Assessing comparability of optional questions

Elizabeth Harrison

Summary

Optionality is any feature in an examination that allows different students to achieve the same qualification without responding to exactly the same set of questions. Where optionality is offered within an exam paper, it is intended that the options should be equivalent in demand. Since we cannot measure demand statistically, we instead consider the relative difficulty of the different options. However, as the options may have appealed to different ability groups, it is not sufficient to simply compare the mean marks of each option; a more sophisticated analysis is needed. The Willmott-Nuttall index can be used to facilitate the comparison of the difficulty of optional questions but it does not assess whether any differences between options is consistent across the ability range.

This report illustrates possible analyses using two example papers where optionality was offered in the form of a choice between two sets of questions. The various analyses indicate that the optional sets for one of the papers were not of equal difficulty. The difference in performance of these two options was much more noticeable for low-ability students than for very high-ability students, indicating that a simple adjustment to align the options could not easily be made. The analyses also show that this is not a straightforward problem to assess statistically.

To meet regulatory requirements regarding optionality, awarding organisations need to ensure that no group of students is disadvantaged by any inconsistency in demand. It is suggested that, in future, where a qualification contains within-paper choices, specific optionality analysis could be performed. The possibility of awarding different grades for each route may need to be considered (i.e. treating each combination of options as if it were a separate optional paper). This would ensure that the awarding committee assesses work at each judgemental grade boundary for all options in order to address any inconsistency.

Introduction

Optionality is any feature in an examination that allows students to achieve a qualification without responding to exactly the same set of questions. Optionality should not unfairly disadvantage students who chose an option that was unintentionally more difficult than another option, or vice versa. When optionality is offered by using alternative papers, different grade boundaries are set for each paper, which addresses any lack of comparability. However, where optionality is offered within an exam paper, it is intended that the options should be equivalent in demand; hence, only one set of grade boundaries is used. Ideally, careful test construction and marking standardisation mean that no adjustments need to be made to enable students to be compared fairly.

Ofqual's General Conditions of Recognition require that optional questions are of equal demand and should not result in any inconsistency that might disadvantage a group of students. If this does occur, it is expected that a reasonable alteration will be made and applied uniformly to the marking (i.e. any adjustment should be fair and should not alter the rank order of the students

within an option). The current Ofqual (2016) guidelines pertinent to optionality are reproduced in the appendix.

Example data

Two example papers are used here to illustrate possible analyses. Within both papers, optionality was offered in the form of a choice between two sets of questions. A student answers either all of set A or all of set B as well as a compulsory section. The structures, assessment objectives and the breakdown of marking are identical for both sets of questions. Ideally, each pair of alternative questions should provide an equal level of demand for students. However, it is more important that the sets of optional questions provide the same overall level of demand (it is possible that differences between alternative questions might cancel each other out). Hence, the analyses presented here will compare total scores for the two options.

Possible statistics to assess comparability

In the judgement of the exam writers, the within-paper options are of equivalent demand; however, this cannot be confirmed statistically. Once the exam has been taken, the actual difficulty of the options, for the cohort who took them, can be analysed (Pollitt, Ahmed, & Crisp, 2007, p. 168).

The mean scores of two options could be compared using an unpaired t-test. If there are several options to compare, a multi-comparison correction would be needed. However, although optional questions are intended to be of equal demand, they may have appealed to different ability groups. If this is the case, it is not sufficient to simply compare the mean performances for each option; a more sophisticated analysis is needed. A fair arbiter of ability needs to be found that can enable a correction to be applied to the means. This arbiter could be the performance on a common element within the paper, an additional paper without optionality, or an external ability measure such as average KS2 score or average GCSE score. Average prior attainment scores are used to create predictions that maintain standards over time, including when there are optional components. Therefore, their use as a potential arbiter is consistent with current practice.

Various possible approaches to analysing the comparability of optional questions are described below. Each will be applied to the example data in order to assess whether any method can be usefully employed at the time of awarding to identify potential problems.

Willmott-Nuttall statistics

The Willmott-Nuttall (WN) index was designed for use with optional questions (Willmott & Hall, 1975). It can be used to estimate the average difficulty of an optional question by adjusting for the difference between the overall performance of all students and the overall performance of those selecting that option. This adjustment provides some correction for the effect produced by an optional question proving more popular with able or less able students. Any remaining differences between the WN indices of each option might then be due to the relative difficulty of each option. The adjustments made are based on averages; individual performances are not considered.

The WN index for option *j* is derived using equation 1:

$$WN_j = \bar{x}_j + (\bar{x}_w - \bar{x}_{wj}) \quad (1)$$

where:

$\bar{x}_j$ = mean score on option *j* for those students who took option *j* (expressed as a percentage of the maximum mark for the option)

$\bar{x}_w$ = mean score on the whole paper for all students (expressed as a percentage of

the maximum mark for the paper)

$\bar{x}_{wj}$ = mean score on the whole paper for those students who took option j (expressed as a percentage of the maximum mark for the paper).

The WN index is usually reported as a percentage, but interpretation of it may sometimes be easier if it were also reported in terms of marks. A mark difference is easier to relate to scores for an option or to eventual grade boundaries when assessing the impact on individuals.

