Teacher Moderation Systems

Martin Taylor

August 2005

This report was commissioned by the National Assessment Agency

TEACHER MODERATION SYSTEMS

Summary

There are two parts to this study. The first part is concerned with internally-assessed coursework. In most current GCE, VCE, GCSE and GNVQ examinations, each centre's standard of marking of coursework is monitored by inspecting a sample of candidates' work. Where a centre's marking is found to be at odds with the national standard, adjustments can normally be made to the marks based on the evidence from the sample. This procedure is known as moderation by inspection. A possible alternative method of monitoring and, where necessary, adjusting centres' marking is to compare the marks of the internally-assessed component with those of one or more externally-assessed components. This procedure is known as statistical moderation, and the study investigated how it might affect candidates' marks and grades in a number of GCE and GCSE examinations.

It was found that different methods of statistical moderation had surprisingly similar effects. In all of them there was considerable variation, across centres and candidates, in the differences between the actual marks for the internally-assessed component in the operational examination and the marks generated by statistical moderation. By including an appropriate term (called the 'allowed difference') in the statistical moderation formula, it was possible to ensure that the *mean* of these differences (across all candidates) was approximately zero, with the consequence that there was little effect on the number of candidates at each grade at subject level, but many candidates would have obtained *different* grades (some higher and some lower), particularly in specifications in which coursework accounts for a large proportion of the total assessment.

The second part of the study was concerned with teacher assessment. In recent years this term has come to mean the assessment by teachers of the general work of candidates in part or all of their course of study, rather than the type of assessment associated with current coursework components, which usually involve a particular task or set of tasks. This more general form of teacher assessment was modelled by using centres' estimated grades, which are routinely provided for GCE units. The study investigated the effect on candidates' overall (subject) grades of replacing one or more of the six unit marks awarded operationally by centres' estimates. The effect of applying statistical moderation to the estimates was also considered.

The outcomes varied but could be placed in a small number of categories. For example, in some cases replacing operational marks by estimated grades (teacher assessments) tended to cause bunching (for example, more candidates gaining C and D and fewer gaining A and U), while in others it caused inflation (for example, the numbers of candidates gaining grades A, B and C increased). To some extent, the outcomes depended on the number of units for which the operational marks were replaced by teacher assessments. Even where the numbers of candidates per grade did not change much when teacher assessment was introduced, there were often many candidates who would have obtained a different subject grade (some higher and some lower). When statistical moderation was applied, using one of the procedures from the first part of the study, it often increased the numbers of candidates who would have obtained different subject grades under teacher assessment. It also tended

to give rise to poorer subject grades, although this effect could have been reduced by using a different statistical moderation procedure or by modifying the unit grade boundaries.

TEACHER MODERATION SYSTEMS

1. INTRODUCTION

In most current GCE, VCE, GCSE and GNVQ specifications, coursework comprises one or more defined tasks (either set by the awarding body or based on criteria defined by the awarding body) and is marked by centres. Because the assessment is of an end-product, the work of a sample of candidates from each centre can be re-marked by a moderator, and adjustments to the centre's marks of all candidates at that centre can be determined, where necessary, based on a comparison of the centre marks and moderator marks for the candidates in the sample. This is known as *moderation by inspection*.

Although major changes to coursework are not now expected, the Working Group on 14-19 Reform (2004) had favoured a move away from set-piece coursework tasks to a more open-ended style of teacher assessment, possibly based on the general work of candidates during the course. One of the consequences of such a move would have been to remove the possibility of moderation in the manner described above. Thoughts therefore turned to the use of statistical methods as part of the monitoring process, to check whether centres appear to be marking to the correct standard.

The present study has two parts. The first part investigates the use of one or more externally-assessed components to moderate an internally-assessed coursework component, using statistical methods. The second part uses centres' estimated grades as a proxy for teacher assessment and investigates the effect on candidates' overall results of replacing operational marks with these 'teacher assessments', either moderated or unmoderated.

2. SCOPE OF THIS STUDY

The study is intended to investigate the effects of using statistical moderation and teacher assessment, by modelling with data which are available from current GCE and GCSE examinations, and to consider issues which arise from the modelling. It is not intended as a review of national or international experience with statistical moderation or of previous research evidence concerning different methods of statistical moderation – such aspects were recently covered in some detail by Wilmut and Tuson (2004). In undertaking this study, reference has been made to earlier surveys of methods of statistical moderation, such as Kingdon (1980) and Birch (1991), which describe a variety of methods in some detail, and also to Australian websites such as www.vcaa.vic.edu.au/vce/exams/statisticalmoderation/statmod.html It is believed that (apart from a few minor variations and refinements) this study covers most of the methods which can be used without the availability of additional information such as scores on a specially-constructed reference test. The study does not seek to investigate the accuracy of centres' estimated grades, as this matter has also been covered comprehensively in the past, for example by Delap (1995) and

Dhillon (2005), although evidence of how these estimates vary across units in one subject is presented in Appendix D.

3. PART 1: STATISTICAL MODERATION

3.1 Methods used

The following methods of statistical moderation were used in the modelling:

(i) *adjustment of centre mean marks* (ie where the internally-assessed marks for the centre are adjusted to have the same mean as the moderating instrument marks for the centre);

(ii) *linear scaling* (ie where the internally-assessed marks for the centre are adjusted to have the same mean and standard deviation as the moderating instrument marks for the centre);

(iii) *linear regression within centre* (ie where the regression line of moderating instrument marks on internally-assessed marks for the centre is used to calculate adjusted internally-assessed marks – this should not be confused with the use of regression to adjust a centre's internally-assessed marks in the context of moderation by inspection[1]);

(iv) *mapping ranks* (ie where each candidate's moderated internally-assessed mark is set to the mark on the moderating instrument which is in the appropriate position in the rank order).

The moderating instrument, which may be a single (externally-assessed) component or the aggregation of two or more components, must be scaled to have the same maximum as the internally-assessed component.

Methods (i), (ii) and (iii) are all variations on a theme in the sense that method (i) takes account of centre means only, method (ii) also adjusts for spread and method (iii) takes account of the correlation between the internal assessment marks and the moderating instrument marks at the centre (so would be expected to produce moderated internally-assessed marks which are close to the moderating instrument mean if the correlation is low). Different versions of these methods were modelled, for example using a tolerance (whereby no adjustments are made if the differences between the original centre marks and the statistically moderated marks are within a pre-defined tolerance).

A further method was considered, involving the use of linear regression, across all centres, to calculate the marks for the internally-assessed component from the marks of the other components. This method is part of the screening process currently used for Speaking & Listening in GCSE English to identify centres for possible moderator visits. However, it ignores altogether the marks awarded for the component which needs to be

[1] The use of regression to make adjustments in this context is merely automating a process which is carried out judgementally in some awarding bodies. Crucially, it still relies on the re-marking of a sample of work by a moderator and should therefore not be regarded as a form of statistical moderation.

moderated. Therefore, while it may be used as a screening device to identify centres for further investigation, it does not provide adjusted marks, and it was therefore excluded from the study.

3.2 Example

Table 3.2a Marks at a centre for an internally-assessed component and for the moderating instrument

cand	centre mark for int-ass component (max 50) *(c)*	moderating instrument (max 150)
1	42	87
2	2	20
3	36	89
4	17	43
5	37	69
6	23	70
7	16	74
8	24	78
9	38	101
10	9	58
11	12	43

	int-ass component (max 50)	moderating instrument (max 150)
mean for centre	23.27 ($\bar{c}$)	66.55
standard deviation for centre	13.39 (s_c)	23.80
mean for all centres	29.69 (μ_c)	71.56
correlation for centre	0.84 (*r*)	

In order to use the moderating instrument data, they must be converted to the same maximum as the internally-assessed data. Since the maximum for the centre-assessment and the moderating instrument are, respectively, 50 and 150, the moderating instrument data have to be multiplied by $^{50}/_{150}$, equivalent to dividing by 3. Thus, summary statistics for the scaled moderating instrument marks *w* are as follows:

mean for centre	= 66.55 ÷ 3 = 22.18	($\bar{w}$)
standard deviation for centre	= 23.80 ÷ 3 = 7.93	(s_w)
mean for all centres	= 71.56 ÷ 3 = 23.85	(μ_w)

In addition to the notation already defined, denote (for the internally-assessed component) moderated marks by *y* (centre marks are denoted by *c*).

No tolerance

First, methods (i), (ii) and (iii) are used with no tolerance applied.

Under method (i) $y = c - \bar{c} + \bar{w}$,

under method (ii) $y = (c - \bar{c}) \times s_w / s_c + \bar{w}$,

and under method (iii) $y = (c - \bar{c}) \times r \times s_w / s_c + \bar{w}$

(equivalent to $y = s_{cw} / s_c^2 \times (c - \bar{c}) + \bar{w}$) .

The moderated marks generated in this case are shown in Table 3.2b.

Table 3.2b Statistically moderated marks under methods (i), (ii) and (iii) with no tolerance applied

cand	internally-assessed component mark			
	centre mark	moderated mark		
		method (i)	method (ii)	method (iii)
1	42	41	33	32
2	2	1	10	12
3	36	35	30	29
4	17	16	18	19
5	37	36	30	29
6	23	22	22	22
7	16	15	18	19
8	24	23	23	23
9	38	37	31	30
10	9	8	14	15
11	12	11	16	17

With tolerance applied

Second, methods (i), (ii) and (iii) are used again, now applying the normal tolerance currently used operationally in the moderation of internally-assessed components (defined as 6% of the maximum component mark, rounded up, and equal to 3 in this example). The formulae for calculating the moderated marks are the same as in Table 3.2b. Then, for each centre, the tolerance is used in two ways:

(1) where the difference between $\bar{c}$ and $\bar{w}$ is within tolerance, the moderated marks are set equal to the centre marks;

(2) where all of the calculated moderated marks are within tolerance of the centre marks, these calculated marks are ignored and the moderated marks are set equal to the centre marks.

The moderated marks generated in this case are shown in Table 3.2c. Note that, in all cases, the moderated marks revert to centre marks because the difference between $\bar{c}$ and $\overline{w}$ is within tolerance.

Table 3.2c Statistically moderated marks under methods (i), (ii) and (iii), with tolerance applied

cand	internally-assessed component mark			
	centre mark	moderated mark		
		method (i)	method (ii)	method (iii)
1	42	42	42	42
2	2	2	2	2
3	36	36	36	36
4	17	17	17	17
5	37	37	37	37
6	23	23	23	23
7	16	16	16	16
8	24	24	24	24
9	38	38	38	38
10	9	9	9	9
11	12	12	12	12

With tolerance and 'allowed difference' applied

Third, methods (i), (ii) and (iii) are used again, now incorporating an 'allowed difference' between the centre means ($\bar{c}$ and $\overline{w}$) equal to the difference between the corresponding means for all centres (μ_c and μ_w). For example, if the difference between the all-centre means is 10 marks, the *expectation* is that the difference between the means for individual centres should be 10 marks, with the potential for adjustments to be made to the internally-assessed marks if not. A tolerance is again applied in two ways, as in Table 3.2c:

(1) where ($\bar{c}$ - $\overline{w}$) is within tolerance of (μ_c - μ_w), the moderated marks are set equal to the centre marks;

(2) where all of the calculated moderated marks are within tolerance of the centre marks, these calculated marks are ignored and the moderated marks are set equal to the centre marks.

The moderated marks generated in this case are shown in Table 3.2d. In fact, the tolerance has no effect in this instance (because neither (1) nor (2) is satisfied). Note that the moderated marks are always 6 marks higher (or sometimes 5 marks higher, due to the effects of rounding) than the corresponding marks in Table 3.2b because the 'allowed difference' (μ_c - μ_w) is 5.84.

Table 3.2d Statistically moderated marks under methods (i), (ii) and (iii), with an 'allowed difference' between centre means and tolerance applied

cand	internally-assessed component mark			
	centre mark	moderated mark		
		method (i)	method (ii)	method (iii)
1	42	47	39	37
2	2	7	15	17
3	36	41	36	34
4	17	22	24	25
5	37	42	36	35
6	23	28	28	28
7	16	21	24	24
8	24	29	28	28
9	38	43	37	35
10	9	14	20	21
11	12	17	21	22

Method (iv) – mapping ranks

Finally, method (iv) is used. This method is quite different from methods (i), (ii) and (iii). In Table 3.2e each candidate's moderated mark is set equal to the moderating instrument mark which has the same rank as the centre mark for that candidate. For example, candidate 3 is in position 4 in the rank order for centre marks, so his/her moderated mark is the mark which is in position 4 in the moderating instrument marks, ie 26. (The moderating instrument marks in Table 3.2e have been scaled from the corresponding marks in Table 3.2a by dividing by 3, as described earlier.) Tied ranks in the centre marks (not present in this example) are easy to deal with – for example, where ranks 6 and 7 are tied the candidates receive the mean of the moderating instrument marks in positions 6 and 7. It is less obvious how tied ranks in the moderating instrument marks should be treated. In Table 3.1.2e candidates 10 and 11, who are in positions 10 and 9 respectively in the centre marks, both receive the mark which is at the tied 9-10 position in the moderating instrument rank order. There are other, possibly better, ways of dealing with tied ranks in the moderating instrument marks, but it would be futile in this project to pay undue attention to this issue, which has no more than a marginal effect on the outcomes.

For comparability with methods (i), (ii) and (iii), the same three approaches are used – first with no tolerance and no 'allowed difference', second with the tolerance applied as defined in (1) and (2) of Table 3.2c and third with tolerance and an 'allowed difference' applied as in Table 3.2d. Referring to centre mean marks seems somewhat contrived when mapping ranks – perhaps using medians instead of means would have been more appropriate, but again there would have been a marginal effect on the outcomes.

Table 3.2e shows the moderated marks generated using mapping ranks with the three approaches. Note that, when the tolerance is applied, the moderated marks revert to centre marks (as in Table 3.2c) and that the incorporation of an allowed difference increases the moderated marks by 6 (as in Table 3.2d).

Table 3.2e Statistically moderated marks under method (iv)

cand	centre mark for internally-assessed component	rank for internally-assessed component	mod inst mark (scaled)	rank for mod inst mark	moderated mark for int-ass component with no tolerance	moderated mark for int-ass component with tolerance applied	moderated mark for int-ass component with allowed difference and tolerance
1	42	1	29	3	34	42	40
2	2	11	7	11	7	2	13
3	36	4	30	2	26	36	32
4	17	7	14	9.5	23	17	29
5	37	3	23	6.5	29	37	35
6	23	6	23	6.5	23	23	29
7	16	8	25	5	19	16	25
8	24	5	26	4	25	24	31
9	38	2	34	1	30	38	36
10	9	10	19	8	14	9	20
11	12	9	14	9.5	14	12	20

3.3 Specifications used in the analysis of statistical moderation

The AQA specifications included in this part of the study are listed below. Mark data from Summer 2004 were used. The approximate proportions of centres which, operationally, had their marks for the internally-assessed component adjusted in Summer 2004 are shown in brackets.

GCSE Business Studies Specification A (17%)

GCSE Design & Technology (Food) (12%)

GCSE History Specification A (17%)

GCSE History Specification B (16%)

GCSE Humanities (16%)

GCSE Music (two components – 42% and 27%)

GCSE Religious Studies Specification A (15%)

GCE Biology Specification A (14%)

GCE Business Studies (38%)

GCE French (38%)

GCE Geography Specification A (see below)

GCE Geography Specification B (4%)

GCE Psychology Specification A (27%)

In the case of GCE Biology, only the AS units were included in the study.

Each of the GCE Geography specifications has a coursework component but in Specification A the coursework is externally-assessed (therefore there is no figure for the proportion of centres adjusted). Although the coursework marks should not therefore need to be moderated, this specification was included in the study in order to compare the effects of statistical moderation of the coursework component in the two specifications.

4. PART 2: TEACHER ASSESSMENT

4.1 Background

In GCE, centres currently provide estimated grades (A-E or U) for every unit. These estimates, though on a rather coarse scale, can be used as a proxy for teacher assessment, thus giving the opportunity to model the effects of replacing operational marks by teacher assessments.

In the modelling, candidates' uniform marks for selected units were replaced by the centres' estimated grades ('teacher assessments'). These estimated grades were converted to a numerical scale proportional to the uniform mark scale, with a nominal maximum of 5 (see Table 4.1a). Note that the numerical values are placed at approximately the mid-point of the grade range (except for U, which is placed at zero). For example, grade B is converted to 3.75, which as a percentage of the nominal maximum (5) is 75%.

