Calculating the reliability of complex qualifications

Tom Benton Research Division

Introduction

In order for qualifications to be meaningful they typically need to cover a greater range of curriculum material than could reasonably be assessed within a single paper. For this reason, all current GCSE and A level qualifications consist of multiple assessment components. Candidate achievement across these different elements is then combined in order to determine the final grade they will be awarded.

Estimating the reliability of qualifications that are examined through a composite of multiple assessments creates some challenges as any estimate of reliability must adequately account for the different amounts of weight given to different components. Some possible approaches to this issue are discussed by He (2009). However, a bigger problem arises when candidates have multiple options regarding which assessments will count towards their overall qualification grade. Such a situation could arise due to candidates working towards the same qualification being able to choose between:

- Different tiers
- Different papers covering different optional topics
- Different examination sessions for individual components as was possible in unitised assessment schemes

Previous work examining how the reliability of qualifications with multiple possible routes may be estimated, such as that by Bramley and Dhawan (2013), have addressed this issue by simply focussing on the most common set of options chosen by candidates to achieve a given qualification. The aim of this article is to demonstrate a relatively simple, and highly intuitive method of calculating reliability for such qualifications that includes the results of candidates across all possible routes. This method is exemplified for a very complicated qualification to show the power of the method in circumstances where it would not be feasible to derive estimates of composite reliability for each possible route.

The Qualification

This article focusses on OCR Mathematics A level specification 7890 and the candidates that certificated for this qualification in June 2012. To be awarded an A level, candidates needed to complete four compulsory units (Core Mathematics 1 to 4) and two out of a possible six optional units (two in each of Mechanics, Probability and Statistics and Decision Mathematics). For their optional papers they could either take both papers within the same optional subject area (e.g. both Mechanics papers), or the first paper in two different subject areas (e.g. Mechanics 1 and Decision Mathematics 1). To make matters more complicated they had the option to take a version of these papers within any of four examination sessions (January 2011, June 2011, January 2012 and June 2012)[1].

1. In theory candidates could also take any of these units prior to 2011 but this was rare for those candidates that completed the A level in 2012.

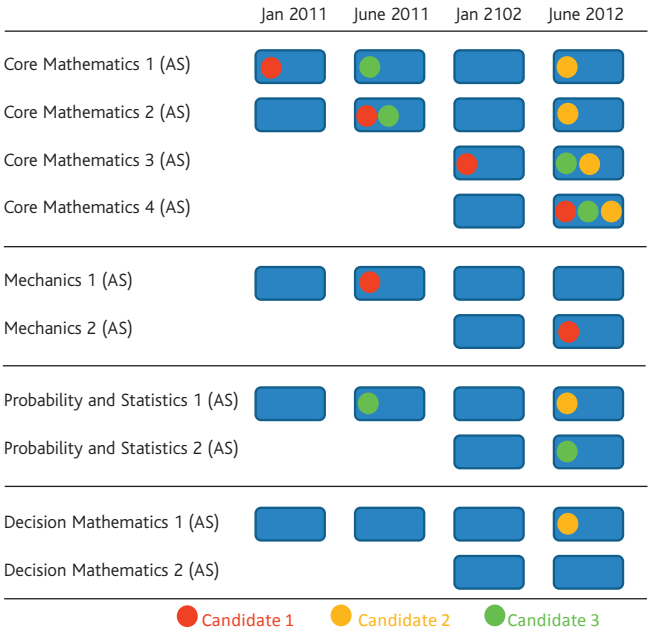


Figure 1: Possible routes through OCR Mathematics A level specification 7890 (assuming all A2 units taken in 2012)

Figure 1 illustrates some possible routes through this qualification. The 30 rectangles represent 30 of the papers available to candidates within this qualification. Note that there were no common questions across the 30 papers. The circles illustrate three different possible combinations of papers that would lead to completion of the A level. Candidate 1 takes the A level in a progressive modular fashion; taking one core unit in each available session and one optional unit (in Mechanics) each June. In contrast, candidate 2 takes a fully linear approach taking all core units and both optional units, this time split across Probability and Statistics and Decision Mathematics, in June 2012. Another option is illustrated by candidate 3; a modular approach but limited to using the June examination sessions[2].