The WN index is a biased estimate of the performance on an option because the correction includes the scores obtained by students for the option itself (in $\bar{x}_{wj}$). It is therefore not an independent measure of the level of attainment of students for that option (Willmott & Hall, 1975, p. 48). An alternative was proposed by Willmott and Hall, which they describe as an unbiased WN index (uWN). This index uses an adjustment that compares overall student performance with performance on the rest of the paper:

$$uWN_j = \bar{x}_j + (\bar{x}_w - \bar{x}_{rj}) \quad (2)$$

where:

$\bar{x}_{rj}$ = mean score on the rest of the paper (excluding the option *j*) for those students who took option *j* (expressed as a percentage of the maximum mark for the rest of the paper).

In some situations (though not for the example papers), the 'rest of the paper' might include another option. If, as here, a compulsory element is available, equation 2 could be modified to consider only this common performance in the adjustment. This is shown in equation 3:

$$uWN_j = \bar{x}_j + (\bar{x}_c - \bar{x}_{cj}) \quad (3)$$

where:

$\bar{x}_c$ = mean score on the compulsory element of the paper for all students (expressed as a percentage of the maximum mark for the compulsory element of the paper)

$\bar{x}_{cj}$ = mean score on the compulsory part of the paper (excluding the option *j*) for those students who took option *j* (expressed as a percentage of the maximum mark for the compulsory part of the paper).

In the case of both the example papers, the two versions of the unbiased WN will give the same value. However, for a paper with a different structure (e.g. one that requires a student to choose two options), the values will differ; in this scenario, the version of uWN in equation 3 would be preferable, if the compulsory element is of sufficient size.

The indices are an estimate of the 'true' difficulty. It is useful, where possible, to also report a range of uncertainty within which the true estimate is likely to lie. As this is not possible for the WN or uWN index, the values for these indices may give a false impression of accuracy.

Livingstone's ad-hoc index

Livingstone (1988) asked whether making adjustments based on a common element is justified if correlations to option scores are poor. He proposed a solution based on the strength of the correlations between the common and optional elements, the means and standard deviations of each element and each student's actual option score. This is a more sophisticated approach than the WN or uWN indices, but is likely to produce a more conservative estimate because the strength (or otherwise) of the correlations is also incorporated. Hence, where correlations are weak, adjustments will be small. The calculation is outlined here in relation to a paper with two alternative optional questions.

To find a student's predicted score (y_{pred}) for an alternative option, Livingstone proposes a weighted calculation using an imputed score (y_{imp}) based on the correlations and the actual score achieved on their chosen option (y_{act}). If the correlation is weak, only a slight adjustment will be made to y_{act}.

$$y_{pred} = \rho y_{imp} + (1 - \rho) y_{act}$$

where:

ρ = correlation between the performance on the common element and the performance on the alternative option question (for those students who took the alternative).

The imputed score is calculated assuming a perfect relationship between the common and optional elements. Hence, for a student taking option 1, y_{imp} is calculated so that it is in the same relative position on the compulsory element as y_{act}:

$$y_{imp} = m_{y2} + \frac{s_{y2}}{s_{x2}}(m_{x1} - m_{x2}) + \frac{s_{y2}s_{x1}}{s_{y1}s_{x2}}(y_{act} - m_{y1})$$

where:

m_{yi}, s_{yi} = mean and standard deviation on option *i* for students taking option *i* = 1, 2

m_{xi}, s_{xi} = mean and standard deviation on compulsory (common) element for students taking option *i* = 1, 2.

Finally, the adjusted score for a student taking option 1 is the average of their actual score for option 1 and their predicted score for option 2:

$$y_{adj} = (y_{pred} + y_{act})/2$$

The average adjusted score can be calculated for each candidate, and the mean adjusted scores for each option can then be compared; this enables inferences about the relative difficulty of the options. The method can be generalised when there are more than two options. As with the Willmott-Nuttall indices, there is no means of measuring the accuracy of the estimates of option difficulty.

ANCOVA

Using analysis of covariance (ANCOVA), a regression model can be fitted that links the scores achieved for the optional questions to the arbiter of student ability (e.g. marks on a compulsory question within the paper, marks on a second paper, or the average prior attainment score) (Fearnley, 2002). With this approach, both an individual student's option score and their arbiter score are considered. Three possible regression models can be fitted and compared using F-tests:

1. A null model – a single line of best fit through all the data, relating the option scores to the ability measure without differentiating between the options:

$$y_{ij} = \mu + \beta x_i$$

where:

μ = where the line crosses the y-axis

β = the slope of the fitted line

y_{ij} = mark on the optional question, of student *i* on option *j*

x_i = score on ability measure for student *i*.

2. A parallel lines model – a separate line is fitted per option, which then models the average difference in performance:

$$y_{ij} = \mu + \beta x_{ij} + \alpha_j$$

where:

μ = where a line for option A crosses the y-axis

α_j = the difference in the intercept of a separate line for option B; α_j then gives the average difference, in marks, between option A and option B.

If this model is a significantly better fit to the data than the null model (assessed using an F-test), the null hypothesis that the difference between the options is zero can be rejected. A confidence interval can be given for the α_j parameter.

3. A further non-parallel lines model can be fitted, which allows different slopes for each option (this tests for homogeneity of regression[1]):

$$y_{ij} = \mu + \beta_j x_{ij} + \alpha_j$$

where:

β_j = the slope of the line for option *j*.

If this model is a significantly better fit than the parallel lines model, the difference between the options will vary with the mark for the compulsory section. This means it would not be appropriate to apply a single transformation based on an average difference. Therefore, model 2 should not be used to determine the size of the differences between options.