Table 4.1a Conversion of estimated grades to a numerical scale

grade	uniform mark range (as percentage of the maximum uniform mark)	numerical equivalent for grade
A	80-100	4.5
B	70-79	3.75
C	60-69	3.25
D	50-59	2.75
E	40-49	2.25
U	0-39	0

Following suitable scaling, these teacher assessments were aggregated with the uniform marks from the remaining units and the resulting grades were compared with the grades issued operationally. In each case, overall grades were generated using both the raw teacher assessments and the teacher assessments moderated using method (i) from Part 1 of the study

(adjustment of the teacher assessments to have the same centre mean as the moderating instrument – see section 3.1).

The use of teacher assessments for various numbers of units was trialled. Where the uniform marks for just one unit were replaced by teacher assessments, there was likely to be a relatively small effect on candidates' overall grades, but the scale for the teacher assessments was very coarse (with only six points, corresponding to U, E, …, A). Where (at the other extreme) the uniform marks for five units were replaced by teacher assessments, the effects on candidates' overall grades were likely to be greater, but in this case there was a twenty-six point scale for the aggregated teacher assessments (with a nominal maximum of 25).

Two procedures for carrying out statistical moderation were trialled:

(i) the teacher assessments for the other component(s) were compared with the actual uniform marks for the moderating instrument;

(ii) the teacher assessments for the moderating instrument were compared with the actual uniform marks for the moderating instrument.

Although approach (ii) appears more logical, it could in practice lead to abuse. Because the teacher assessments for the moderating instrument would not directly affect candidates' results (they would be used only for determining adjustments), centres could manipulate the system by giving unduly low assessments for this element, causing upward adjustments to be made to the teacher assessments for the other component(s) (ie the components for which the teacher assessments mattered). Approach (i) does not lend itself to abuse or manipulation but has the disadvantage that the teacher assessments and external assessments which are being compared are for different components. However, this is the procedure which was used in Part 1 of the study (see section 3.1).

For this part of the study, it is largely irrelevant whether a unit is internally-assessed or externally-assessed.

4.2 Example

In this example, estimated grades are used as a proxy for teacher assessment for units 1, 2 and 4 (combined) in a six-unit GCE. The teacher assessments for these units are aggregated with the uniform marks for units 3. 5 and 6 to produce overall A level grades. This aggregation is carried out twice, for both the unmoderated and the moderated teacher assessments.

For statistical moderation of the teacher assessments, units 3, 5 and 6 (combined) are used as the moderating instrument. Under approach (i) in section 4.1 above, the teacher assessments for units 1, 2 and 4 are statistically moderated by making an adjustment based on a comparison of the teacher assessment *for these units* and the actual (total) uniform mark for

units 3, 5 and 6. Under approach (ii), estimated grades are also used as a proxy for teacher assessment for units 3, 5 and 6. The teacher assessments for units 1, 2 and 4 are statistically moderated by making an adjustment based on a comparison of the teacher assessment for units 3, 5 and 6 and the actual (total) uniform mark for these units.

Table 4.2a shows the initial details for one centre of five candidates. Further details of the outcomes of statistical moderation and aggregation follow in later tables.

Table 4.2a Teacher assessments at a centre (using estimated grades as a proxy) and uniform marks for the moderating instrument (units 3, 5 and 6 combined).

cand	tch ass for units 1+2+4 (max 15)	tch ass for units 3+5+6 (max 15)	total um for units 3+5+6 (max 270)
1	10.25	10.75	141
2	10.25	10.75	157
3	12.00	12.00	122
4	9.25	9.75	161
5	11.00	10.75	180
mean for centre	10.55	10.80	152.20

In order to use the moderating instrument data, they must be converted to the same maximum as the teacher assessments. Since the maxima for the teacher assessment and the moderating instrument are, respectively, 15 and 270, the moderating instrument data have to be multiplied by $^{15}/_{270}$ (equivalent to dividing by 18). Thus

scaled mod inst mean for centre = 152.20 ÷ 18 = 8.46.

In Table 4.2b statistical moderation is carried out by comparing the actual uniform marks for the moderating instrument (units 3+5+6) with the teacher assessments for the other components (units 1+2+4). The difference in means is

(8.46 – 10.55) = -2.09,

so each candidate's teacher assessment (for units 1+2+4) is reduced by 2.09. In the fourth and fifth columns the A level grade is generated by replacing the uniform marks for units 1+2+4 with the teacher assessments for these units (appropriately scaled).

Table 4.2b **Effect on subject grades of replacing the uniform marks for units 1+2+4 with teacher assessments (estimated grades), moderated by comparing the centre's mean teacher assessment for units 1+2+4 with the centre's mean external mark for units 3+5+6 (the moderating instrument)**

cand	unmoderated tch ass for units 1+2+4	moderated tch ass for units 1+2+4	overall A level grade		
			with unmod tch ass	with mod tch ass	actual
1	10.25	8.16	C	D	C
2	10.25	8.16	C	D	C
3	12.00	9.91	C	D	D
4	9.25	7.16	C	D	D
5	11.00	8.91	B	C	C

In Table 4.2c statistical moderation is carried out by comparing, for the moderating instrument (units 3+5+6), the teacher assessments with the actual uniform marks. The difference in means is

(8.46 – 10.80) = -2.34,

so each candidate's teacher assessment (for units 1+2+4) is reduced by 2.34. In the fourth and fifth columns the A level grade is generated by replacing the uniform marks for units 1+2+4 with the teacher assessments for these units (appropriately scaled).

Table 4.2c **Effect on subject grades of replacing the uniform marks for units 1+2+4 with teacher assessments (estimated grades), moderated by comparing the centre's mean teacher assessment for units 3+5+6 with the centre's mean external mark for units 3+5+6 (the moderating instrument)**

cand	unmoderated tch ass for units 1+2+4	moderated tch ass for units 1+2+4	overall A level grade		
			with unmod tch ass	with mod tch ass	actual
1	10.25	7.91	C	D	C
2	10.25	7.91	C	D	C
3	12.00	9.66	C	D	D
4	9.25	6.91	C	D	D
5	11.00	8.66	B	C	C

4.3 Specifications used in the analysis of teacher assessment

The AQA specifications included in this part of the study (all GCE) are listed below. Mark data from Summer 2004 were used.

Biology Specification A

Business Studies

Communication Studies

Computing

English Literature Specification A

English Literature Specification B

French

Geography Specification A

Geography Specification B

ICT

Law

Psychology Specification A

As in Part 1 of the study, only the AS part of Biology was included.

5. STATISTICAL MODERATION: RESULTS

A huge volume of data has been collected and analysed. It would be impracticable to present the findings exhaustively. Instead, a certain amount of detail is provided for one of the specifications included in the study (GCSE History Specification A) and the main patterns are identified for the other specifications.

There were two stages in the analysis. First, the statistically moderated marks were calculated for the centre-assessed component (as explained in sections 3.1 and 3.2) as well as the differences between those marks and the operational marks. Second, the subject grades obtained when using statistical moderation for the centre-assessed component were compared with the operational subject grades. Any differences between statistically moderated marks and operational marks would be of little consequence if there was negligible effect on candidates' overall grades.

5.1 GCSE History A

This specification has two written components (each with 37½% weighting) and a centre-assessed coursework component (with 25% weighting). It is untiered. One of the written components has four options; only the option with the largest number of candidates was considered. For the purpose of the study, each candidate's marks for the two written papers were added (using the appropriate scaling factors for the specification), and the aggregated marks for these components were used as the moderating instrument for the coursework component.

The correlation between marks for the aggregated written papers and marks for the coursework component was found to be 0.72. The maximum mark for the coursework component is 50.

5.1.1 Differences between statistically moderated marks and operational marks

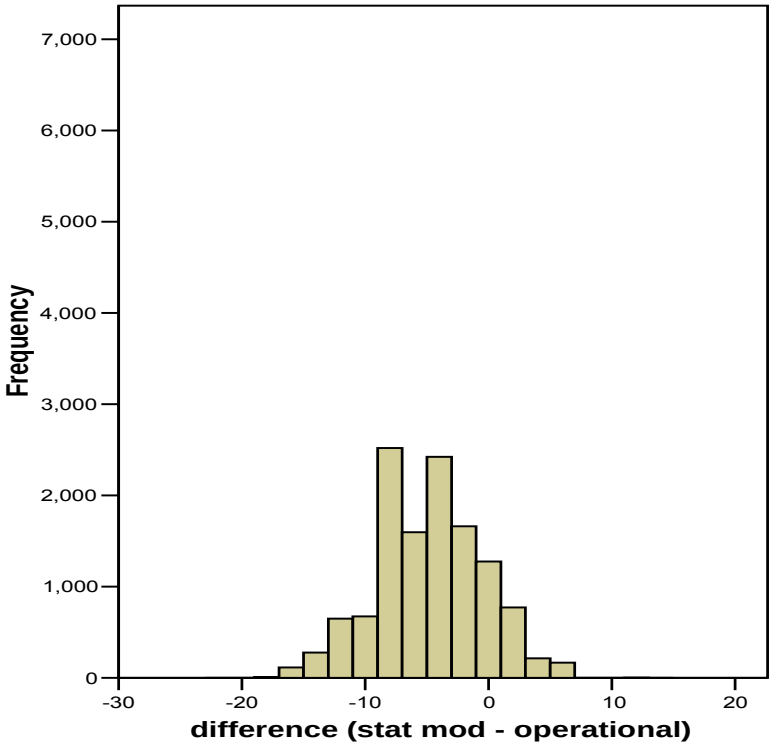
Table 5.1.1a shows summary statistics for the differences between statistically moderated marks and operational marks for the coursework component when method (i) (adjustment of centre mean marks) is used with no tolerance applied (see section 3.1). Please note that negative differences indicate that the statistically moderated marks were lower than the operational marks (ie the statistical moderation was more severe than the moderation by inspection which was used operationally).

Table 5.1.1a GCSE History Specification A: differences between statistically moderated marks and operational marks for the coursework component under method (i) with no tolerance applied

Number of candidates	12355
Maximum mark for coursework	50
Weighting (%)	25%
Mean difference	-5.3
Mode	-8
Standard deviation	4.3

Figure 5.1.1a shows the differences diagrammatically. A scale of 0-7000 on the vertical axis is used in order to assist comparison with Figures 5.1.1b and 5.1.1c.

Figure 5.1.1a GCSE History Specification A: differences between statistically moderated marks and operational marks for the coursework component under method (i) with no tolerance applied



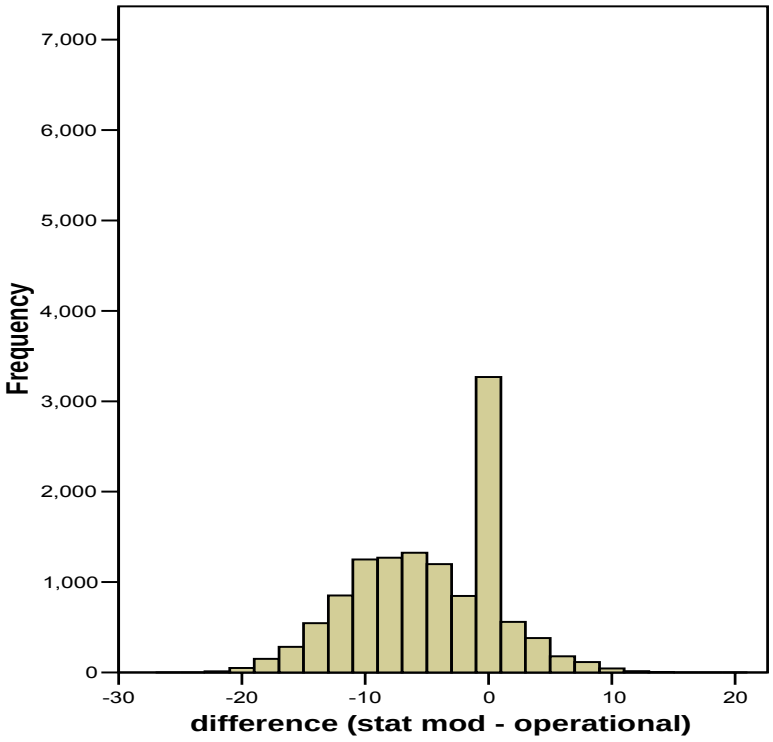
The outcomes under methods (ii), (iii) and (iv) (linear scaling, linear regression within centre and mapping ranks, respectively) were similar. In line with a general pattern observed for most of the specifications considered, the spread of the differences was largest for method (iii) followed by (iv) and then (ii), with (i) the smallest. However, the means and modes were all about the same (see Table 5.1.1b).

Table 5.1.1b GCSE History Specification A: comparison of statistics for methods (i)-(iv) with no tolerance applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	12355			
Maximum mark for coursework	50			
Weighting (%)	25%			
Mean difference	-5.3	-5.4	-5.4	-5.3
Mode	-8	-8	-5	-7
Standard deviation	4.3	5.1	6.0	5.4

A second approach was to apply the normal moderation tolerance (3 marks for the present coursework component) to the statistically moderated marks. Again, the details are explained in section 3.1. Although the means and standard deviations were similar to those in Table 5.1.1a, the modes transfer to zero because of the application of the tolerance. Figure 5.1.1b illustrates for method (iii). The other methods gave similar outcomes, although method (i) in particular had a lower frequency at the mode.

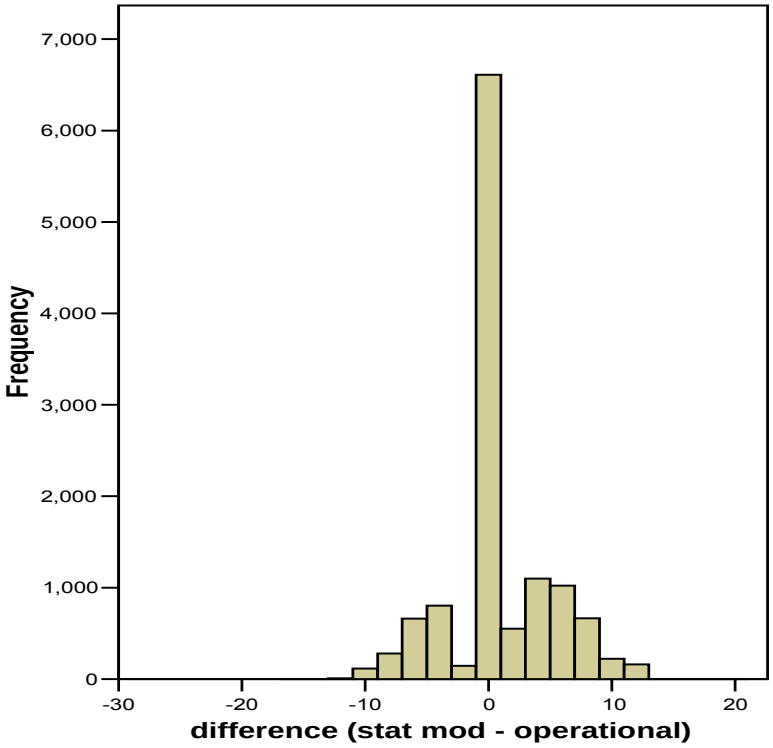
Figure 5.1.1b GCSE History Specification A: differences between statistically moderated marks and operational marks for the coursework component under method (iii) with tolerance applied



Most of the 3000 or so candidates with a zero difference are those who would have received the mark awarded by the centre, both operationally and under statistical moderation.

When a so-called allowed difference is introduced, systematic variations between coursework mean marks and written paper mean marks are eliminated. The mean differences under all of methods (i)-(iv) are close to zero and the modes are zero. The standard deviations are a little smaller than before. Figure 5.1.1c shows the differences under method (i).

Figure 5.1.1c GCSE History Specification A: differences between statistically moderated marks and operational marks for the coursework component under method (i) with an 'allowed difference' and tolerance applied



5.1.2 Effect on subject grades

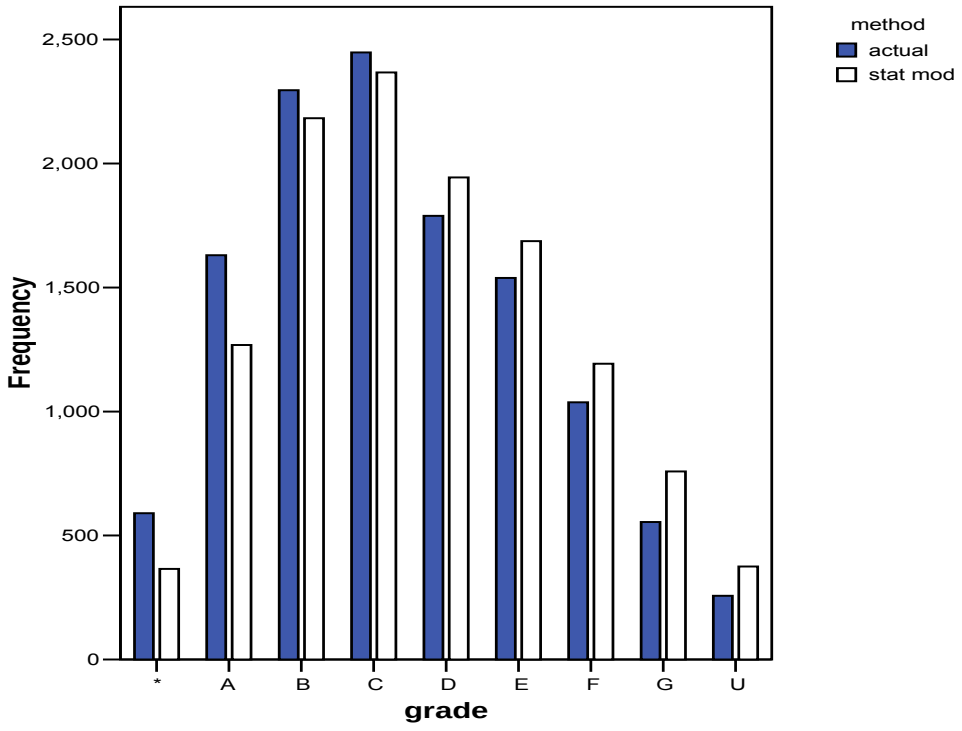
Table 5.1.2a and Figure 5.1.2a illustrate how subject grades were affected when the operational marks for the coursework component were replaced by statistically moderated marks. The data are for method (i) (adjustment of centre mean marks) with no tolerance or allowed difference applied.