As can be seen from Figure 1, there were an enormous number of options available to candidates. Even if we assume that all of the (more challenging) A2 units were taken towards the end of the course (that is, in 2012 rather than 2011), and that at least one of these A2 units was taken in June 2012 itself, there remains a total of 3,648 possible different combinations of papers that would have led to a Mathematics A level. Note that if, instead, all units had been required to be taken in a linear fashion in June 2012 then the number of possible routes would have reduced substantially, but would still have left six possible combinations of papers leading to the same qualification.

Given that each of the possible routes through the A level will lead to candidates being awarded the same qualification at the same time, it is

2. Note that for the purposes of this article resits are not relevant. For the purposes of this article we need only consider a candidates best performance within any element of the A level. This means that for each candidate we can restrict ourselves to exactly six examination scores.

of interest to calculate the overall reliability of the qualification. Whilst numerous techniques exist for evaluating the reliability of any one of the 30 papers listed in Figure 1 individually, there is little consensus regarding how the reliability of the qualification as a whole should be calculated.

In order to calculate reliability we first need to define what it means. In some senses this is a general problem with any reliability estimate with usual decisions including whether we are interested in reliability in terms of which questions are included – that is, how much difference the exact choice of questions within an assessment has upon the performance of candidates – or marking reliability – that is, how much difference it would make if the same set of responses from each candidate were marked by a different marker. For the purposes of this article we are only interested in reliability relating to the choice of questions[3]. In addition to the usual decisions over the definition of reliability there are some that are particular to the problem in hand. Specifically, we need to determine whether we are interested in:

- how much difference it would make if different questions had been used within each paper but that each candidate's route remained constant,
- or how much difference it would make if candidates had chosen a different route through the same qualification and answered different questions as a result.

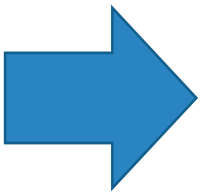
In this article we will concern ourselves with the former of these. That is, we are interested in evaluating how much difference it would make to results if each candidate's route through the qualification remained constant but a different set of questions were included in each of the 30 papers. Putting it another way, we wish to calculate what percentage of the variance in candidates' scores is attributable to their underlying mathematical ability as would be demonstrated if they were able to answer an infinitely large number of questions covering the skills assessed by their chosen route through the qualification.

The Idea

In order to evaluate the reliability of our qualification we will make use of the method of split-halves. Given that the usual definition of assessment reliability is "the consistency of…measurements when the testing procedure is repeated on a population of individuals" (AERA & NCME, 1999), the most intuitive way we might seek to measure reliability is to get candidates to take two versions of the same test and compare their scores. However, this would require candidates to spend additional (and possibly unnecessary) time taking a second version of the same test. To circumvent this issue the split-half procedure instead splits a single question paper into two halves, and then explores the extent to which test scores are 'repeated' from one half of the question in a test to another[4]. If all candidates tend to have similar scores across both halves of the test then we infer that the exact choice of questions has little impact on achievement as the set of questions in one half give a similar result to the entirely different set in the other. Thus, we can be confident that had we written another version of the test and got candidates to take that instead, their results would still be largely unaffected. Conversely, a massive difference between scores on different halves would indicate that candidates' performances were highly dependent upon the precise choice of questions, so that another version of the test may have led to very different results. The formulae used to convert comparisons of scores in different halves into an overall reliability coefficient rely on the correlation (or covariance) of scores between halves, and have been in existence for more than 70 years (see Rulon, 1939).

A simple example of how a test might be split into halves is shown in Figure 2. This figure is based upon the scores for one particular candidate taking the Core Mathematics 1 paper in June 2012. In this case their total score on 10 questions out of 72 is split into two total scores each based on 5 questions and out of 36. Although, this particular candidate has raw scores that are similar across the two halves, a full calculation of reliability would require separate scores on each half to be calculated for each candidate and an estimation of the correlation (or covariance) between the two.