For all of these models, the validity of the F-test and any confidence intervals calculated depends on normality assumptions and homogeneity of residuals. However, the F-test is fairly robust to abuse of the normality assumption when samples are large and similar in size (Lumley, Diehr, Emerson, & Chen, 2002).

When reporting estimated average differences for the parallel lines model, the following statistics can also be given:

- R^2 for the parallel and non-parallel models. R^2 lies between 0 and 1; low values would indicate that there is considerable variability in the data that has not been explained by the model, which means that inferences are likely to be unreliable
- 95% confidence intervals of the difference estimate α_j
- The significance (p value) of the F-test comparing the parallel lines model to the null model. This represents a test of whether the difference between options is significantly different to zero
- The significance (p value) of the F-test comparing the parallel lines model to the more complex non-parallel lines model (also known as a test for homogeneity of regression). This represents a test of whether the difference between options is constant.

Each of these models can be extended to include more than one arbiter (e.g. both the compulsory element score and the average prior attainment score). This method can also be generalised when there are more than two options.

[1] Homogeneity of regression: a statistical term indicating that the slopes are the same, i.e. the lines are parallel. Heterogeneity of regression exists when the slopes cannot be assumed to be the same.

Equating

The comparability of the scores for a pair of options could be viewed as an equating problem, using the ability measure as an anchor in a non-equivalent groups design (Kolen & Brennan, 2004).

Equating could be performed using the partial credit form of the Rasch model. However, Rasch modelling imposes unidimensionality on the data and this may not be a valid assumption, particularly in papers that comprise a few high-tariff questions. It is possible that the compulsory element and the optional section measure different constructs. Taylor (2009), in her analysis of English GCSE data, observed that while the first dimension of the data explained 70% of the variability, the second dimension appeared to describe the difference between the compulsory and optional elements.

Chained equipercentile equating (CEPE) could be used to assess the comparability of the options. CEPE involves two steps: the total marks on each option are equipercentile equated to the arbiter scores; option marks that are linked to the same arbiter value are then deemed to be equivalent (Kolen & Brennan, 2004, p. 14). The relationship between the options is likely to be non-linear but can be compared to an ideal relationship of one to one. Bootstrapping gives a measure of the confidence about a pair of equated marks; confidence might be low (indicated by a wide interval) when there are a few observations on a particular part of the scale, and high when there are a considerable number of observations (shown by a narrow interval). If a particular interval does not include the ideal equating, there might be grounds for concern about that region of the mark range. The differences between the ideal equating and the actual equating could be summarised as a simple average and used as an alternative to the WN indices.

Results

The results for each of the example papers are presented below, together with some interpretation.

Paper 1

Figure 1 shows a box plot that summarises the spread of marks for each section, by option choice. Table 1 gives summary statistics for each option, together with the WN, uWN and Livingstone indices. These show that option B is considerably more popular than option A. Students choosing option B tend to score higher on the optional section (by 5 marks on average), and have a similar mark on the compulsory section to those choosing option A. The marks associated with option A have a higher standard deviation.

The WN index indicates that the difference between the options is actually quite small, at -1.2 marks. However, the unbiased version suggests that the difference is nearer -4.2 marks. The Livingstone index suggests this difference is -2.6 marks; this measure has incorporated both the modest correlations and the differences in standard deviations seen in Table 1.

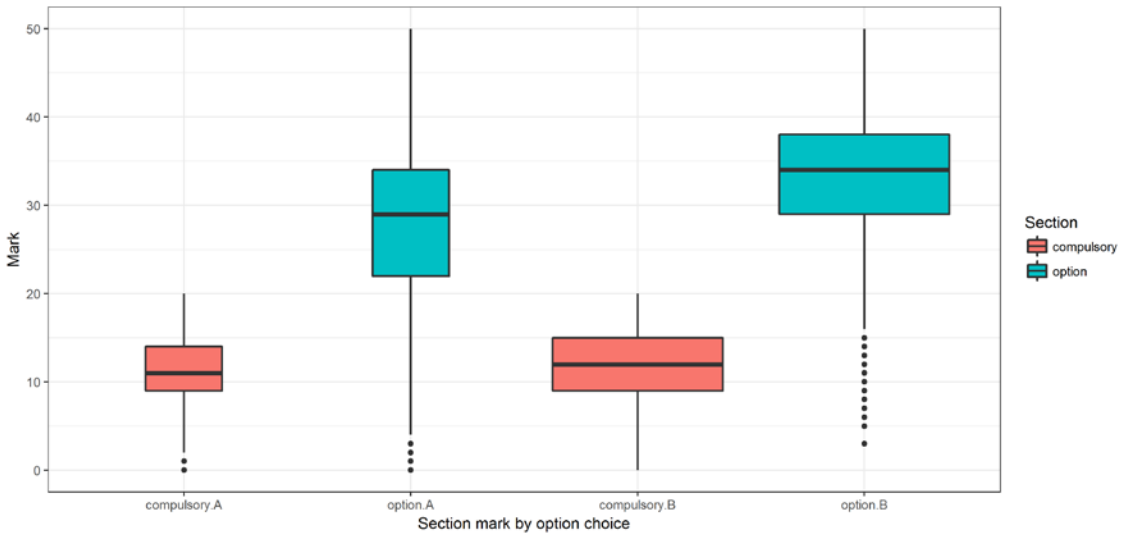

Figure 1 **Box plot for Paper 1, by option choice**[2]