Table 5.1.2a GCSE History Specification A: crosstabulation showing effect on subject grades of replacing operational marks for the coursework component with statistically moderated marks (method (i), no tolerance or allowed difference)

(Table shows numbers of candidates)

		Grade with statistical moderation for the coursework component									Total
		A*	A	B	C	D	E	F	G	U	
Actual grade	A*	357	233	0	0	0	0	0	0	0	590
	A	8	1019	603	0	0	0	0	0	0	1630
	B	0	16	1561	719	0	0	0	0	0	2296
	C	0	0	19	1619	807	3	0	0	0	2448
	D	0	0	0	29	1113	644	3	0	0	1789
	E	0	0	0	0	24	1021	493	1	0	1539
	F	0	0	0	0	0	19	686	332	0	1037
	G	0	0	0	0	0	0	11	417	126	554
	U	0	0	0	0	0	0	0	8	249	257
Total		365	1268	2183	2367	1944	1687	1193	758	375	12140[2]

Figure 5.1.2a GCSE History Specification A: effect on subject grades of replacing operational marks for the coursework component with statistically moderated marks (method (i), no tolerance or allowed difference)



[2] The shortfall compared with Table 5.1.1a is because of candidates who did not have valid marks for both components.

The pattern shown here is repeated in many of the specifications considered, with fewer candidates obtaining the higher grades and more candidates obtaining the lower grades under statistical moderation of the coursework component. This pattern is clearly to be expected, because the data in section 5.1.1 above showed that statistical moderation generally gives rise to lower marks for the coursework component than those obtained operationally.

The crosstabulation shows that, while many candidates would lose a grade, a few candidates would gain, and very few would change by more than one grade. This pattern was not the same across all of the specifications considered.

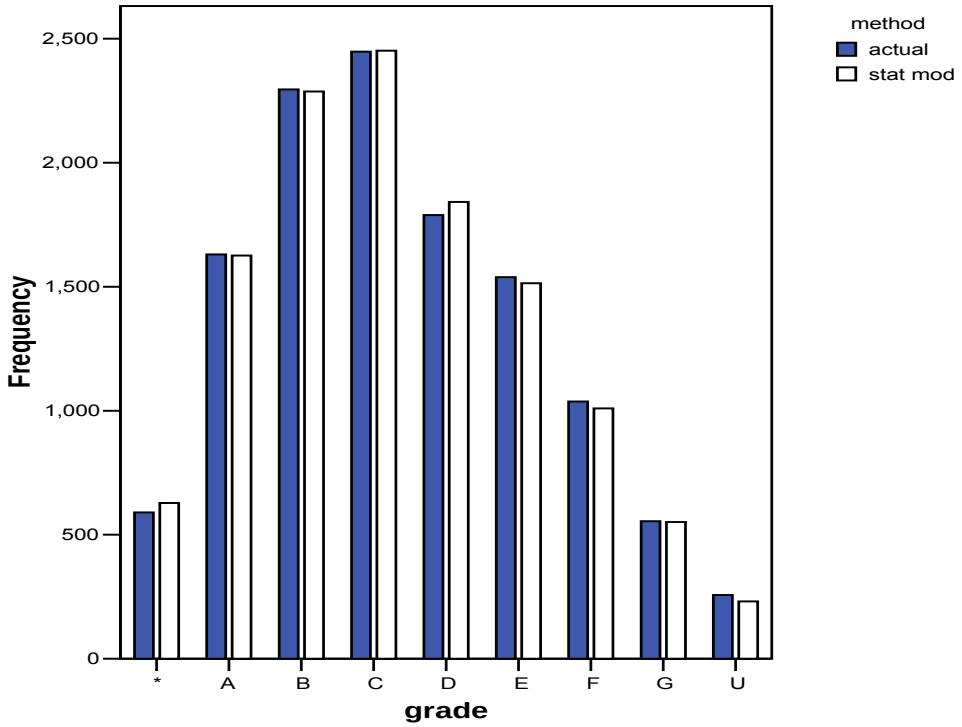
As illustrated in section 5.1.1 above, the statistically moderated marks are the same as the operational marks for many candidates when an 'allowed difference' is applied. Therefore, far fewer subject grade changes would be expected. However, the crosstabulation shows that more candidates would change grades than might be expected from inspecting the diagram (see Table 5.1.2b and Figure 5.1.2b).

Table 5.1.2b GCSE History Specification A: crosstabulation showing effect on subject grades of replacing operational marks for the coursework component with statistically moderated marks (method (i), with tolerance and allowed difference applied)

(Table shows numbers of candidates)

		Grade with statistical moderation for the coursework component									
		A*	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**	**Total**
Actual grade	**A***	548	42	0	0	0	0	0	0	0	590
	A	80	1450	100	0	0	0	0	0	0	1630
	B	0	134	2018	144	0	0	0	0	0	2296
	C	0	0	169	2126	153	0	0	0	0	2448
	D	0	0	0	182	1489	118	0	0	0	1789
	E	0	0	0	0	200	1250	89	0	0	1539
	F	0	0	0	0	0	146	829	62	0	1037
	G	0	0	0	0	0	0	91	443	20	554
	U	0	0	0	0	0	0	0	46	211	257
Total		628	1626	2287	2452	1842	1514	1009	551	231	12140

Figure 5.1.2b GCSE History Specification A: effect on subject grades of replacing operational marks for the coursework component with statistically moderated marks (method (i), with tolerance and allowed difference applied)



5.2 Other specifications with similar properties

Several other untiered GCSE specifications with similar weightings for the centre-assessed coursework component were investigated. These were History Specification B, Humanities and Religious Studies Specification A. As can be seen from Table 3.4, all had almost identical proportions of centres whose coursework marks were adjusted in 2004 (about 16%). The outcomes were similar to those for GCSE History A.[3] For the record, summary statistics are shown in Appendix A.

5.3 Tiered GCSE specifications

Two tiered GCSE specifications were considered: Business Studies Specification A and Design & Technology (Food). Each of these has a tiered written paper and an untiered coursework component. In Business Studies the coursework accounts for 25% of the total weighting while in Design & Technology it accounts for 60%.

[3] Where there is just one written component, this is used as the moderating instrument for statistical moderation. Where there is more than one, these components are aggregated to form the moderating instrument (as in GCSE History Specification A).

The nature of the assessment pattern in tiered specifications was expected to inflate the coursework marks for Foundation tier candidates and deflate them for Higher tier candidates. The reason for the expectation for the Foundation tier is explained in the next paragraph, for the simplest method of statistical moderation involving adjustment of centre mean marks with no tolerance or allowed difference applied.

Suppose that the maximum mark for both the (Foundation tier) written paper and the (untiered) coursework component is 50. The mean mark for the written paper might be expected to be about 25 and the mean coursework mark for the same (Foundation tier) candidates about 15. The lower coursework mean would be expected because this component covers the whole grade range while the written paper covers only grades C-G. With these overall means, which would generally be reflected in centre mean marks, statistical moderation would cause the coursework marks to be increased by an average of 10.

In fact, the expectations were not realised. Further investigation revealed that, while for the Higher tier the mean written paper marks were indeed lower than the mean coursework marks, the same was also true for the Foundation tier. The outcomes from statistical moderation were therefore similar to those in the untiered specifications (where the written paper means were also lower than the coursework means), although for Higher tier candidates the downward adjustments were greater.

Tables 5.3a and 5.3b show summary statistics for Business Studies Foundation tier and Business Studies Higher tier when no tolerance or allowed difference is applied. Figures 5.3a and 5.3b show the outcomes diagrammatically for method (i).

Table 5.3a GCSE Business Studies Specification A Foundation tier: summary statistics for methods (i)-(iv) with no tolerance or allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	8926			
Maximum mark for coursework	63			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	-3.9	-3.9	-3.9	-3.9
Mode	-4	-2	-2	-2
Standard deviation	5.8	7.0	9.3	7.4

Table 5.3b **GCSE Business Studies Specification A Higher tier: summary statistics for methods (i)-(iv) with no tolerance or allowed difference applied**

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	10000			
Maximum mark for coursework	63			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	-12.4	-12.4	-12.4	-12.3
Mode	-13	-13	-14	-16
Standard deviation	6.4	7.4	9.2	7.8

Figure 5.3a **GCSE Business Studies Specification A Foundation tier: differences between statistically moderated marks and operational marks for the coursework component under method (i) with no tolerance or allowed difference applied**

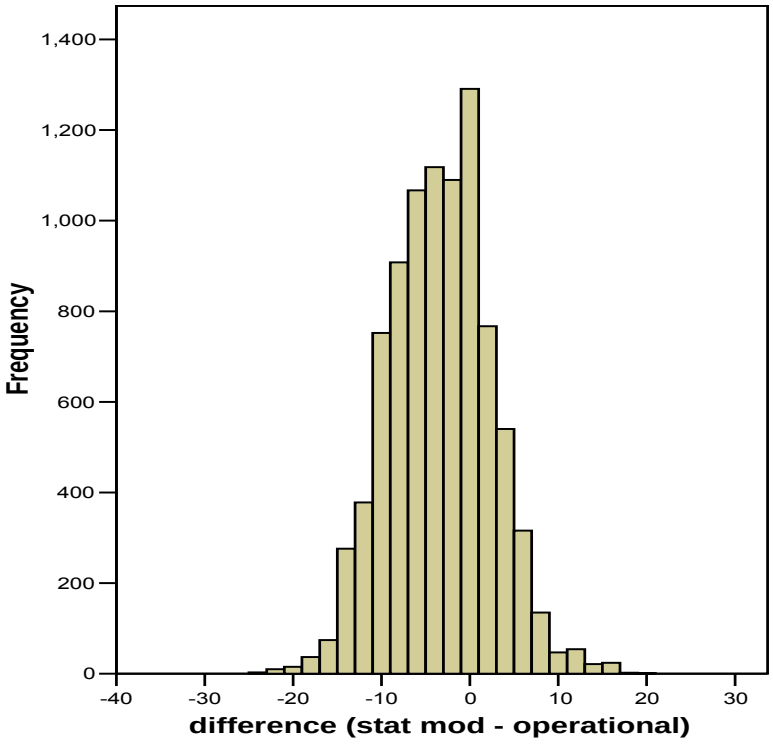
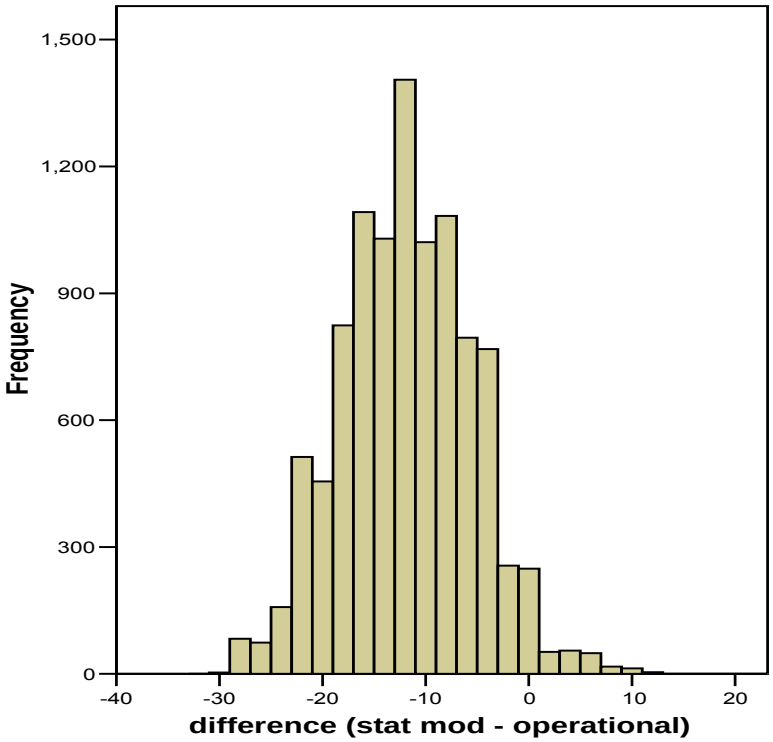


Figure 5.3b GCSE Business Studies Specification A Higher tier: differences between statistically moderated marks and operational marks for the coursework component under method (i) with no tolerance or allowed difference applied



A further issue with respect to Design & Technology is the high weighting for coursework – it accounts for 60% of the total assessment in contrast to 20%-30% in most of the other specifications investigated. The consequence is that changes brought about by statistical moderation have a much greater effect on subject grades. Figures 5.3c and 5.3d compare the outcomes for Business Studies Higher tier and Design & Technology Higher tier, for which the effects of statistical moderation were fairly similar in terms of mean difference (in fact slightly greater for Business Studies). Figure 5.3d shows the draconian effect on candidates' grades in Design & Technology.

Figure 5.3c **GCSE Business Studies Higher tier: effect on subject grades of replacing operational marks for the coursework component with statistically moderated marks (method (i), with no tolerance or allowed difference applied)**

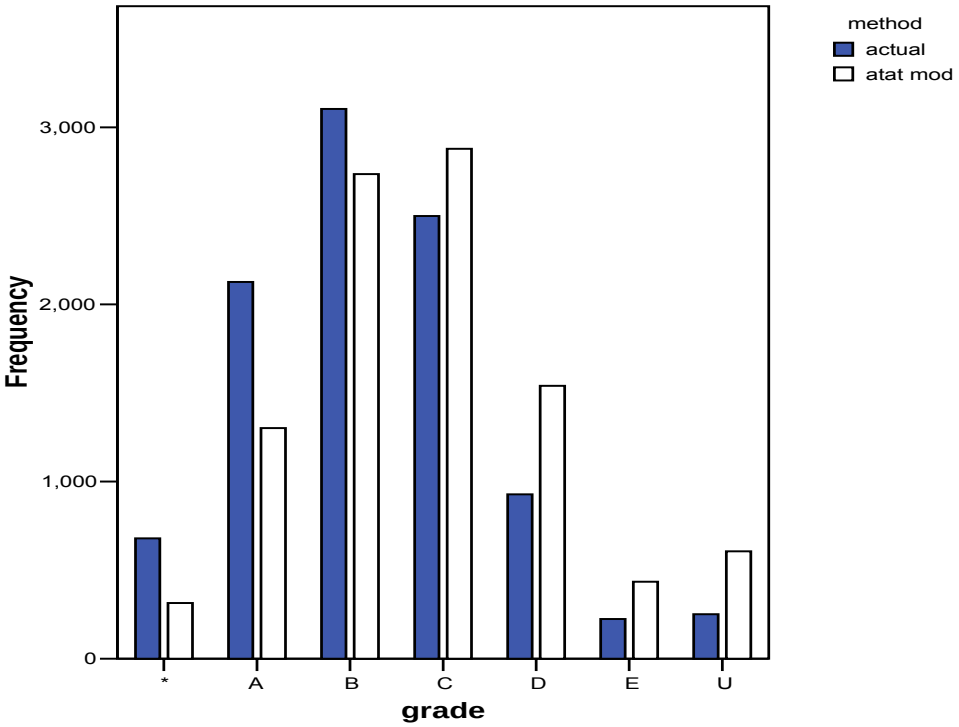
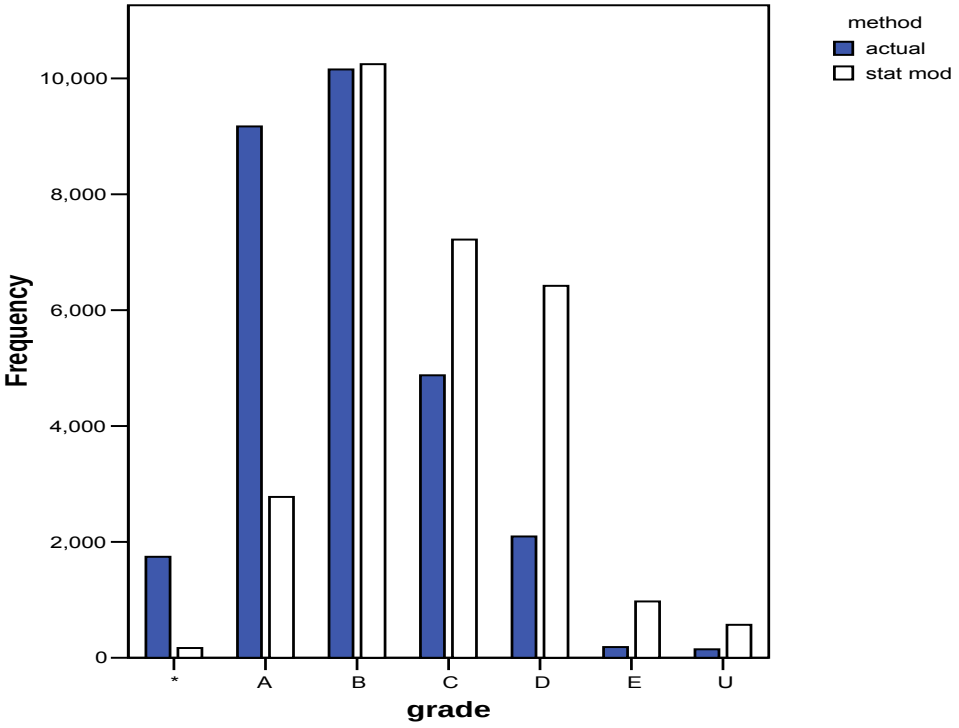


Figure 5.3d **GCSE Design & Technology (Food) Higher tier: effect on subject grades of replacing operational marks for the coursework component with statistically moderated marks (method (i), with no tolerance or allowed difference applied)**



Full summary statistics for Business Studies and Design & Technology are provided in Appendix B.