Question	Score
1 (3 marks)	3
2 (5 marks)	3
3 (5 marks)	4
4 (6 marks)	6
5 (6 marks)	4
6 (7 marks)	5
7 (6 marks)	4
8 (8 marks)	7
9 (11 marks)	8
10 (15 marks)	15
Total (out of 72)	59



	Half 1	Half 2
		3
	3	
		4
	6	
	4	
		5
		4
	7	
	8	
		15
Totals (out of 36)	28	31

Figure 2: Example of split half scores for one candidate taking Core Mathematics 1 in June 2012

The great advantage of the split halves technique in our scenario is that it automatically handles the issue of multiple routes through a qualification. Note that all of the possible routes through the Mathematics A level, as defined by Figure 1, require candidates to take exactly six assessments. Thus, if we were to split all 30 assessments into halves thus creating 30 half 1s and 30 half 2s, regardless of a candidate's route through the qualification, they will have scores from exactly six half 1s and six half 2s. Thus if we add up all 'half 1' scores to make one total and all the 'half 2' scores to make another, we can produce two 'half A level' scores for each candidate. These scores can then be compared in the usual way to estimate reliability for the A level as a whole. In applying this technique we are only examining the *reliability* of the scores, that is, the extent to which the achieved scores would be replicable if different questions were used in each paper. The question of whether all possible routes through a qualification are equally valid is not addressed. Furthermore, the overall reliability coefficient generated in this way will essentially provide an average level of reliability across all the possible routes. It does not examine whether particular routes provide a more reliable final score than others.

Note that the technique suggested here could equally well be used to examine the reliability of scores comprised of results in different subjects. For example, we could theoretically apply a similar method to examine the reliability of candidates' UCAS scores that combine A level performance across numerous subjects and are used for university

3. Although as discussed by Benton (2013b), because each question must be marked, it is likely that such estimates will also account for a proportion of marking unreliability.
4. Technically we tend to be interested in changes in standardised scores rather than raw scores. That is, the change in each candidate's score after accounting for any changes in the overall mean and standard deviation of scores across all candidates. This value is of more interest as such overall changes to the score distribution are likely to be accounted for within the process of grade awarding in any case.

applications in the UK. This would address the question of the extent to which applicants' UCAS scores are dependent upon the precise set of questions included in their particular examinations. This would not address the issue of whether all UCAS scores are equally valid predictors of university performance or whether all subjects provide equally reliable scores; it would simply provide an average reliability coefficient across the different subject choices chosen by candidates.

Which split half?

As can be seen from Figure 2, there are numerous possibilities for how we should split a single test into two parts (from now on referred to as 'halves' even though they may not be of equal size). In fact, if we imagine that question 1 is always in half 2, then each other question is either in the same half as question 1 or the opposite side. Thus there are 2^9=512 ways to split this test into halves minus the one split with all questions on the same side. In Figure 2 we have focussed on ensuring that the same number of items and the same number of marks are available in each half. However, although this is intuitively appealing, ensuring similar coverage in terms of the curriculum content and skills required by each half is probably more important.

If we make the (reasonable) assumption that scores on questions measuring the same skills are likely to have stronger associations, then we can encourage each half to measure similar skills by looking for the split that maximises the association between scores on one half and scores on the other. Maximising the association, as measured by covariance rather than by correlation, is advantageous in that it will encourage each half to have a similar score distribution. This approach has been adopted by numerous authors and the resulting reliability coefficient is sometimes referred to as Guttman's λ_4 (after Guttman 1945, see Callender & Osburn 1977; Ten Berg & Socan 2004).

Compared to Cronbach's alpha, Guttman's λ_4 is less likely to underestimate the reliability of a test (Ten Berg & Socan, 2004). On the other hand, there is a danger that, in small samples or with very large numbers of available items, it may grossly overestimate reliability (Ten Berg & Socan, 2004). That is, because it focusses on finding the best split half, it may overestimate the likely similarity between candidates' scores on two real parallel versions of a test. However, this issue was investigated further by Benton (2013a) and is unlikely to be a concern in our scenario as all of the 30 papers investigated contained ten questions or fewer and all but three had sample sizes numbering in the thousands.

