^2 is the covariance of items i and j

$$\lambda_1 = 1 - \frac{\sum_{i=1}^{k} S_i^2}{S_P^2} \quad (5)$$

$$\lambda_2 = \lambda_1 + \frac{\sqrt{\frac{2k}{k-1}\sum_{j}^{k}\sum_{i<j}^{k} v_{ij}^2}}{S_P^2} \quad (6)$$

$$\lambda_3 = \frac{k}{k-1}\lambda_1 \quad (7)$$

$$\lambda_4 = \lambda_1 + \frac{4\sum_{j}^{k}\sum_{i<j}^{k} v_{ij}^2}{S_P^2} \quad (8)$$

$$\lambda_5 = \lambda_1 + \frac{2\sqrt{\max\sum_{j}^{k} v_{ij}^2}}{S_P^2} \quad (9)$$

$$\lambda_6 = 1 - \frac{\sum_{i=1}^{k} \varepsilon_i^2}{S_P^2} \quad where\ \varepsilon_i^2 = \frac{1}{\left(V^{-1}\right)_{ii}}$$ or the variance of the errors for item i (10)

$$\omega_t = 1 - \frac{\sum_{i-1}^{k} \sigma_i^2\left(1 - h_i^2\right)}{S_P^2} \quad (11)$$

where h_i^2 is the communality of item i from a factor analysis of the k items

Spearman-Brown prediction formula:

$$\rho^{*}_{xx'} = \frac{N\rho_{xx'}}{1+(N-1)\rho_{xx'}} \quad (12)$$

where N is the number of tests combined and $\rho_{xx'}$ is the reliability of the current test.

Note: non-integer values of N can be used to lengthen or shorten the test

Appendix 2: Bootstrap macro

This macro is an adaptation of one given in the SPSS help file.

```
DEFINE reliability_bootstrap (samples=!TOKENS(1)
                              /Modl=!TOKENS(1)
                    /Scl=!TOKENS(1)
                    /indvars=!CMDEND)

COMPUTE dummyvar=1.
AGGREGATE
  /OUTFILE=* MODE=ADDVARIABLES
  /BREAK=dummyvar
  /filesize=N.
!DO !other=1 !TO !samples
SET SEED RANDOM.
WEIGHT OFF.
FILTER OFF.
DO IF $casenum=1.
- COMPUTE #samplesize=filesize.
- COMPUTE #filesize=filesize.
END IF.
DO IF (#samplesize>0 and #filesize>0).
- COMPUTE sampleWeight=rv.binom(#samplesize, 1/#filesize).
- COMPUTE #samplesize=#samplesize-sampleWeight.
- COMPUTE #filesize=#filesize-1.
ELSE.
- COMPUTE sampleWeight=0.
END IF.
WEIGHT BY sampleWeight.
FILTER BY sampleWeight.
RELIABILITY
  /VARIABLES=!indvars
  /SCALE('ALL VARIABLES') ALL
  /MODEL=!Modl.
!DOEND
!ENDDEFINE.
```

The macro can then be called using:

```
reliability_bootstrap
samples=500
Modl = Guttman
Scl = All
indvars=q01 q02…….qnn.
```

Appendix 3: Test structures

Maximum available mark	Number of assessments	
	GCSE	GCE
40	43	7
50	38	6
60	10	3
70	3	39
75	-	54
80	11	31
90	5	64
100	19	5
110	6	-
120	1	-
126	6	-
Total	142	209

Number of items	Number of assessments	
	GCSE	GCE
4 - 5	14	5
6 - 10	2	62
11 - 15	3	38
16 - 20	10	56
21 - 25	11	12
26 - 30	25	9
31 - 35	36	10
36 - 40	14	6
41 - 45	10	9
46 - 50	5	1
51 - 55	5	1
56 - 60	2	
61 - 65	1	
66 - 70	3	
71 - 75		
76 - 80	1	
Total	142	209

Appendix 4: Glossary of assessment terminology

Term	Description
Qualification	GCSE, AS and A level are the only qualifications referred to in this report.
Assessment	A component of a qualification.
GCSE	General Certificate of Secondary Education. Usually taken by 16 year olds.
GCE	General Certificate of Education. Consists of AS and A level.
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. 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.
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).
Centre	The examination centre that the candidate is registered with.
Script	The physical paper or digital image containing a candidate'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 candidate responses.
Raw score	The score obtained by adding up the marks obtained by the candidate on the questions in the unit/component.
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.
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.

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 2013

This publication is also available on our website at www.ofqual.gov.uk

Any enquiries regarding this publication should be sent to us at:

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