5.4 GCSE Music

This specification is considered separately because it has a different assessment structure from the other GCSE specifications in this study (in particular it has two centre-assessed coursework components) and because the proportions of centres which had adjustments to their coursework marks in 2004 were higher than in the other specifications (41.9% for Composing coursework and 26.8% for Performing coursework). Attention in this section is concentrated on Composing, since it had the higher proportion of centres with adjustments.

Table 5.4a shows summary statistics for Composing when no tolerance or allowed difference is applied. Figure 5.4a shows the outcomes diagrammatically for method (i).

Figure 5.4a GCSE Music Composing: differences between statistically moderated marks and operational marks under method (i) with no tolerance or allowed difference applied

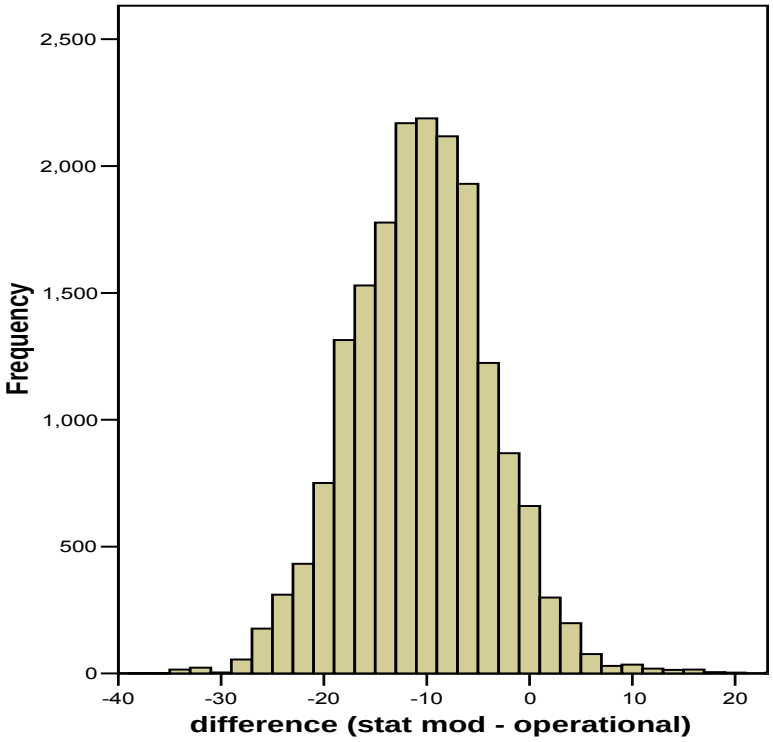


Table 5.4a GCSE Music Composing: statistics for methods (i)-(iv) with no tolerance or allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	18245			
Maximum mark for coursework	60			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	-11.0	-11.1	-11.1	-11.0
Mode	-10	-9	-10	-10
Standard deviation	6.7	7.6	8.6	8.0

In fact, the outcomes for Music Composing are similar to those for the Higher tier in both Business Studies and Design & Technology, where there was also a large negative mean difference.

Two further investigations were carried out for Music Composing. First, just those centres whose centre marks were accepted operationally were considered (ie the final marks for the component were centre marks[4]). Arguably (and under the assumption that the outcomes of the operational moderation by inspection process were correct), any adjustments made to these marks by statistical moderation are wrong. However, Table 5.4b shows that statistical moderation made substantial adjustments. In fact, the mean differences in Table 5.4b are *greater* than those in Table 5.4a.

Table 5.4b GCSE Music Composing (centres for which centre marks were accepted): statistics for methods (i)-(iv) with no tolerance or allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	11407[5]			
Maximum mark for coursework	60			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	-12.4	-12.4	-12.4	-12.4
Mode	-10	-11	-12	-13
Standard deviation	6.2	7.5	8.6	7.9

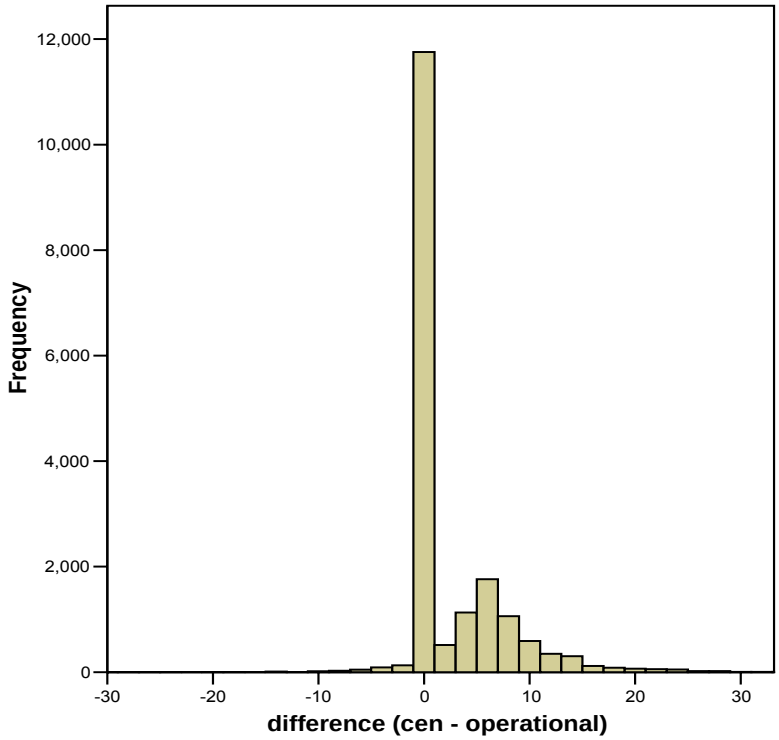
The second additional investigation carried out in Music Composing was to repeat the calculations with the statistically moderated marks replaced by centre marks. In this case, despite the large proportion of adjustments made operationally, consistency with the operational outcomes is greater than when statistical moderation is used (see Table 5.4c and Figure 5.4b).

[4] ie the marks originally awarded by the centre
[5] This number is not 41.9% (the proportion of adjustments, quoted earlier) of the number of candidates in Table 5.4a, because this proportion applies to the number of *centres* rather than the number of *candidates.*

Table 5.4c GCSE Music Composing: differences between centre marks and operational marks

Number of candidates	18245
Maximum mark for coursework	60
Weighting (%)	25%
Mean difference	2.3
Mode	0
Standard deviation	4.6

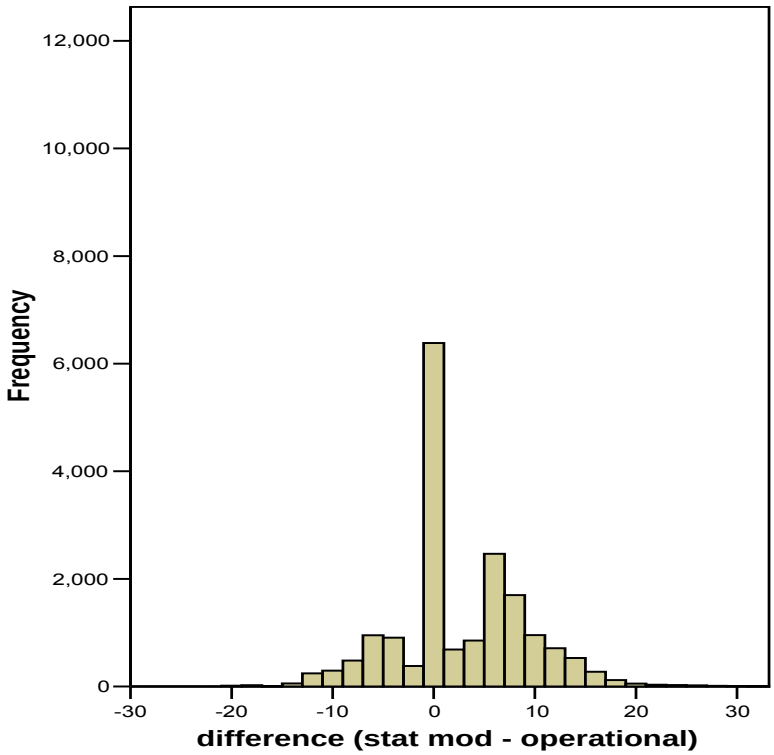
Figure 5.4b GCSE Music Composing: differences between centre marks and operational marks



It is clear that, although the use of centre marks produces a positive mean difference, the differences are generally closer to zero than when statistical moderation is used.

It was seen in section 5.1.1 above that, when an allowed difference is used, the statistically moderated marks were much more in line with the operational marks. Figure 5.4c shows the outcomes for statistically moderated marks in GCSE Music when an allowed difference was used (the vertical scale is the same as in Figure 5.4b to aid comparison).

Figure 5.4c GCSE Music Composing: differences between statistically moderated marks and operational marks under method (i) with tolerance and allowed difference both applied



Figures 5.4d and 5.4e show that replacing the operational marks for Composing coursework with centre marks (5.4d) and with statistically moderated marks with an allowed difference applied (5.4e) both had little effect on the numbers of candidates in each grade. However, inspection of the corresponding crosstabulations shows that many more candidates would change grade under statistical moderation (Tables 5.4d and 5.4e).

Figure 5.4d **GCSE Music: effect on subject grades of replacing operational marks for Composing coursework with centre marks**

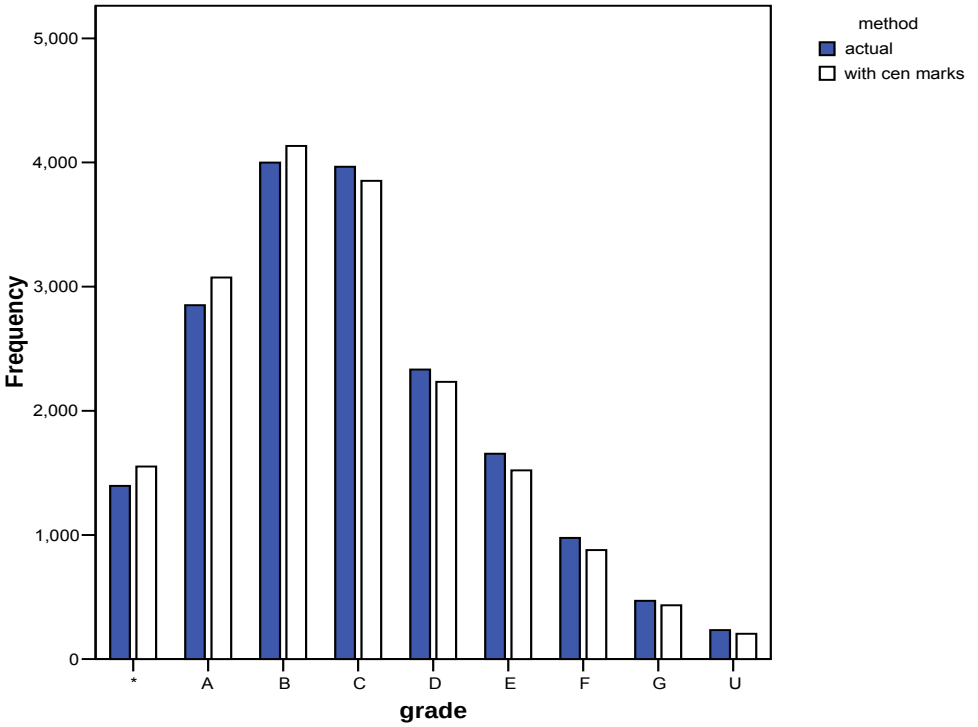


Figure 5.4e **GCSE Music: effect on subject grades of replacing operational marks for Composing coursework with statistically moderated marks (method (i), with tolerance and allowed difference applied)**

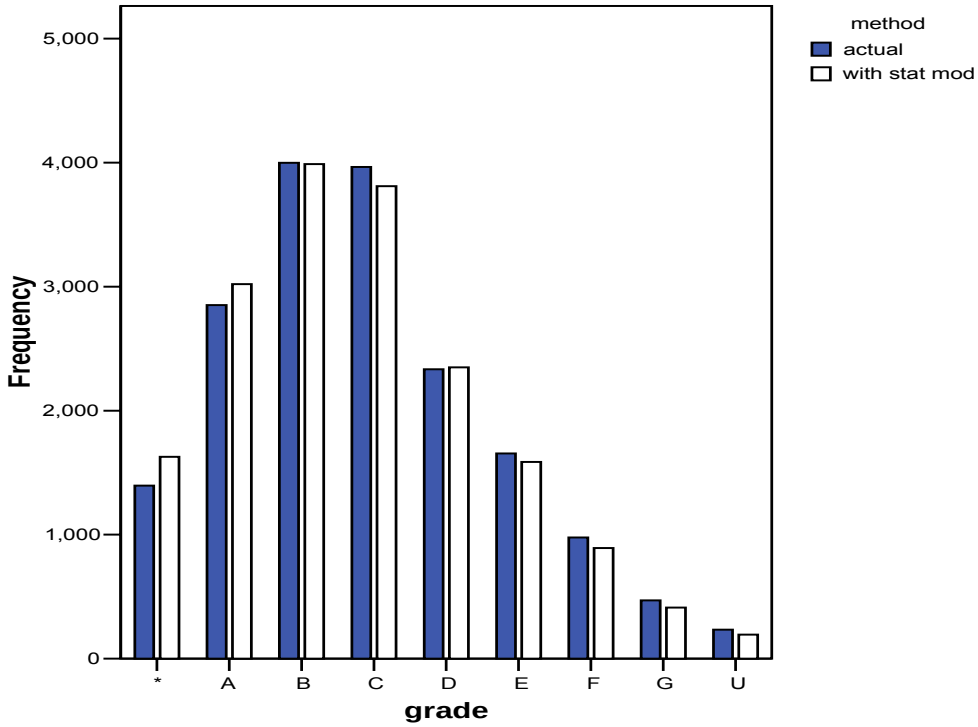


Table 5.4d GCSE Music: crosstabulation showing effect on subject grades of replacing operational marks for Composing coursework with centre marks

(Table shows numbers of candidates)

		Grade with centre marks for Composing coursework									Total
		*	A	B	C	D	E	F	G	U	
Actual grade	A*	1387	7	0	1	0	0	0	0	0	1395
	A	164	2671	14	2	0	0	0	0	0	2851
	B	0	393	3589	17	0	0	0	0	0	3999
	C	0	3	523	3421	19	0	0	0	0	3966
	D	0	0	8	391	1920	12	2	0	0	2333
	E	0	0	0	20	283	1342	10	0	0	1655
	F	0	0	0	0	11	164	792	9	1	977
	G	0	0	0	0	0	3	74	389	4	470
	U	0	0	0	0	0	0	0	35	199	234
Total		1551	3074	4134	3852	2233	1521	878	433	204	17880

Table 5.4e GCSE Music: crosstabulation showing effect on subject grades of replacing operational marks for Composing coursework with statistically moderated marks (method (i), with tolerance and allowed difference applied)

(Table shows numbers of candidates)

		Grade with statistical moderation for Composing coursework									
		A*	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**U**	**Total**
Actual grade	**A***	1300	95	0	0	0	0	0	0	0	1395
	A	327	2347	176	1	0	0	0	0	0	2851
	B	1	576	3155	267	0	0	0	0	0	3999
	C	0	3	654	3040	267	2	0	0	0	3966
	D	0	0	4	489	1679	161	0	0	0	2333
	E	0	0	0	14	394	1158	85	4	0	1655
	F	0	0	0	0	9	262	667	38	1	977
	G	0	0	0	0	0	3	139	311	17	470
	U	0	0	0	0	0	0	0	58	176	234
Total		1628	3021	3989	3811	2349	1586	891	411	194	17880

5.5 GCE specifications

In many of the GCE specifications considered in this study, statistical moderation produced similar outcomes to those for the GCSE specifications. The main difference is that in some cases the standard deviation of the differences was lower under method (iii) than under methods (ii) and (iv) (cf Table 5.1.1b). However, in a few specifications the mean difference between the statistically moderated marks and the operational marks for the coursework component was negligible, a feature not found in any of the GCSE specifications considered.