Finding the best split half

As explored by Benton (2013a), there are several possible algorithms for identifying the best split half. For the purposes of this study we used the 'start-then-improve' algorithm. As suggested by the name, this algorithm begins with an initial split of the items into two halves (such as based upon an odd and even split) and then examines whether swapping any pair of items from opposite sides will improve the strength of association between the two sides. In essence this means that items that are found to be more strongly associated with the overall score on their own half than the overall score on the opposite half are likely to be swapped across. Once a swap has been made, the algorithm looks for further swaps that may improve the association between scores on opposite sides. This continues until there are no remaining swaps that will improve this association any further.

An example of how this algorithm works in practice is shown in Table 1. Initially the questions are split such that all the odd numbered questions are in half 1 and all the even numbered questions are in half 2[5]. The top section of the Table 1 (labelled 'step 1') examines the improvement in the covariance between scores on the two halves that would result from swapping any question in half 1 with any other question in half 2. For example, swapping question 1 to half 2 and question 2 the other way would increase the covariance between halves by 1.27. Note that, questions cannot swap with themselves and that questions cannot swap from a half if they are not already included in that half. This leads to the regular pattern of 0s in the matrix. Note that, the algorithm also considers swapping any question to the opposite half without moving another question in the opposite direction. This possibility is explored in the last row and last column within each step in Table 1.

All swaps that would lead to a positive change are highlighted in blue and the swap leading to the greatest improvement (item 1 swapping with item 8) is highlighted in green. Once this is done, we can then recalculate the improvement in covariance from any subsequent swaps ('step 2'). In fact, only one swap (item 5 with item 4) leads to any improvement. Once these items are swapped, there are no possible improvements from further swaps ('step 3'). This means the final split has questions 3, 4, 7, 8 and 9 in half 1 and questions 1, 2, 5, 6 and 10 in half 2.

Results

The algorithm described above was applied to each of the 30 papers detailed in Figure 1. The half scores on each paper were rescaled so that, for each candidate, the total of their scores on the two halves equalled their total UMS score[6] for each paper as a whole rather than their total raw score[7]. This means that, for each candidate, the total of their 12 half-paper scores equalled their total UMS score for the A level as a whole – the score used to determine their final grade.

Rather than simply comparing the total of the scores on all the first halves with the total of the scores on all the second halves, we applied a best split of best splits method to ensure both halves are representative of a full A level. Having applied this method, we finally have scores on two 'half A levels' each comprising of total scores across a mixture of half 1 and half 2 scores from different units so as to maximise the association.

Figure 3 compares the scores on each half A level for a random sample of 1000 candidates. As can be seen, there is a very strong relationship between the two scores with the majority of candidates displaying close agreement. Table 2 displays the mean and standard deviation of scores for the whole cohort on each half. As can be seen, the distribution of UMS scores is fairly similar on each half.

The reliability of Mathematics A level as a whole is estimated via the association between the two halves. Overall, there was a very strong correlation between halves (0.928). This can be combined with the Spearman–Brown formula to generate an overall reliability estimate of 0.963. An almost identical reliability coefficient can be generated based upon the covariance between the two halves and using the formula of Rulon (1939); that is, the usual formula for Guttman's λ_4.

5. In practice, more than one initial starting split is used in order to ensure that the optimal split is identified.
6. Uniform Mark Scale. See AQA (2009) and Gray and Shaw (2009) for details.
7. In order to achieve this, the each candidate's total UMS score was divided between the two halves according to the proportion of their total raw score that was achieved on each half.