Table 1 **Summary statistics, WN, uWN and Livingstone for Paper 1**

	Option A	Option B	Option A (marks)	Option B (marks)	Difference (marks)
Counts	1953	9659			
Correlations to comp	0.65	0.53			
Overall paper mean	62.85%				
Overall compulsory mean	59.21%				
Paper mean by option	56.35%	64.17%			
Comp mean (out of 20)	57.63%	59.52%	11.53	11.90	-0.37
Compulsory sd			3.85	3.70	
Option mean (out of 50)	55.83%	66.02%	27.92	33.01	-5.09
Option sd			8.64	6.82	
WN	62.34%	64.71%	31.17	32.36	-1.19
uWN	57.40%	65.71%	28.70	32.86	-4.16
Livingstone			29.07	31.63	-2.56

The ANCOVA results, using different measures of ability as covariates, are summarised in Table 2. The ability measures used are either internal (compulsory section mark) or external (average prior attainment score), or both measures are used together. Each ANCOVA indicates that the average difference between the options is -4.7 to -4.4 marks. Analysis 3, which uses both measures of ability, explains more of the variability in the data than analyses 1 or 2 (as the R^2 is higher), but the estimate of the difference between the options is similar. All the analyses indicate that the difference between the options varies across the ability range, as there is not

[2] A box plot is used to graphically summarise the spread of values seen in the data. The width of each box indicates the relative popularity of each option. The horizontal lines of the box show the lower, median and upper quartile marks within each section; 50% of the data lies within the box. The 'whiskers' show the minimum and maximum values, excluding extreme values (those beyond 1.5 × the height of the box), which are shown as dots.

homogeneity of regression (i.e. the non-parallel models are always a significantly better fit to the data).

Table 2 ANCOVA results for Paper 1

Analysis	Arbiter	Parallel lines model			Non-parallel lines model	
		R^2	Significance of option effect	Option effect (95%confidence interval)	R^2	Significance of heterogeneity
1	Compulsory	0.35	<0.001	-4.70 (-4.99, -4.41)	0.36	<0.001
2	Av. prior attain	0.36	<0.001	-4.39 (-4.71, -4.06)	0.37	<0.001
3	Av. prior & comp	0.43	<0.001	-4.38 (-4.68, -4.06)	0.44	<0.001

The non-parallel lines version of Analysis 1 is illustrated in Figure 2. It shows that the difference between the option marks decreases as student ability increases. To apply the same average correction across the mark range would disadvantage the weaker students and reward the stronger ones. The figure demonstrates that there is considerable variability in the data; this is also shown by the fairly low values of R^2 in Table 2 and the modest correlations between components given in Table 1. The confidence intervals are fairly narrow, and statistically significant effects can be found because of the large sample size.

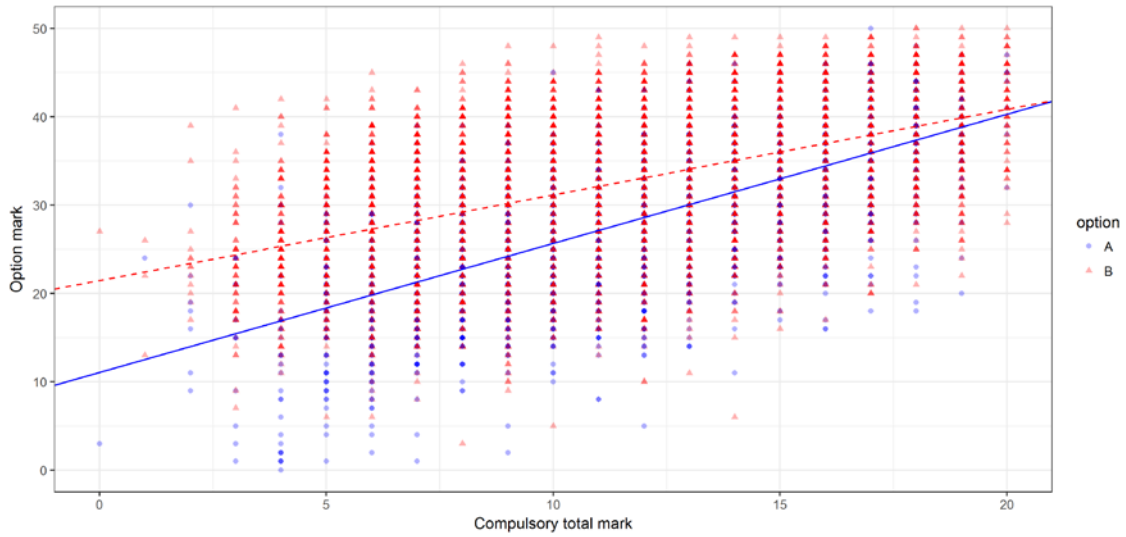

Figure 2 Non-parallel lines model fit for Paper 1

The results of CEPE, using the compulsory mark as an anchor, are shown in Figure 3 with bootstrap intervals at each possible option A mark. The bootstrap intervals are wide where there are low numbers of students. It can be seen that each option A mark equates to a higher option B mark, indicating that option A is more difficult. The ideal one-to-one equating is shown as a solid line. The simple average difference between the observed equating and the ideal equating is -4.74 marks. However, this difference is not the same across the mark range and is close to zero for high-performing students, as observed in the ANCOVA analysis. When using average prior attainment score as the anchor, CEPE gives a simple average difference of -4.73 marks.