Section 5.5.1 below describes the outcomes for the specifications which had negligible mean differences. In these and the other GCE specifications considered there is just one coursework unit; the other units were aggregated to form the moderating instrument. For Psychology, the use of a single unit as the moderating instrument was also investigated. The outcomes are described in section 5.5.2.

GCE Geography Specification A was included in the study although it does not have a centre-assessed unit. The same statistical moderation procedures were applied to the externally-assessed coursework component in this specification and the marks compared with the operational marks. The adjustments were in fact larger than those in almost any other specification, even though (under the assumption that the marking of this unit was correct) there should be no adjustments. This is similar to the situation described for GCSE Music in Table 5.4b where only those centres whose marks were accepted without adjustment in the operational examination were included in the analysis.

Summary statistics for all of the GCE specifications considered are provided in Appendix C.

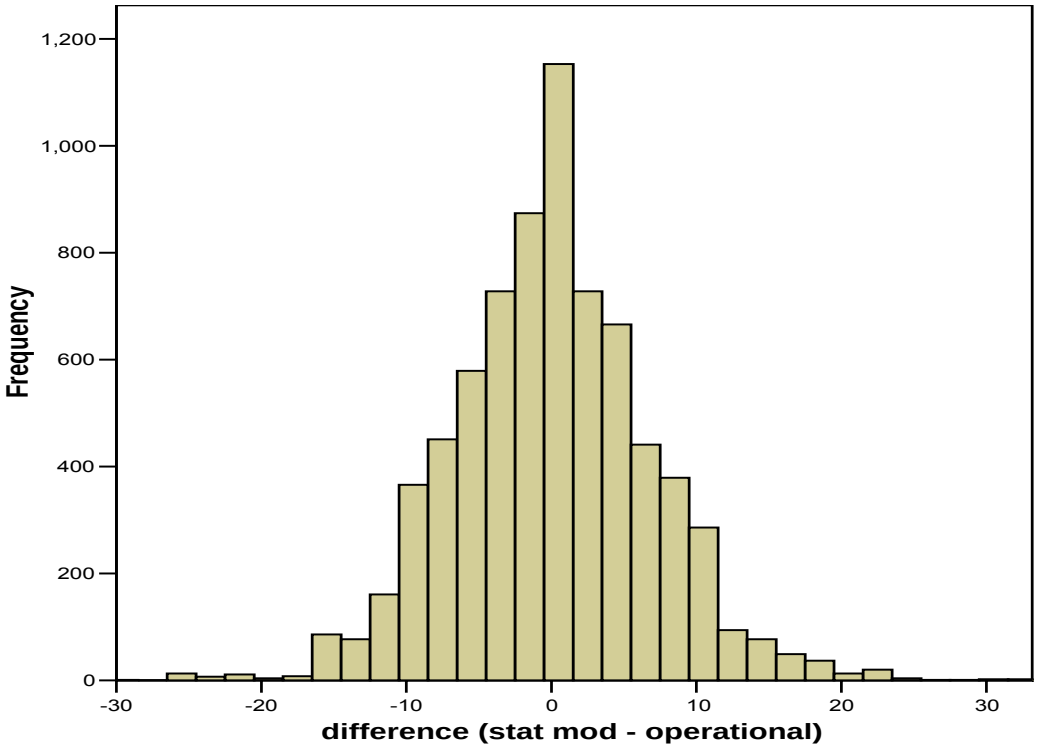
5.5.1 GCE Business Studies and GCE French

For these two specifications there were negligible mean differences between the statistically moderated marks and the operational marks for the coursework component. Table 5.5.1a and Figure 5.1.1a show the outcomes for GCE Business Studies with no tolerance or allowed difference applied.

Table 5.5.1a GCE Business Studies: statistics for methods (i)-(iv) with no tolerance or allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	7323			
Max mark for coursework unit	84			
Weighting (%)	15% (of the total A level assessment)			
Tolerance	6 marks			
Mean difference	0.0	0.1	0.1	0.1
Mode	1	-1	0	0
Standard deviation	6.9	7.4	8.8	8.0

Figure 5.5.1a GCE Business Studies: differences between statistically moderated marks and operational marks under method (i) with no tolerance or allowed difference applied



It can be seen both from the standard deviations in Table 5.5.1a and from the spread of differences evident in Figure 5.5.1a that, despite the mean difference of zero, the coursework marks received by candidates under statistical moderation would be very different from those received operationally. However, because of the relatively low weighting of coursework in this specification, relatively few candidates would change subject grade, as shown in the crosstabulation in Table 5.5.1b.

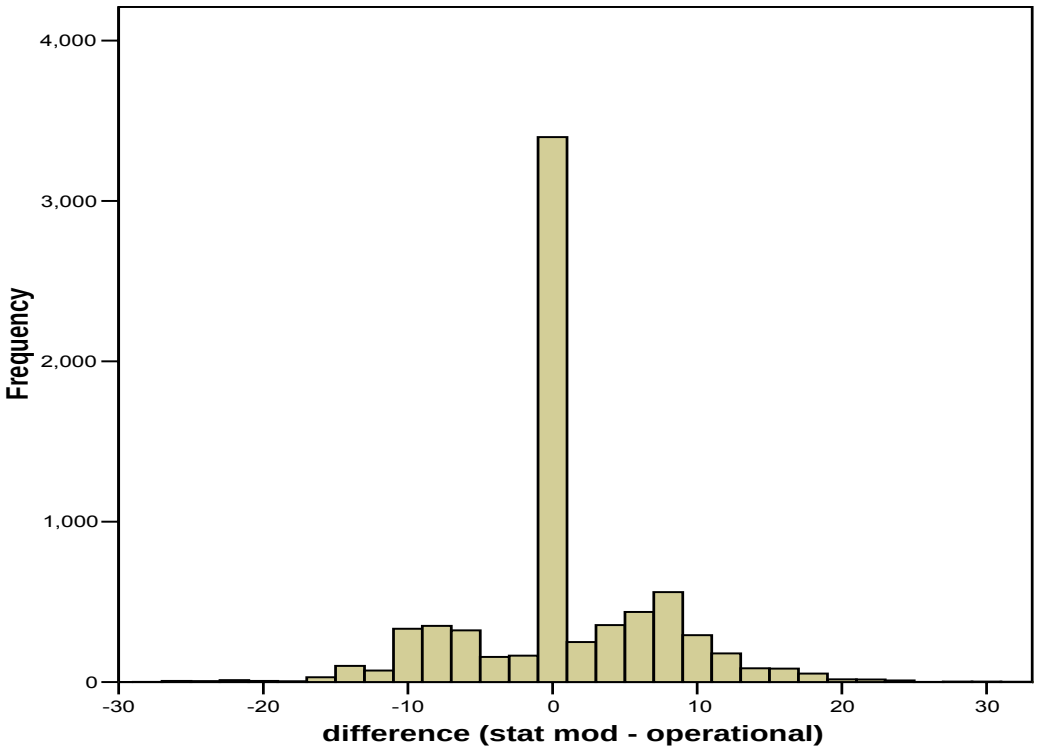
Table 5.5.1b GCE Business Studies: crosstabulation showing effect on subject grades of replacing operational marks for the coursework unit with statistically moderated marks (method (i), with no tolerance or allowed difference applied)

(Table shows numbers of candidates)

		Grade with statistical moderation for the coursework unit						
		A	**B**	**C**	**D**	**E**	**U**	**Total**
Actual grade	**A**	1140	49	0	0	0	0	1189
	B	93	1623	95	0	0	0	1811
	C	0	104	1754	123	0	0	1981
	D	0	0	87	1340	71	0	1498
	E	0	0	0	40	638	43	721
	U	0	0	0	0	7	116	123
Total		1233	1776	1936	1503	716	159	7323

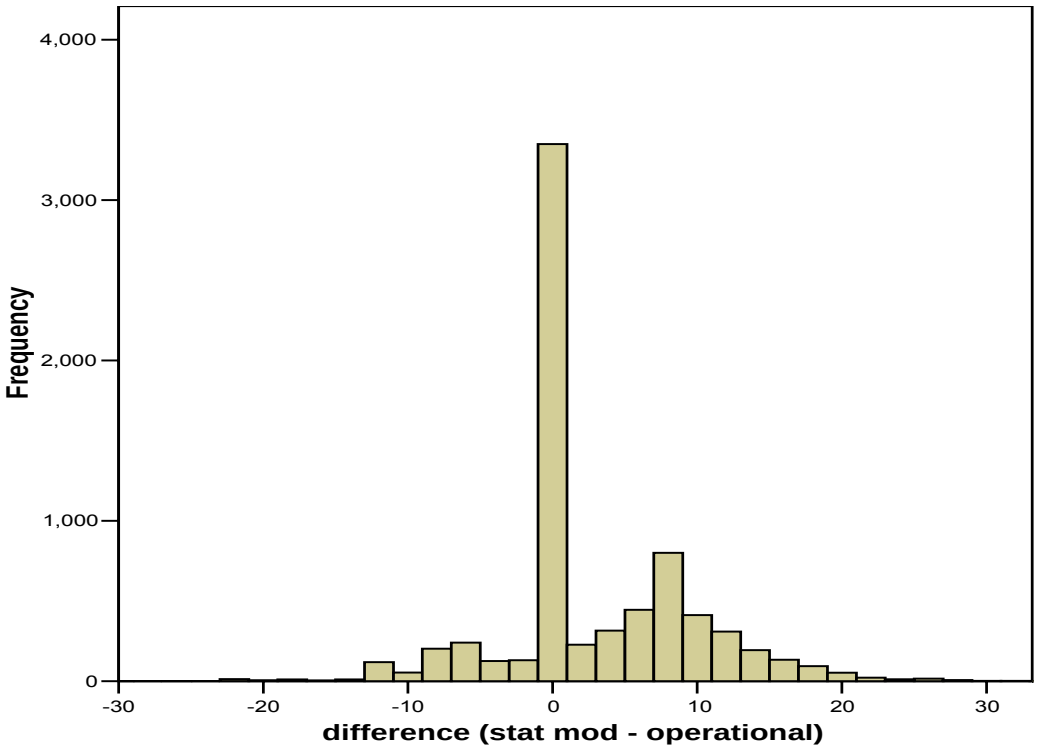
Application of tolerance tended to reduce to zero the differences which were close to zero, although the standard deviation of the differences changed very little. Figure 5.5.1b illustrates.

Figure 5.5.1b GCE Business Studies: differences between statistically moderated marks and operational marks under method (i) with tolerance applied but no allowed difference



When tolerance and allowed difference were both applied, the mean difference rose to 2.6 under all four methods but the standard deviations remained about the same. Figure 5.5.1c illustrates.

Figure 5.5.1c GCE Business Studies: differences between statistically moderated marks and operational marks under method (i) with tolerance and allowed difference applied



It appears that in this specification the effect of this method of statistical moderation was generally to increase candidates' marks for the coursework unit.

The outcomes for GCE French were very similar. Summary statistics are provided in Appendix C.

5.5.2 GCE Psychology

For this specification, use of a single unit as moderating instrument, as well as the aggregation of all of the written units, was investigated. Using a single unit gave rise to a numerically larger (negative) mean difference and a larger standard deviation. Tables 5.5.2a and 5.5.2b contain summary statistics and the outcomes are shown diagrammatically in Figures 5.5.2a and 5.5.2b.

Table 5.5.2a GCE Psychology (moderating instrument is aggregation of all written units): statistics for methods (i)-(iv) with no tolerance or allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	24504			
Max mark for coursework unit	60			
Weighting (%)	15%			
Tolerance	4 marks			
Mean difference	-8.7	-8.6	-8.6	-8.6
Mode	-8	-9	-9	-7
Standard deviation	4.8	5.7	5.2	6.0

Table 5.5.2a GCE Psychology (moderating instrument is Unit 5): statistics for methods (i)-(iv) with no tolerance or allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	24504			
Max mark for coursework unit	60			
Weighting (%)	15%			
Tolerance	4 marks			
Mean difference	-14.7	-14.6	-14.6	-14.6
Mode	-9	-16	-11	-18
Standard deviation	7.3	9.4	7.8	10.0

Figure 5.5.2a GCE Psychology (moderating instrument is aggregation of all written units): differences between statistically moderated marks and operational marks under method (i) with no tolerance or allowed difference applied

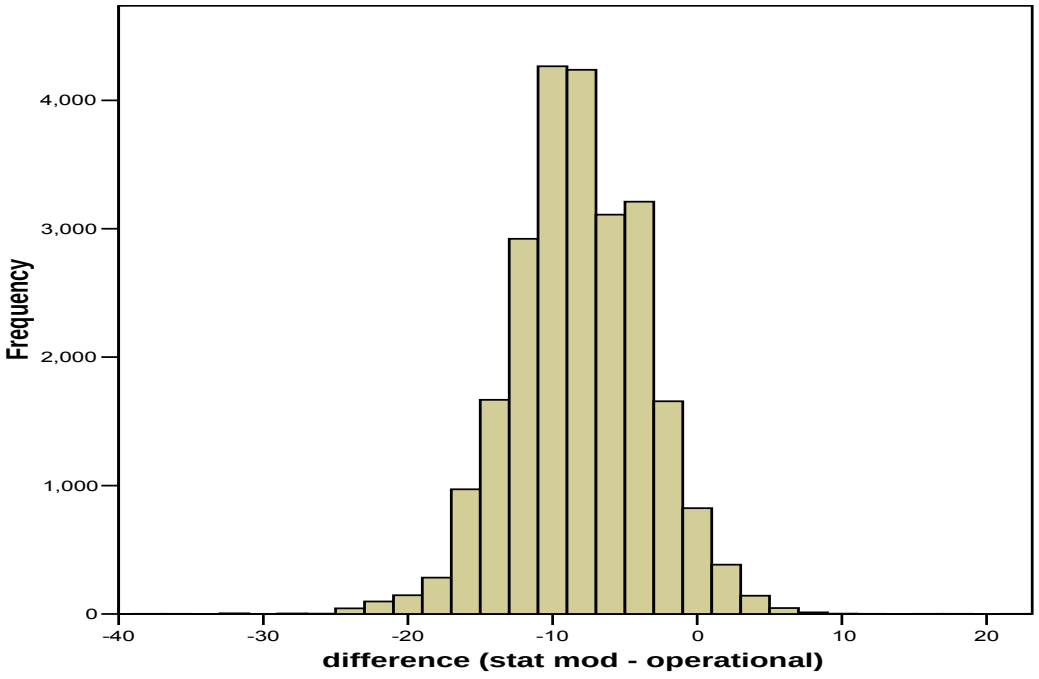
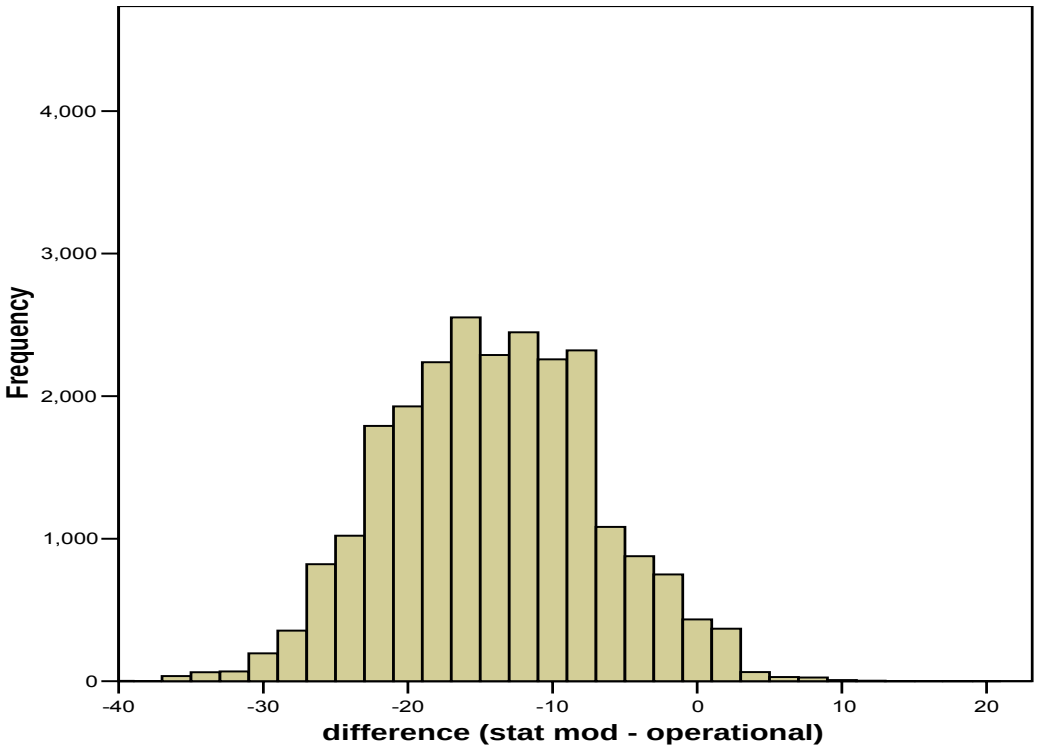


Figure 5.5.2b GCE Psychology (moderating instrument is Unit 5): differences between statistically moderated marks and operational marks under method (i) with no tolerance or allowed difference applied



In both cases, but particularly where Unit 5 was used as moderating instrument, many candidates would lose a grade, as shown in the crosstabulation in Table 5.5.2c.