Table 1: Example of the algorithm used to find the optimal split for Core Mathematics 1 in June 2012

			Question to swap from half 2										
			1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*None*
	Step 1	*1*	0	1.27	0	1.80	0	1.94	0	2.19	0	-1.64	-0.84
		2	0	0	0	0	0	0	0	0	0	0	0
		3	0	-1.06	0	-0.62	0	1.17	0	1.20	0	1.25	-5.36
		4	0	0	0	0	0	0	0	0	0	0	0
		5	0	-0.72	0	0.03	0	1.46	0	1.60	0	0.63	-4.56
		6	0	0	0	0	0	0	0	0	0	0	0
		7	0	-2.32	0	-1.65	0	0.80	0	0.58	0	1.38	-7.26
		8	0	0	0	0	0	0	0	0	0	0	0
		9	0	-3.90	0	-3.42	0	-0.49	0	-0.50	0	2.12	-9.82
		10	0	0	0	0	0	0	0	0	0	0	0
		None	0	1.31	0	1.89	0	1.48	0	1.85	0	-3.18	0
	Step 2	*1*	0	0	0	0	0	0	0	0	0	0	0
		2	0	0	0	0	0	0	0	0	0	0	0
		3	-0.99	-0.22	0	-0.23	0	-0.78	0	0	0	-5.38	-1.67
Question		*4*	0	0	0	0	0	0	0	0	0	0	0
to swap		*5*	-0.59	-0.13	0	0.17	0	-0.75	0	0	0	-6.25	-1.13
from		*6*	0	0	0	0	0	0	0	0	0	0	0
half 1		*7*	-1.61	-0.42	0	-0.19	0	-0.09	0	0	0	-4.18	-2.50
		8	-2.19	-0.92	0	-0.39	0	-0.25	0	0	0	-3.82	-3.03
		9	-2.69	-0.90	0	-0.87	0	-0.28	0	0	0	-2.34	-3.97
		10	0	0	0	0	0	0	0	0	0	0	0
		None	-0.33	-1.55	0	-1.41	0	-4.17	0	0	0	-13.50	0
	Step 3	*1*	0	0	0	0	0	0	0	0	0	0	0
		2	0	0	0	0	0	0	0	0	0	0	0
		3	-1.16	-0.23	0	0	-0.40	-0.68	0	0	0	-5.17	-1.96
		4	-0.76	-0.30	0	0	-0.17	-0.91	0	0	0	-6.42	-1.30
		5	0	0	0	0	0	0	0	0	0	0	0
		6	0	0	0	0	0	0	0	0	0	0	0
		7	-2.05	-0.69	0	0	-0.36	-0.26	0	0	0	-4.23	-3.06
		8	-2.32	-0.89	0	0	-0.56	-0.12	0	0	0	-3.58	-3.28
		9	-3.10	-1.14	0	0	-1.04	-0.42	0	0	0	-2.37	-4.49
		10	0	0	0	0	0	0	0	0	0	0	0
		None	-0.22	-1.26	0	0	-1.58	-3.78	0	0	0	-12.99	0

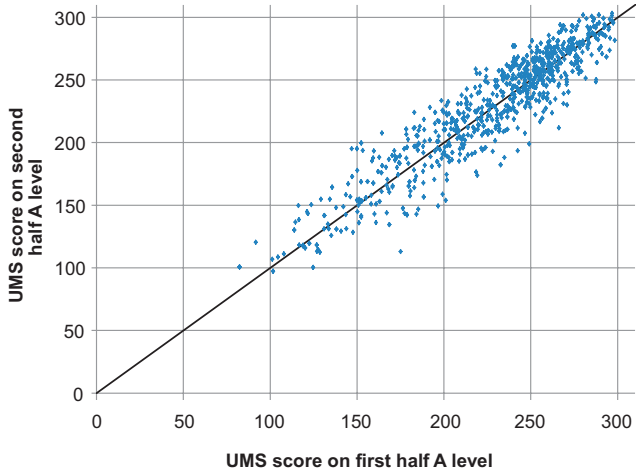


Figure 3: The relationship between scores on overall split halves of a Mathematics A level (n=1,000)

Table 2: Descriptive statistics for each "half A-level" (n=11,771)