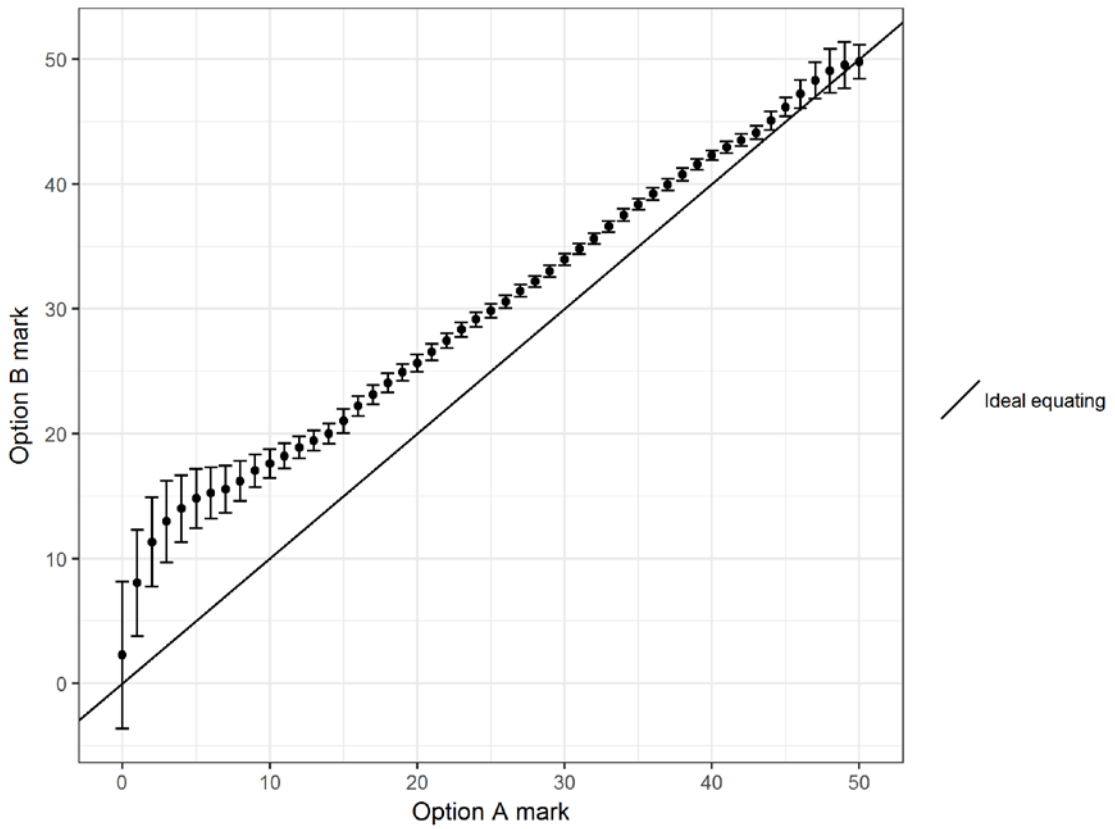

Figure 3 CEPE of optional questions for Paper 1, using an internal anchor

Paper 2

Figure 4 shows a box plot for the Paper 2 data and Table 3 shows summary statistics, together with the WN, uWN and Livingstone indices. Students choosing option A tend to score higher on the optional section (by 4.4 marks on average), and on the compulsory section (by 2.2 marks on average). The WN index indicates that there is little or no difference in the difficulty of the options (-0.4 marks). However, the unbiased version indicates that there is a difference of about -1.5 marks, which suggests that option B is in fact easier. The Livingstone index reflects the low correlations and is similar to the actual difference between the options, at 4.7 marks.