Table 5.5.2c GCE Psychology (moderating instrument is Unit 5): crosstabulation showing effect on subject grades of replacing operational marks for the coursework unit with statistically moderated marks (method (i), with no tolerance or allowed difference applied)

(Table shows numbers of candidates)

		Grade with statistical moderation for the coursework unit						
		A	**B**	**C**	**D**	**E**	**U**	**Total**
Actual grade	**A**	2620	1901	2	0	0	0	4523
	B	20	2580	2683	3	0	0	5286
	C	0	6	2830	2832	2	0	5670
	D	0	0	8	2551	2107	1	4667
	E	0	0	0	3	1799	1033	2835
	U	0	0	0	0	0	1073	1073
Total		2640	4487	5523	5389	3908	2107	24054

The outcomes with tolerance and/or allowed difference applied were similar to those for other specifications, with the spread of differences when the moderating instrument was Unit 5 alone continuing to be greater than when the moderating instrument was the aggregation of all written units. Full summary statistics are provided in Appendix C.

6. TEACHER ASSESSMENT: RESULTS

As in section 5, it would be impracticable to provide full details of the findings. However, the outcomes followed a limited number of distinct patterns and it has been possible to place most of the cases investigated into categories corresponding to these patterns. Thus, all of the outcomes can be illustrated using relatively few examples.

As indicated in section 4.3, only GCE specifications were considered in this part of the study.

As explained in section 4.1, candidates' uniform marks for a certain unit or units were replaced by the centres' estimated grades ('teacher assessments') and the resulting subject grades were compared with those obtained operationally. The details, including the unit(s) for which teacher assessments were used, were varied from one specification to another, partly because of differing assessment structures. Teacher assessment was used both for just a single unit and for several units. For example, in Communication Studies separate investigations were carried out using teacher assessment for Units 1-5, for Units 1, 2 and 4, and for Unit 6. Both unmoderated and (statistically) moderated teacher assessments were used. The moderation was carried out using one of the methods from Part 1 of the study, namely adjustment of centre mean marks with no tolerance or allowed difference applied (this was the most straightforward of the methods investigated). As described in section 4.1, two ways of making the comparisons between teacher assessment and moderating instrument were used. For example, when teacher assessment was used for Units 1-5 (and the moderating instrument was Unit 6), moderation was carried out first by comparing the teacher assessment mean for the centre for Units 1-5 with the actual (external) mean for the centre for Unit 6 and second by comparing the teacher assessment mean for the centre for Unit 6 with the actual mean for the centre for Unit 6. Thus, letting *c* be the teacher assessment mark for Units 1-5, *d* the teacher assessment mark for Unit 6, *w* the external mark for Unit 6 and *y* the moderated teacher assessment mark for Units 1-5,

$$y = c - \bar{c} + \bar{w} \quad \text{under the first method}$$

and $$y = c - \bar{d} + \bar{w} \quad \text{under the second method,}$$

where $\bar{c}$, $\bar{d}$ and $\bar{w}$ are the means for the centre of *c*, *d* and *w*.

The second method superficially appears more logical, but (as noted in section 4.1) in practice it would require centres to provide teacher assessments for Unit 6 merely for the purpose of moderation and could therefore lead to abuse.

The outcomes of this part of the study depend heavily on both the accuracy of centres' estimated grades and on the effectiveness of the statistical moderation. It was noted earlier that the study does not seek to investigate the accuracy of estimated grades, but for the record the evidence in one specification (GCE Business Studies) are is presented in Appendix D.

6.1 Category 1

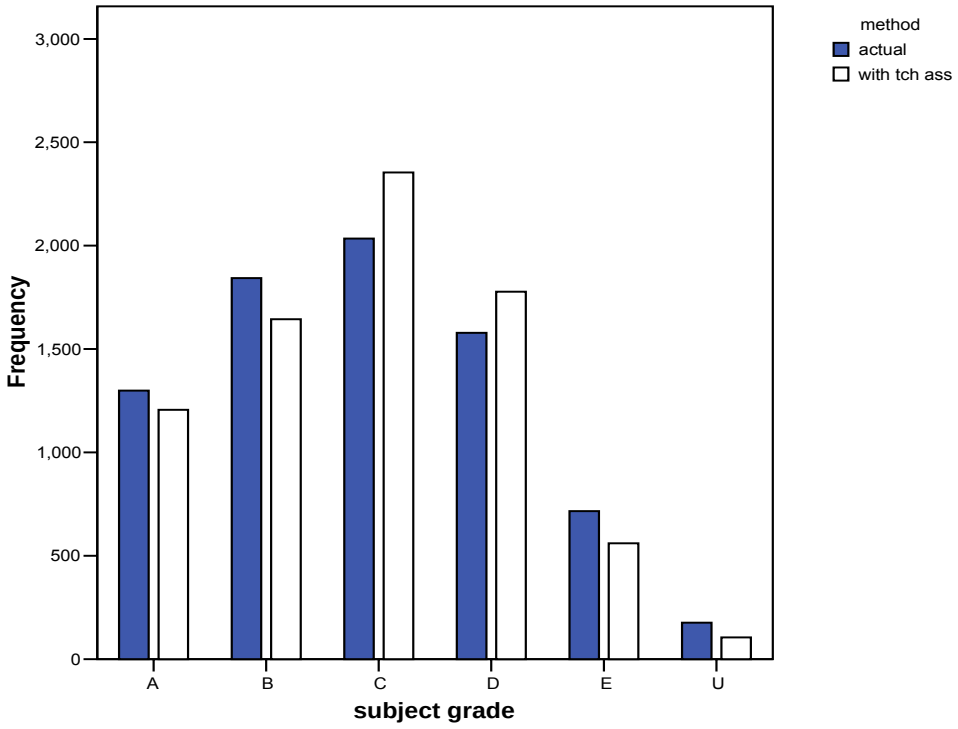
In this category teacher assessment tends to create bunching. Business Studies, with teacher assessment for Units 1-5, is used to illustrate.

Table 6.1a GCE Business Studies: crosstabulation showing effect on subject grades of replacing operational marks for Units 1-5 with unmoderated teacher assessments

(Table shows numbers of candidates)

		Grade with teacher assessments						
		A	B	C	D	E	U	Total
Actual grade	A	849	373	74	3	0	0	1299
	B	328	843	605	64	3	0	1843
	C	29	394	1140	443	24	4	2034
	D	0	32	481	874	168	23	1578
	E	0	2	49	360	265	40	716
	U	0	0	5	33	100	38	176
Total		1206	1644	2354	1777	560	105	7646

Figure 6.1a GCE Business Studies: bar chart showing effect on subject grades of replacing operational marks for Units 1-5 with unmoderated teacher assessments



The main feature here is that, when teacher assessment was incorporated, the numbers of candidates in a grade were lower at the top and bottom but higher in the middle. The crosstabulation shows that many candidates would change grade, by up to three grades.

Other specifications and configurations with a similar pattern are listed in Table 6.1b. The trend at the middle grades is not exactly the same in all of these cases (for example, at grade B the 'with teacher assessment' number was sometimes higher than the actual number and at grade D the opposite was sometimes true) but it was always the same at grades A, E and U.

Table 6.1b Specifications with the same pattern as in Figure 6.1a

Specification	Units with teacher assessment
Business Studies	1-5
Communication Studies	1-5
Communication Studies	1, 2, 4
Geography A	1-5
ICT	1-4, 6
Psychology A	1-4, 6

6.2 Category 2

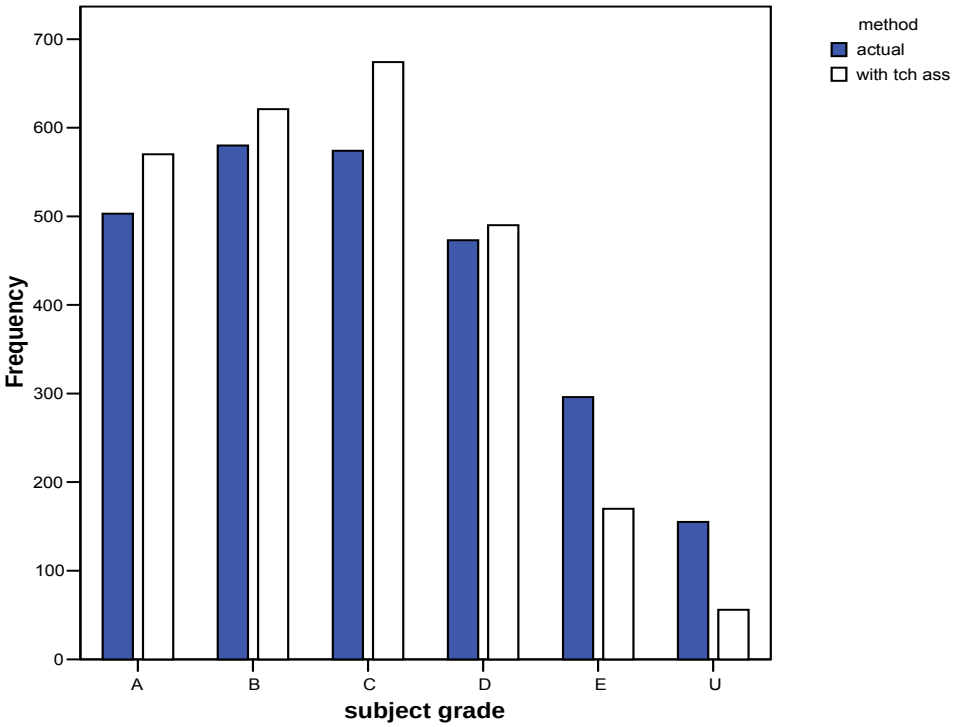
In this category teacher assessment tends to create grade inflation. Law, with teacher assessment for Units 1-5, is used to illustrate.

Table 6.2a GCE Law: crosstabulation showing effect on subject grades of replacing operational marks for Units 1-5 with unmoderated teacher assessments

(Table shows numbers of candidates)

		Grade with teacher assessments						
		A	B	C	D	E	U	Total
Actual grade	A	396	96	10	1	0	0	503
	B	153	292	124	11	0	0	580
	C	19	196	273	81	3	2	574
	D	2	33	203	198	30	7	473
	E	0	3	52	159	67	15	296
	U	0	1	12	40	70	32	155
Total		570	621	674	490	170	56	2581

Figure 6.2a GCE Law: bar chart showing effect on subject grades of replacing operational marks for Units 1-5 with unmoderated teacher assessments



The main feature here is that, when teacher assessment is incorporated, the numbers of candidates in a grade are higher at the top but lower at the bottom. As in Category 1, the crosstabulation shows that many candidates change grade, here by as much as four grades.

Other specifications and configurations with a similar pattern are listed in Table 6.2b. The trend at the middle grades is not exactly the same in all of these cases (for example, at grade B the 'with teacher assessment' number was sometimes lower than the actual number) but it was always the same at grades A, E and U.

Table 6.2b Specifications with the same pattern as in Figure 6.2a

Specification	Units with teacher assessment
Communication Studies	6
English Literature A	1-5
English Literature A	6
French	1, 2, 3, 5, 6
Geography B	1-5
Geography B	1-4, 6
ICT	1-4, 6
Law	1-5

6.3 Category 3

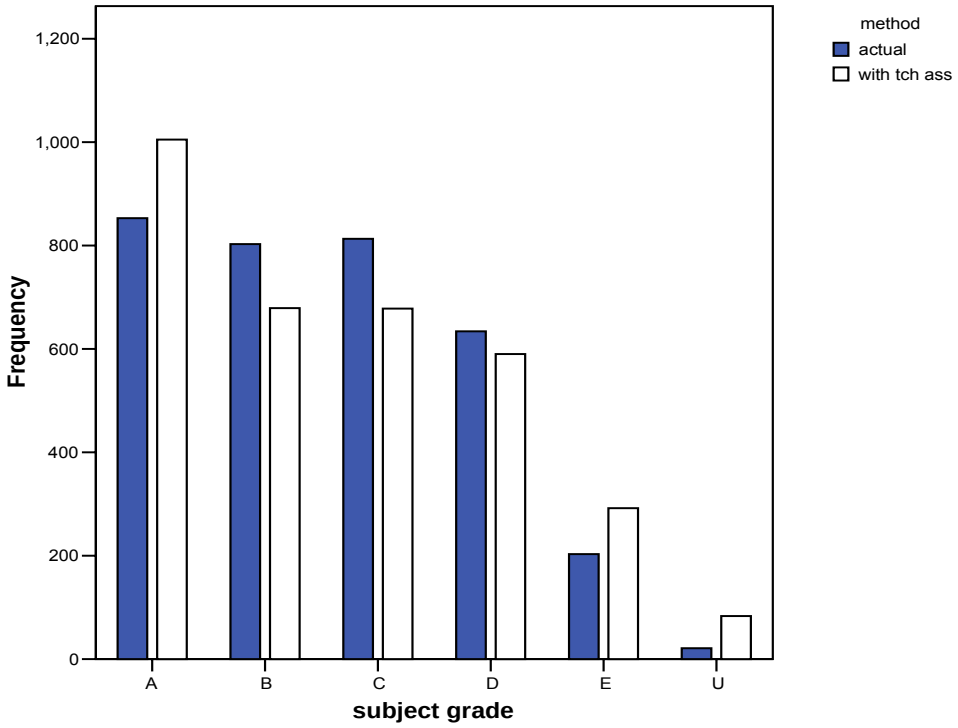
In this category teacher assessment tends to create greater dispersion or discrimination. English Literature Specification B, with teacher assessment for Units 3 and 4, is used to illustrate.

Table 6.3a GCE English Literature B: crosstabulation showing effect on subject grades of replacing operational marks for Units 3 and 4 with unmoderated teacher assessments

(Table shows numbers of candidates)

		Grade with teacher assessments						
		A	**B**	**C**	**D**	**E**	**U**	Total
Actual grade	A	843	10	0	0	0	0	853
	B	162	584	57	0	0	0	803
	C	0	85	582	146	0	0	813
	D	0	0	39	437	158	0	634
	E	0	0	0	7	133	63	203
	U	0	0	0	0	1	20	21
Total		1005	679	678	590	292	83	3327

Figure 6.3a GCE English Literature B: bar chart showing effect on subject grades of replacing operational marks for Units 3 and 4 with unmoderated teacher assessments



The main feature here is that, when teacher assessment was incorporated, the numbers of candidates in a grade were higher at the top and bottom but lower in the middle. The crosstabulation shows that the maximum change

was one grade, much less dramatic than in Tables 6.1a and 6.2a. There is a simple reason: teacher assessment was used for only two units, in contrast to Tables 6.1a and 6.2a where it was used for five units.

The only other example found which followed the same pattern was AS English Literature B, with teacher assessment for Unit 3.