Half	*Mean score (UMS)*	*Standard Deviation*
First	225.3	45.0
Second	226.5	45.7

Summary

This article has demonstrated how a method of split halves can be used to calculate the reliability of a complex qualification. In contrast to some approaches to this problem, based upon item response or classical test theory, the recommended approach does not start with a pre-conceived model for the way in which scores on different items will relate to one another. Rather, the underlying model is implicitly built up through the search for the most appropriate splits. This allows us to directly examine the extent to which scores, possibly representing skills across multiple domains, remain consistent across different sets of questions. This same method could be used to estimate reliability for any qualification where multiple routes are possible, including qualifications with options with regard to topic choices.

In the case of Mathematics A level, the analysis reveals an extremely high level of reliability with almost 97% of the variance in scores attributable to the underlying mathematical ability of candidates as would be demonstrated if they were able to answer an infinitely large number of questions covering the skills assessed by their chosen route through the qualification. This implies that the impact of the exact selection of questions seen by any candidate is extremely small.

Any internal estimate of reliability requires some assumptions. In particular, the split-half approach recommended in this paper assumes that the skills measured by one half are equivalent to those measured

by the other. This requires that there are no particular skills that are only assessed by a single question in any exam paper, as a single question can by definition only occur in one half. Thus, although our approach is intended to maximise the similarity between halves, we cannot be certain that the two halves of any given paper measure exactly the same set of skills. In this technical sense, the reliabilities derived via this method may be viewed as a lower bound on the true level of reliability. Having said this, as demonstrated by Benton (2013b), genuine alternative versions of the same test may also be less 'parallel' than would be desirable in a technical sense. In this way, from a practical perspective, it may be more reasonable to view the estimates as accurate, but slightly optimistic.

One limitation of the suggested method is that it only works if all units, taken within a qualification can be split into parts. This is not universally the case, for example, if one unit of the qualification comprises of a single, non-dividable mark for coursework. However, provided such elements only comprise a minority of the qualification, it will still be possible to provide a reasonably accurate estimate of reliability by adding the score from this non-dividable element to one of the two 'half A level' scores derived for the remainder of the qualification as described above. This approach would provide a reliability estimate at least as good as the classical test theory composite reliability approach suggested (amongst other approaches) by Bramley and Dhawan (2013).

As stated earlier, whilst the method suggested here estimates an overall reliability coefficient, it does not investigate whether all routes provide equally valid, or even equally reliable test scores. However, users of test scores such as employers, or university admissions, are unlikely to be aware of the route an applicant has taken through a qualification and will only see their final result. From their point of view, the ability to quantify the general level of reliability for a qualification, and in the case of Mathematics A level verify an extremely high level of reliability, may be important, even if it does not necessarily apply to every possible route. Alongside this is the ongoing duty of qualification providers to ensure that all of the individual assessments underlying the different routes through a qualification are themselves individually reliable.

References

American Educational Research Association and National Council on Measurement in Education, (1999). *Standards for Educational and Psychological Testing.* Washington, D C: AERA.

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

Benton, T. (2013a) *An empirical assessment of Guttman's Lambda 4 reliability coefficient.* Paper presented at the 78th Annual Meeting of the Psychometric Society, July 2013. Available from: http://www.cambridgeassessment.org.uk/Images/141299-an-empirical-assessment-of-guttman-s-lambda-4-reliability-coefficient.pdf.

Benton, T. (2013b). Exploring equivalent forms reliability using a key stage 2 reading test, *Research Papers in Education, 28*(1), 57–74.

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

Callender, J. and Osburn, H.G. (1977). A method for maximizing split-half reliability coefficients, *Educational and Psychological Measurement, 37*, 819–825.

Gray, E. and 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.

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

He, Q. (2009). *Estimating the Reliability of Composite Scores.* Coventry: Ofqual. Available from http://ofqual.gov.uk/documents/estimating-reliability-composite-scores/.

Rulon, P. (1939). A simplified procedure for determining the reliability of a test by split-halves. *Harvard Educational Review, 9*, 99–103.

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