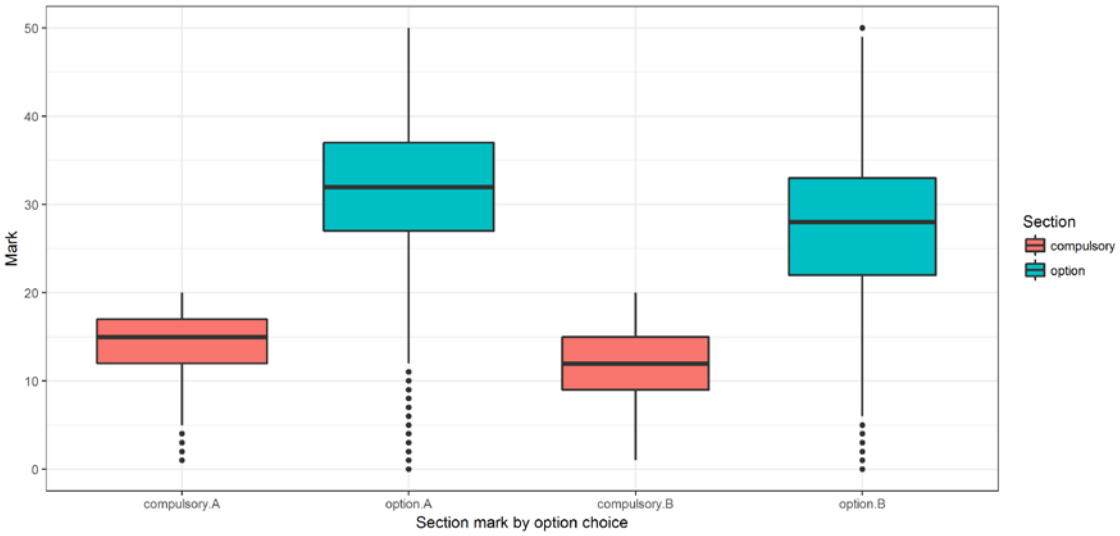

Figure 4 Box plot for Paper 2, by option choice

Table 3 Summary statistics, WN, uWN and Livingstone for Paper 2

	Option A	Option B	Option A (marks)	Option B (marks)	Difference (marks)
Counts	6651	4938			
Correlations to comp	0.58	0.56			
Overall paper mean	61.52%				
Overall compulsory mean	66.32%				
Paper mean by option	65.60%	56.03%			
Comp mean (out of 20)	71.29%	59.64%	14.26	11.93	2.33
Compulsory sd			3.39	3.81	
Option mean (out of 50)	63.32%	54.59%	31.66	27.29	4.37
Option sd			7.44	7.52	
WN	59.25%	60.08%	29.63	30.04	-0.41
uWN	58.36%	61.27%	29.18	30.64	-1.46
Livingstone			31.73	27.08	4.65

ANCOVA results are summarised in Table 4. Each ANCOVA analysis indicates that the average difference between the options is 1.2 to 2.4 marks; the three analyses do not agree on the size of any difference. The differences are in the opposite direction to those given by the WN and uWN indices. Analyses 1 and 3 indicate that the difference between the options varies across the mark range (i.e. there is heterogeneity of regression), but this is not evident when the average prior attainment score is used as the arbiter. The non-parallel lines version of Analysis 1 is illustrated in Figure 5. It shows that the difference between the options increases as student ability improves, and that to apply the same correction across the mark range would reward the weaker students and disadvantage the stronger ones.

Table 4 ANCOVA results for Paper 2

		Parallel lines model			Non-parallel lines model	
Analysis	Arbiter	R^2	Significance of option effect	Option effect (95%confidence interval)	R^2	Significance of heterogeneity
1	Compulsory	0.38	<0.001	1.57 (1.34, 1.81)	0.38	<0.001
2	Av. prior	0.36	<0.001	2.35 (2.08, 2.61)	0.36	0.16
3	Av. prior & comp	0.45	<0.001	1.24 (0.99, 1.49)	0.45	0.007

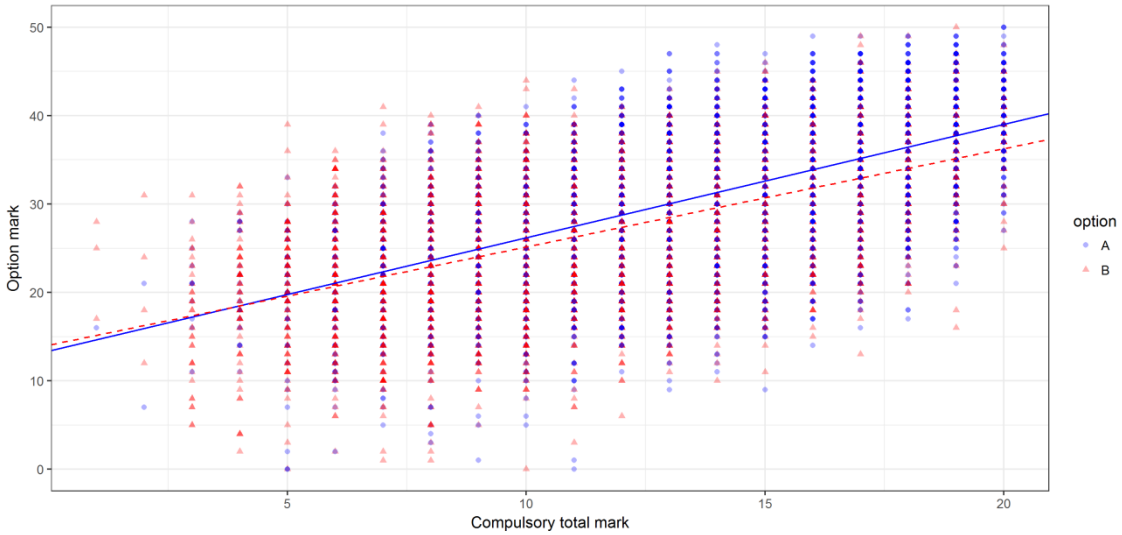

Figure 5 Non-parallel lines model fit for Paper 2

CEPE, using the compulsory mark as the anchor, gives the simple average gap as -0.79 marks. Figure 6 shows the equating with bootstrap intervals. The relationship between the two options appears to be quite complex: option A is easier than option B at low to mid marks and more difficult at high marks. Where a difference exists, it is either small or the bootstrap interval includes a difference of zero, which suggests that any differences between the options are minor. When prior attainment is used as the anchor for CEPE, a different result is produced and the gap is positive at 0.75 marks. Both CEPE results differ from those produced by ANCOVA.