6.4 Category 4

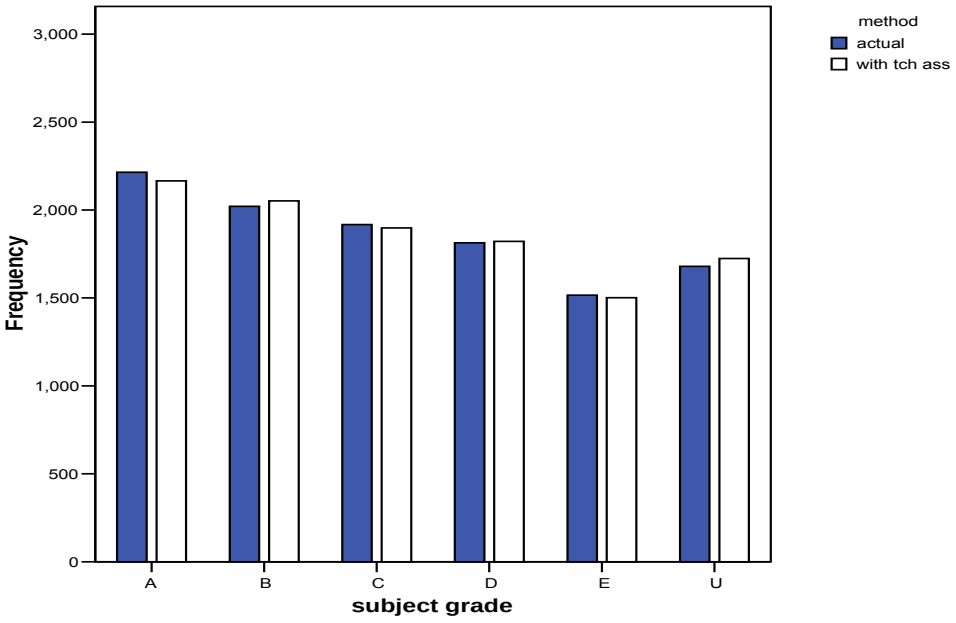
In this category there is little change. AS Biology Specification A, with teacher assessment for the third of the three units[6], was the only example found.

Table 6.4a GCE AS Biology Specification A: crosstabulation showing effect on subject grades of replacing operational marks for the third unit with unmoderated teacher assessments

(Table shows numbers of candidates)

		Grade with teacher assessments						
		A	B	C	D	E	U	Total
Actual grade	A	1997	216	2	0	0	0	2215
	B	169	1644	207	1	0	0	2021
	C	0	190	1529	197	1	0	1917
	D	0	2	160	1453	195	4	1814
	E	0	0	0	171	1195	150	1516
	U	0	0	0	0	110	1570	1680
Total		2166	2052	1898	1822	1501	1724	11163

Figure 6.4a GCE AS Biology Specification A: bar chart showing effect on subject grades of replacing operational marks for the third unit with unmoderated teacher assessments



[6] This is actually Unit 4. Unit 3 is part of the parallel Human Biology specification; candidates for AS Biology take Units 1, 2 and 4.

When teacher assessment was incorporated, the numbers of candidates in a grade did not change much. However, the cross-tabulation shows that there would be a change of one grade for many candidates.

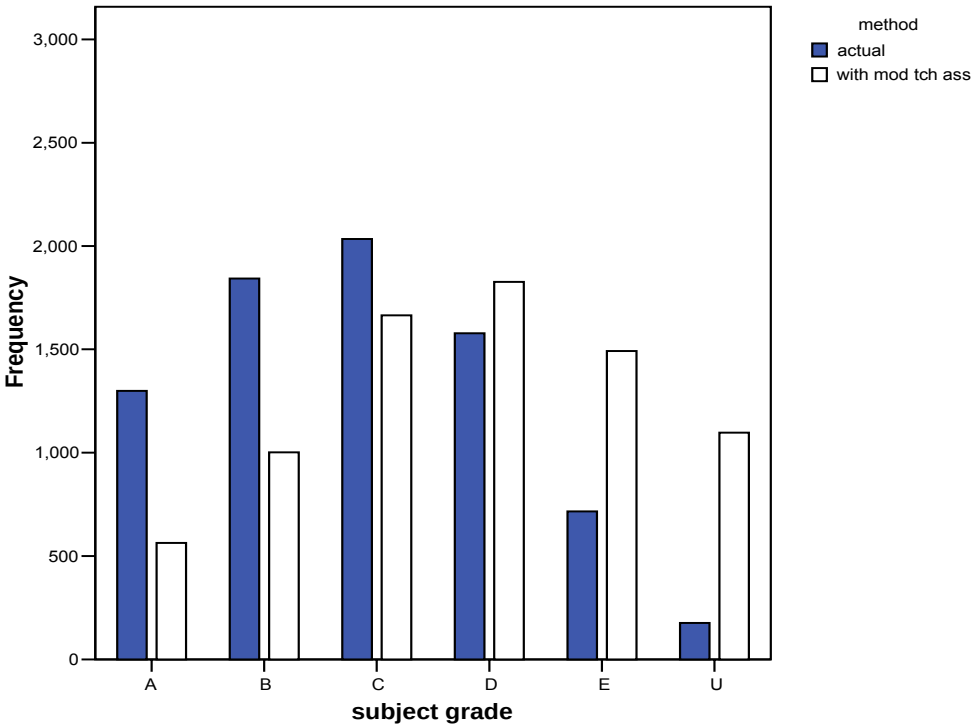
6.5 Effect of moderating the teacher assessments

With a few exceptions, moderation had a similar effect on all of the specifications/configurations within a category. This effect is illustrated by means of the examples in sections 6.5.1 - 6.5.4 below.

6.5.1 Category 1

Figure 6.5.1a shows the outcomes when the teacher assessments for Business Studies (Units 1-5) were moderated. The vertical scale is the same as in Figure 6.1a.

Figure 6.5.1a GCE Business Studies: bar chart showing effect on subject grades of replacing operational marks for Units 1-5 with teacher assessments moderated by comparing the centre's mean teacher assessment for Units 1-5 with the centre's mean external mark for Unit 6



The specifications and configurations in Table 6.5.1a had the same broad pattern of grades, after statistical moderation was carried out, as the above Business Studies example.

Table 6.5.1a Specifications/configurations in Category 1 with the same pattern after statistical moderation as in Figure 6.5.1a

Specification	Units with tch ass	Moderating instrument (Unit(s))	Units whose tch ass mean was compared with the mod inst mean	Max mark for tch ass[7]	Mean diff (moderated tch ass - original tch ass)
Bus Studs	1-5	6	1-5	25	-2.9
Comm Studs	1-5	6	1-5	25	-1.8
Geography A	1-5	6	1-5	25	-0.9
ICT	1-4 & 6	5	1-4 & 6	25	-3.4
Psychology A	1-4 & 6	5	1-4 & 6	25	-3.7

In some other cases the statistical moderation had little effect and the pattern of grades was broadly the same as for the unmoderated teacher assessments. These cases are listed in Table 6.5.1b.

Table 6.5.1b Specifications/configurations in Category 1 where statistical moderation had little effect on the pattern of grades (ie pattern is similar to that in Figure 6.1a)

Specification	Units with tch ass	Moderating instrument (Unit(s))	Units whose tch ass mean was compared with the mod inst mean	Max mark for tch ass	Mean diff (moderated tch ass - original tch ass)
Bus Studs	1-5	6	6	25	-0.5
Comm Studs	1-5	6	6	25	-0.3
Comm Studs	1, 2, 4	3, 5, 6	1, 2, 4	15	-0.5
Comm Studs	1, 2, 4	3, 5, 6	3, 5, 6	15	-0.3
Geography A	1-5	7[8]	7	25	-0.2
ICT	1-4, 6	5	5	25	-0.6
Psychology A	1-4, 6	5	5	25	-0.7

There is a clear pattern here. In Table 6.5.1b (except for Communication Studies) it is the teacher assessment for the moderating instrument (instead of the teacher assessment for the other units) which was used in the moderation process. In the examples studied, this procedure gives rise to a smaller mean difference and thus the moderation has less effect.

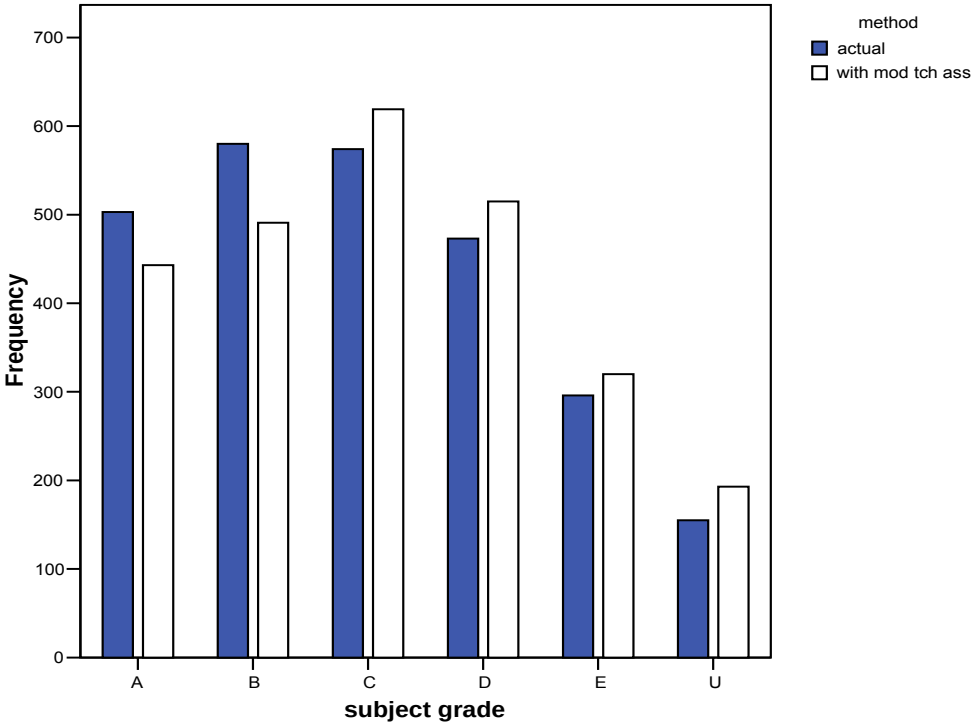
[7] As explained in section 4.1, the teacher assessment for a unit in the model used has a nominal maximum of 5. Thus, where there are five units, the maximum is 25.
[8] Candidates take either Unit 6 or Unit 7. The route consisting of Units 1-5 and 7 was the one considered in this study.

It was seen in sections 6.1 - 6.4 that there were many changes of grade when operational marks were replaced by teacher assessments. Applying statistical moderation to the teacher assessments did not reduce the number of changes (and often increased them).

6.5.2 Category 2

Figure 6.5.2a shows the outcomes when the teacher assessments for Law (Units 1-5) were moderated. The vertical scale is the same as in Table 6.2a.

Figure 6.5.2a GCE Law: bar chart showing effect on subject grades of replacing operational marks for Units 1-5 with teacher assessments moderated by comparing the centre's mean teacher assessment for Units 1-5 with the centre's mean external mark for Unit 6



In fact, the pattern is similar to the Business Studies example in Figure 6.5.1a.

The specifications and configurations in Table 6.5.2a had the same broad pattern of grades, after statistical moderation was carried out, as the above Law example.

Table 6.5.2a Specifications/configurations in Category 2 with the same pattern after statistical moderation as in Figure 6.5.2a

Specification	Units with tch ass	Moderating instrument (Unit(s))	Units whose tch ass mean was compared with the mod inst mean	Max mark for tch ass	Mean diff (moderated tch ass - original tch ass)
Eng Lit A	1-5	6	1-5	25	-2.8
French	1-3, 5, 6	4	1-3, 5, 6	25	-1.4
Geography B	1-4, 6	5	1-4, 6	25	-1.3
ICT	1-4, 6	5	1-4, 6	25	-3.4
Law	1-5	6	1-5	25	-1.4

In fact, the downward shift in grades was more substantial for English Literature and Law (thus, the moderated teacher assessment bar at grade A was lower and that at grade U was higher), as might be expected from the (numerically) larger mean differences.

In some other cases the statistical moderation had little effect and the pattern of grades was broadly the same as for the unmoderated teacher assessments. These cases are listed in Table 6.5.2b.

Table 6.5.2b Specifications/configurations in Category 2 where statistical moderation had little effect on the pattern of grades (ie pattern is similar to that in Figure 6.2a)

Specification	Units with tch ass	Moderating instrument (Unit(s))	Units whose tch ass mean was compared with the mod inst mean	Max mark for tch ass	Mean diff (moderated tch ass - original tch ass)
Comm Studs	6	1-5	6	5	-0.1
Eng Lit A	1-5	6	6	25	-0.5
Eng Lit A	6	1-5	6	5	-0.1
French	1-3, 5, 6	4	4	25	-0.2
Geography B	1-5	6	6	25	-0.2
Geography B	1-4, 6	5	5	25	-1.3

There are several specifications and configurations in Category 2 which are not covered in Tables 6.5.2a or 6.5.2b. In these cases statistical moderation had a significant effect on the teacher assessment outcomes but the pattern was different from that for Law. They are listed in Table 6.5.2c.

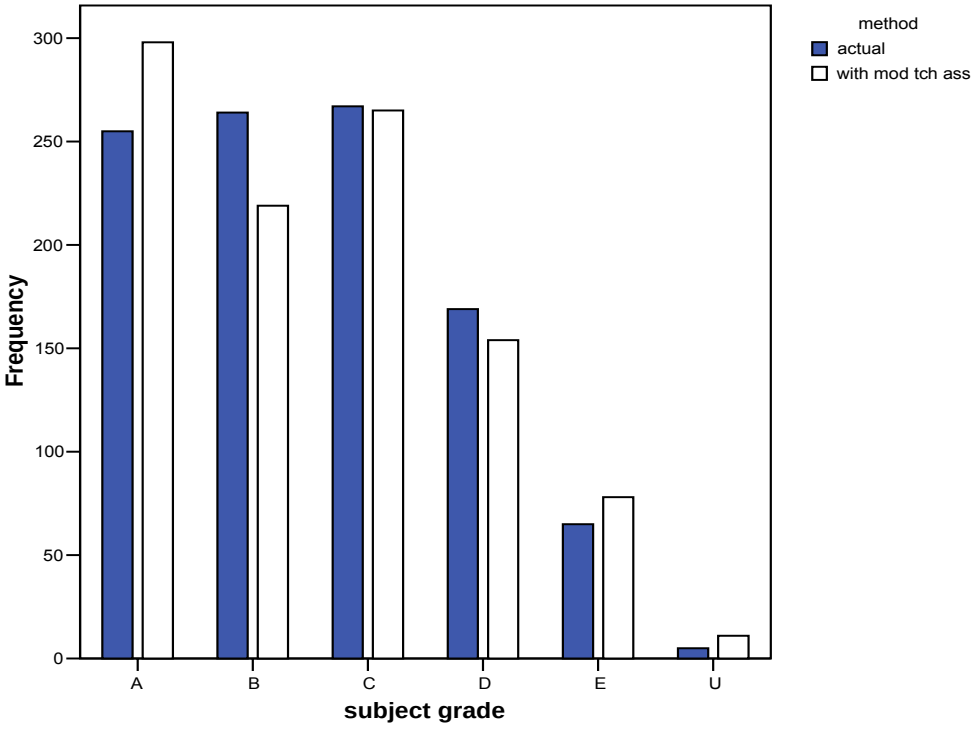
Table 6.5.2c Other specifications/configurations in Category 2

Specification	Units with tch ass	Moderating instrument (Unit(s))	Units whose tch ass mean was compared with the mod inst mean	Max mark for tch ass	Mean diff (moderated tch ass - original tch ass)
Comm Studs	6	1-5	1-5	5	-0.1
Eng Lit A	6	1-5	1-5	5	-0.5
Geography B	1-5	6	1-5	25	-0.8
ICT	1-4, 6	5	5	25	-0.6

For English Literature Specification A and ICT, the numbers of candidates per grade under moderated teacher assessment were similar to the operational numbers (so, in the bar chart, the 'actual' bar and the 'with moderated teacher assessment' bar would be of roughly the same height within each pair). However, as before, large numbers of candidates would receive different grades, so the results under moderated teacher assessment by no means replicated the operational results.

The pattern for Communication Studies is shown in Figure 6.5.2b. Geography B is similar.