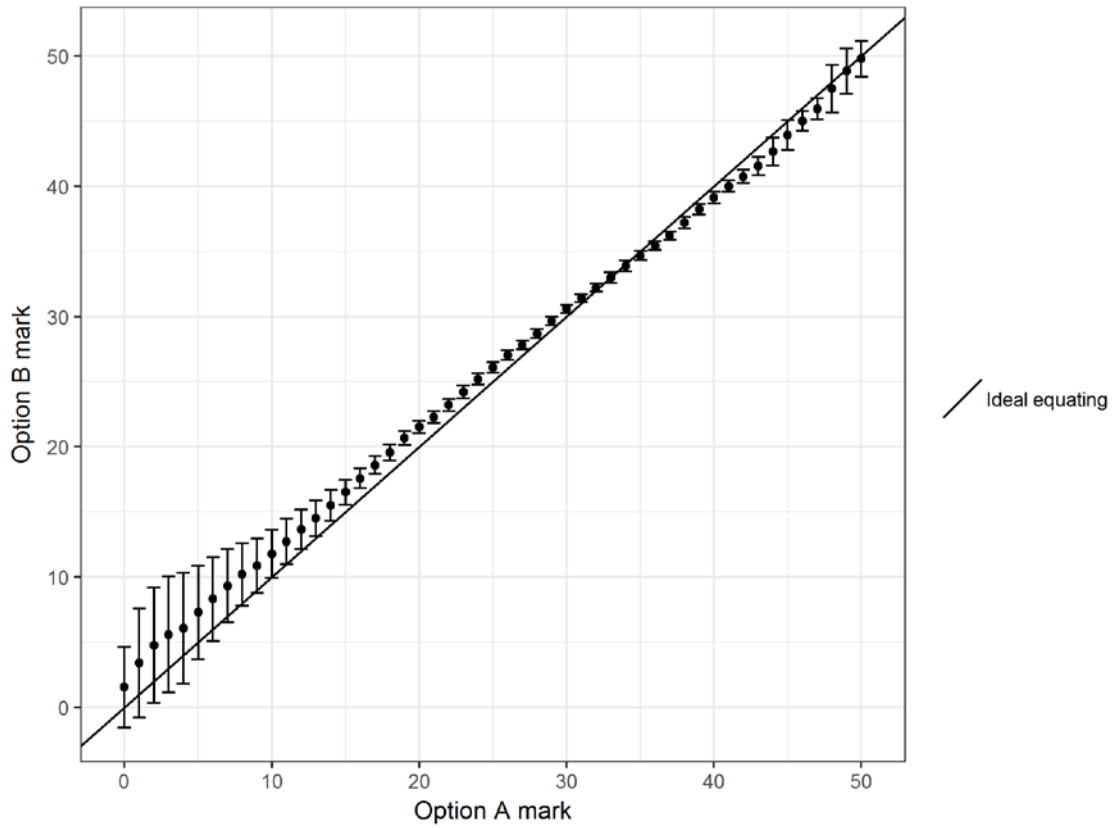

Figure 6 CEPE of optional questions for Paper 2, using an internal anchor

Discussion

This paper has explored the need for a statistic to identify any difference in the difficulty of optional questions. However, all the statistical analyses here assume that a student's choice of option is random; on many occasions it is not. Students choose the option that they think will be better for them, though they do not always do this successfully (Wang, Wainer, & Thissen, 1993). It is probable that the skill of spotting an easier option is linked to knowledge and ability in the subject. Therefore, any difference in option difficulty is likely to be confounded with student choice and will be difficult to quantify.

In the example data, the statistics suggest that for Paper 1 there is a real difference in the difficulty of the options. The WN index does not appear to have adequately captured the scale of this difference, but the other statistical methods agree that an average gap of -5 to -4 marks exists. There is considerable variability in the data but, despite this, it is still possible to show a statistically significant difference in the relationship for each option between the option mark and any arbiter of ability. It is also clear that this difference in achievement decreases as student ability improves and that to apply the same adjustment across the mark range, in order to align the options, would unfairly benefit high-achieving students.

For Paper 2, there appear to be small significant differences in difficulty between the two options, for some abilities. The complex relationship between each of the arbiters of ability and the option mark indicates that any adjustment might only be needed in part of the mark range; the various analyses do not agree on where this should be applied. Any possible adjustment does appear to be small. For example, the CEPE results (see Figure 6) indicate that the change would be 1-2 marks, where needed. However, as the various analyses do not suggest the same adjustment, it is not possible to recommend how to align the options. It may be reasonable to conclude that the inconsistencies in the analyses mean that an educationally significant difference between the options has not been demonstrated and that no adjustment is required.

No single analysis has been able to suggest an adequate solution to the problem of aligning the options:

- the WN and uWN indices give equal weight to the compulsory element and the optional element, as the calculations are performed using percentages. Here, in both papers, the compulsory element accounts for a relatively small proportion of the marks (29%) and may be inadequate to reliably measure true ability in the subject. Both indices can only give an estimate of the average difference between the difficulties of each option; heterogeneity cannot be assessed. In addition, there is no measure of the accuracy of the indices (such as a confidence interval)
- the Livingstone index is more conservative in its assessment of any difference between the options than the WN or uWN. This is because adjustments are moderated by the correlations between the optional and compulsory sections, which are often relatively weak. Again, this index can only estimate an average difference, with no indication of whether heterogeneity exists or of the error in the estimate
- the fitted ANCOVA models can estimate the size of differences in performance between options with confidence limits, indicating statistical significance. However, none of the models give a large R^2 value, which means that any inferences should be treated with some scepticism. Heterogeneity of regression can be assessed, but, where it exists, a simple average correction cannot be applied to align the options
- CEPE also shows differences between the options and provides a great deal of detail about where differences lie, but it may not be desirable or practical to apply a mark-by-mark correction to align the options.

For Paper 1, ANCOVA and CEPE largely agree on the type of differences that exist between the options, and this result is independent of the arbiter used. For Paper 2, these two methods do not agree, and the results depend on which arbiter is used (internal or external).

Implications for awarding

A possible approach to address differences such as those observed in Paper 1 would be to award grades in the same way as a subject with optional routes. That is, it would have been preferable for the paper to have had two sets of judgemental grade boundaries – one for students who took option A, and another for students who took option B. It is quite possible that the bottom grade boundary may have been set at a lower mark for option A, than for option B.

Implications for subjects using within-paper optionality

Several GCSE and GCE subjects use within-paper optionality, and many different approaches are employed. Some subjects have a compulsory element, which could be a multiple choice section (e.g. AS Economics) or a section of short objective questions (e.g. GCSE Geography). Others do not use a compulsory element at all (e.g. GCE English Literature), which means that only an external measure of prior ability could be used as an arbiter. GCE History uses both optional papers (routes) and within-paper optionality; this can result in some small entry sizes, which may make it difficult to detect any option differences. GCSE English Literature also uses optionality within a paper to reflect the different texts that have been studied – students are not expected to be able to answer the questions related to alternative texts. In this situation, optionality has been used to offer choice to schools, not to students; however, it is still important that the options are equivalent in demand. Some subjects offer more than two options (e.g. GCE Sociology), so it is then necessary to consider which pairs of options might not be comparable and make use of multiple comparison corrections.

Suggestions prior to awarding

Prior to awarding, both the option WN and the uWN indices could be considered, together with mean option scores. However, this would not give any indication of the existence of heterogeneity, or any measure of error in the estimates. Additional information such as box plots, the mean of any compulsory section (by option choice), and correlations between optional and compulsory elements would provide further insight. If the WN or uWN indices indicate a lack of comparability, further investigation could be undertaken. ANCOVA or CEPE could be used with an internal or an external arbiter of ability, before considering whether an average correction is needed or whether the subject should be treated as having optional routes requiring separate grade boundaries. Discussions with senior examiners would also be useful to ascertain if they have observed a problem with comparability during the marking, and it would be worth noting any complaints from centres. This preparation would flag the need for scrutiny of sufficient scripts representing each option and a discussion of comparability during the awarding meeting. Input from examiners is important as they may have a particular understanding of the reasons behind student choices. The decision to set separate grade boundaries would then be made based on a combination of examiner judgement and statistical evidence.

GCSE awards use grades A, C and F as judgemental grade boundaries (grades 7, 4 and 1 are used in the reformed GCSEs). The inclusion of a central grade boundary may make assessing comparability of options more straightforward for GCSE examiners than for GCE examiners, who set only two judgemental grades: A and E.

Conclusion

Option comparability is difficult to assess statistically, but it can be considered using the WN and uWN indices, and an ANCOVA can also be performed. If the ANCOVA results indicate an educationally significant problem with option comparability that is consistent across ability (i.e. there is homogeneity of regression), an adjustment could be made to the marks of one option. However, if the ANCOVA results indicate that the problem varies with ability, a compromise might be to treat the subject as having optional routes, i.e. to treat each option (plus the relevant compulsory element) as a separate paper requiring different grade boundaries.

References

Fearnley, A. J. (2002). *Comparability of awards between options within GCE units*. Manchester: AQA Centre for Education Research and Practice.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and Practices* (2nd ed.). New York: Springer.

Livingston, S. A. (1988). *Adjusting scores on examinations offering a choice of essay questions.* Education Testing Service. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev.publhealth.23.100901.140546

Lumley, T., Diehr, P. Emmerson, S. & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, 23, 151–69.

Meadows, M. (2004).*Comparability of the optional questions available in GCSE Latin, GCE Latin, and GCSE Classical Civilisation, summer 2003.* Manchester: AQA Centre for Education Research and Practice.

Pollit, A., Ahmed, A., & Crisp, V. (2007). *The demands of examination syllabuses and question papers.* QCA. Retrieved from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/487056/2007-comparability-exam-standards-g-chapter5.pdf

Ofqual. (2016, June). General Conditions of Recognition. Retrieved from https://www.gov.uk/government/publications/general-conditions-of-recognition

Taylor, R. (2009). *GCSE English A (3702): some insight gained from the Rasch model*. Manchester: AQA Centre for Education Research and Practice.

Wang, X., Wainer, H., & Thissen, D. (1993). *On the viability of some untestable assumptions in equating exams that allow examinee choice.* Education Testing Service. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/j.2333-8504.1993.tb01532.x/pdf

Willmott, A. S., & Hall, C. G. W. (1975). *O level examined: the effect of question choice*. London: Macmillan Education Ltd.

Appendix

Text from Ofqual's General Conditions of Recognition

p. 71 (*Section H: From marking to issuing results*)

Marking options

H1.2 Where –

(a) an awarding organisation offers an option as to tasks which may be completed by a Learner in an assessment or as to assessments which may be completed by the Learner (including units),

(b) the awarding organisation reasonably concludes that there is a material inconsistency between the Level of Demand of two optional tasks or assessments, and

(c) it is likely that the inconsistency will prejudice a group of Learners,

the awarding organisation must make a reasonable alteration to the criteria against which Learners' performance will be differentiated for the optional task or assessment so as to prevent that prejudice from occurring.

H1.3 Where such a reasonable alteration is made for an optional task or assessment, an awarding organisation must ensure that the alteration is applied uniformly in the marking of every task or assessment in relation to which a Learner has taken that option.