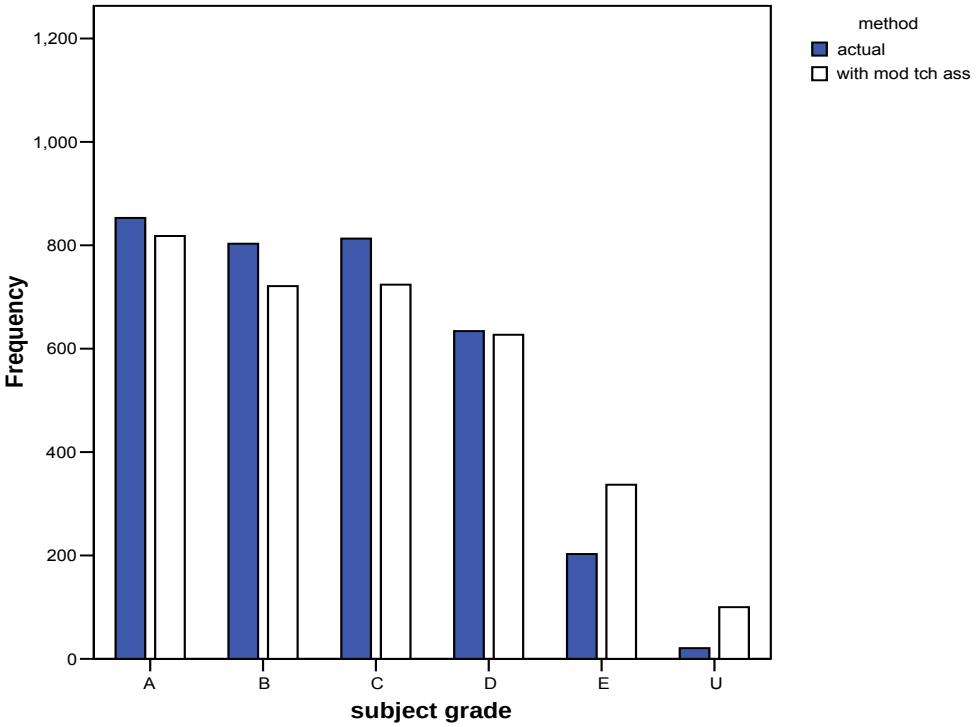
Figure 6.5.2b GCE Communication Studies: bar chart showing effect on subject grades of replacing operational marks with moderated teacher assessments, as in Table 6.5.2c



6.5.3 Category 3

Figure 6.5.3a shows the outcomes when the teacher assessments for English Literature Specification B (Units 3 and 4) were moderated. The vertical scale is the same as in Figure 6.3a.

Figure 6.5.3a GCE English Literature B: bar chart showing effect on subject grades of replacing operational marks for Units 3 and 4 with teacher assessments moderated by comparing the centre's mean teacher assessment for Units 1, 2, 5 and 6 with the centre's mean external mark for Units 1, 2, 5 and 6



The configurations in Table 6.5.3a had the same broad pattern of grades, after statistical moderation was carried out, as the above example.

Table 6.5.3a Specifications/configurations in Category 3 with the same pattern after statistical moderation as in Figure 6.5.3a

Specification	Units with tch ass	Moderating instrument (Unit(s))	Units whose tch ass mean was compared with the mod inst mean	Max mark for tch ass	Mean diff (moderated tch ass - original tch ass)
A level Eng Lit B	3 and 4	1, 2, 5, 6	3 and 4	10	-0.7
A level Eng Lit B	3 and 4	1, 2, 5, 6	1, 2, 5, 6	10	-0.6
AS Eng Lit B	3	2	3	5	-0.4

In all other cases investigated the statistical moderation had little effect and the pattern of grades was broadly the same as for the unmoderated teacher assessments. These cases are listed in Table 6.5.3b.

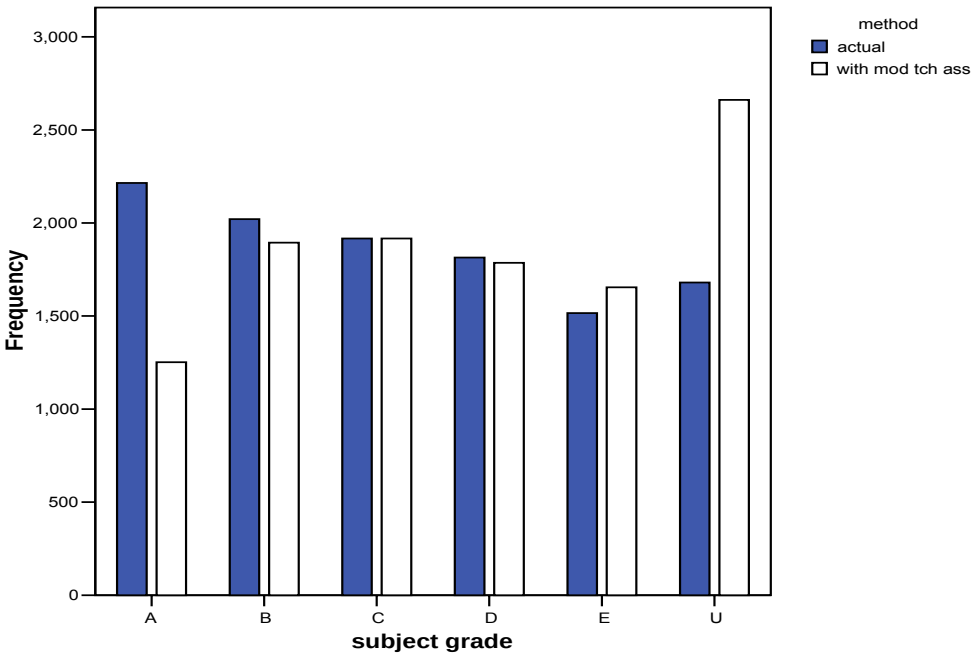
Table 6.5.3b Configurations in Category 3 where statistical moderation had little effect on the pattern of grades (ie pattern is similar to that in Figure 6.3a)

Specification	Units with tch ass	Moderating instrument (Unit(s))	Units whose tch ass mean was compared with the mod inst mean	Max mark for tch ass	Mean diff (moderated tch ass - original tch ass)
AS Eng Lit B	3	1	1	5	-0.1
AS Eng Lit B	3	2	2	5	-0.2
AS Eng Lit B	3	1 and 2	3	5	-0.2
AS Eng Lit B	3	1	3	5	-0.1

6.5.4 Pattern 4

Figure 6.5.4a shows the outcomes when the teacher assessments for AS Biology Specification A (Unit 4[9]) were moderated. The vertical scale is the same as in Figure 6.4a.

Figure 6.5.4a GCE AS Biology A: bar chart showing effect on subject grades of replacing operational marks for Unit 4 with teacher assessments moderated by comparing the centre's mean teacher assessment for Unit 4 with the centre's mean external mark for Units 1 and 2



[9] See footnote 6 on page 41.

All other configurations investigated (listed in Table 6.5.4a) had the same broad pattern of grades, after statistical moderation was carried out, as the above example.

Table 6.5.4a Specifications/configurations in Category 4 with the same pattern after statistical moderation as in Figure 6.5.4a

Specification	Units with tch ass	Moderating instrument (Unit(s))	Units whose tch ass mean was compared with the mod inst mean	Max mark for tch ass	Mean diff (moderated tch ass - original tch ass)
AS Biology A	4	1 and 2	4	5	-0.9
AS Biology A	4	1	4	5	-0.8
AS Biology A	4	2	4	5	-1.1
AS Biology A	4	1	1	5	-0.4
AS Biology A	4	2	2	5	-0.7

As before, the downward changes in grade were greatest for those configurations which had the (numerically) largest mean differences. Thus, in these cases, the number of candidates at grade A under statistical moderation was smaller and the number at grade E bigger.

7. CONCLUSIONS

7.1 Statistical moderation

The findings were disappointing in two respects. First, the outcomes appeared to be very different (at least at candidate level) from those obtained under the current system of moderation by inspection. Except when an 'allowed difference' was applied, statistical moderation generally lowered marks, because under current systems coursework components normally have higher mean marks than written papers. Even where the mean difference between the operational marks and the statistically moderated marks was close to zero, the spread of differences was still quite substantial, indicating that there would be changes to the marks of many candidates under statistical moderation. Where the numbers of candidates per subject grade under statistical moderation were approximately the same as the operational numbers, there were often, nevertheless, large numbers of candidates who would change grade (the numbers going down compensating for the numbers going up) – changes were limited only by the relatively low weighting for centre-assessed coursework. There was a more significant downward effect on grades in GCSE Design & Technology, which has a high coursework weighting (see Figure 5.3d).

The second disappointing feature is the absence of any pattern, across different specifications, with respect to the sizes of the adjustments arising from statistical moderation. For example, under method (i) with no tolerance

or allowed difference applied, the mean difference between statistically moderated coursework marks and operational coursework marks is -7.3/50[10] in GCSE History B, -3.8/60 in GCSE Humanities, -9.0/30 in GCE AS Biology A and -0.3/60 in GCE French (see Appendices A and C). Perversely, one of the largest mean differences (-18.8/100) was in GCE Geography, where the coursework unit is externally-assessed – these external marks are presumably 'correct', so arguably no adjustments should occur at all. However, the sizes of the mean differences are ultimately of no consequence, because current levels of achievement could be maintained by modifying grade boundaries. Of more importance are variations across centres and candidates. If the results under the current system are 'correct', then with statistical moderation many candidates would get 'wrong' results, even if there were no net changes in the numbers of candidates per grade.

The evidence from the investigations in GCSE Music (section 5.4) suggests that it would be more accurate simply to use the marks awarded by the centre instead of carrying out statistical moderation.

The common patterns which do emerge are somewhat overwhelmed by the factors discussed in the paragraphs above. It is perhaps surprising that the outcomes from all four methods of statistical moderation (adjustment of centre mean marks, linear scaling, linear regression within centre and mapping ranks) were so similar, particularly in the case of mapping ranks, which is fundamentally different from the other methods. The mean differences between statistically moderated marks and operational marks were approximately the same under all four methods, with variations only in the spread of these differences. In most specifications the same method (linear regression within centre) produced the greatest spread of differences. Adjustment of centre mean marks always produced the lowest spread. It should be noted in passing that, if statistical moderation had to be carried out on assessments consisting of a rank ordering of candidates rather than marks, only the method of mapping ranks would be available.

Statistical moderation generally fell out of use in UK examinations in the 1980s, and, in the light of the findings in this study, quite rightly. Of course, the underlying assumption is that the marks arising from moderation by inspection are 'correct'. In fact, this assumption is untrue for two reasons. First, an investigation of the reliability of coursework moderation (Taylor 1992) found that moderators often disagreed on a candidate's mark by more than the tolerance (although the effects of the discrepancies on subject grades were generally no greater than in similar re-marking exercises on written papers). Second, because only a sample of work is inspected from most centres, the moderated marks are derived from an adjustment process under which some candidates receive a mark which is different from both the mark awarded by the centre and the mark awarded by the moderator.

A review of internal assessment published when statistical moderation was still in use (Cohen and Deale 1977) contains the following assertion.

[10] This indicates that the mean difference was -7.3, for a component with maximum mark 50.

> 'What does seem necessary is that boards which adopt statistical moderation procedures should make provision for further investigation in cases of doubt, such as may arise when small numbers in a particular school may mean that statistical methods are liable to error.'

'Further investigation' is intended to involve inspection of evidence of candidates' work. Such inspection can occur only if there is a permanent end-product. The revival of interest in statistical moderation at the present time arises because of the possibility of developing teacher assessment based on largely ephemeral evidence. Opportunities for further investigation of such assessments in cases of doubt would be limited.

It is recommended that any further work should involve closer inspection of the outcomes in a small number of specifications, perhaps investigating centre effects in more detail. However, using existing examinations and assessments for this work may give rise to unduly pessimistic outcomes. It is possible that diversion of resources from moderation by inspection to training of teachers would improve the accuracy of marking, and the reliability of statistical moderation might be improved if the general levels of performance on written papers and coursework were to become more balanced, either by encouraging centres to devote more attention to written papers at the expense of coursework or simply by manipulating marking schemes.

7.2 Teacher assessment

Again, the findings were disappointing, mainly because the effects of replacing operational marks by centres' estimated grades (used as proxy for teacher assessments) were unpredictable. For example, why did the number of candidates obtaining grade A *increase* in Law when teacher assessment was used for Units 1-5 but *decrease* in Business Studies (see Figures 6.1a and 6.2a)? Moreover, when teacher assessment was used for Units 3 and 4 in English Literature B, the numbers of candidates at the middle grades decreased, in contrast to the pattern in nearly all other cases investigated. Perhaps these variations are associated with differences in the accuracy of centres' estimated grades across subjects. As reported earlier, this study did not seek to investigate the accuracy of estimated grades. However, Business Studies, Geography A and ICT were considered in passing, and it was found that correlations between estimated grades and actual grades for units were generally low, ranging from 0.41 to 0.49 in Business Studies and from 0.48 to 0.56 in Geography A. In ICT, they varied considerably across units, ranging from 0.35 to 0.74. Most of the values are lower than those reported by Dhillon (2005) for estimated grades versus actual grades at subject level. She investigated six specifications and found correlations ranging from 0.77 to 0.85 at A level and from 0.64 to 0.81at AS.

The present study may be hampered by attempting to use existing data for a purpose for which these data are not intended. For example, Dhillon speculates on the motivation behind teachers' thinking when determining estimated grades. She reports allegations that teachers in Psychology

artificially deflate their estimates in order to improve value-added indicators of their teaching. Although no evidence of this practice was found, most teachers are aware that their estimated grades do not affect their candidates' results and, in the light of this knowledge, some may manipulate the estimates for a variety of purposes.

It was seen that the changes in candidates' subject grades caused by using teacher assessment are often large (up to three or four grades for some candidates), even when the *net* changes in the numbers of candidates per grade are small. Examples where changes in subject grade were less dramatic involved using teacher assessment for fewer units (so that it unsurprisingly had a smaller effect).

Statistical moderation of teacher assessments merely increased the numbers of changes which occur. For example, in Figures 6.1a and Figure 6.5.1a, the number of candidates with grade A under teacher assessment, already lower than the operational number, was further reduced by statistical moderation. A similar effect is apparent in most of the other examples investigated. The use of an 'allowed difference' in statistical moderation (instead of simply adjusting the centre's mean teacher assessment to be equal to the centre's mean mark for the moderating instrument) would have avoided the downward shift in grades but the results from Part 1 of the study suggest that many candidates would nevertheless have changed subject grade.

Again, it is recommended that any further work should involve more detailed investigation of a small number of cases, including the effects of choosing different units within a specification for replacement of the operational marks by teacher assessments. The analysis could be refined to take account of the weightings of the units when aggregating the teacher assessments, although it is unlikely that this refinement would have any appreciable effect, as all weightings are within a fairly narrow range (15% - 20% of the total A level assessment). In addition, higher quality data could perhaps be obtained by asking a sample of centres to provide teacher assessments for certain units specifically for research purposes (so that there would be no reason for teachers to manipulate the information provided) on a broader scale than just A-E and U.

Martin Taylor
30 August 2005

REFERENCES

Birch G. (1991) *Investigation of the use of statistical moderation methods for internally assessed components in SEG GCSE Mode 1 examinations* Research Division, University of Oxford Delegacy of Local Examinations

Cohen L. and Deale R.N. (1977) *Schools Council Examinations Bulletin 37: Assessment by teachers in examinations at 16+* Evans / Methuen Educational

Delap M.R. (1995) *Teachers' estimates of candidates' performances in public examinations* Assessment in Education Vol 2 No 1

Dhillon D. (2005) *Teachers' estimates of candidates' grades: Curriculum 2000 Advanced Level qualifications* British Educational Research Journal Vol 31 No 1

Kingdon J.M. (1980) *Statistical moderation – a mirage?* University of London School Examinations Department

Taylor M. (1992) *The reliability of judgements made by coursework assessors* Internal AEB research paper

Victorian Curriculum and Assessment Authority website www.vcaa.vic.edu.au/vce/exams/statisticalmoderation/statmod.html *Statistical moderation of VCE coursework* accessed 8 February 2005

Wilmut J. and Tuson J. (2004) *Statistical moderation of teacher assessments: a report to the Qualifications and Curriculum Authority* Centre for Developing and Evaluating Lifelong Learning, University of Nottingham

Working Group on 14-19 Reform (2004) *14-19 Curriculum and Qualifications Reform* DfES

APPENDIX A

SUMMARY STATISTICS FOR STATISTICAL MODERATION IN FOUR GCSE SPECIFICATIONS

The specifications included in this appendix are GCSE History Specification A, GCSE History Specification B, GCSE Humanities and GCSE Religious Studies Specification A

GCSE History A: no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	12355			
Maximum mark for coursework	50			
Weighting (%)	25%			
Tolerance	3 marks			
Mean difference	-5.3	-5.4	-5.4	-5.3
Mode	-8	-8	-5	-7
Standard deviation	4.3	5.1	6.0	5.4

GCSE History A: with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	12355			
Maximum mark for coursework	50			
Weighting (%)	25%			
Tolerance	3 marks			
Mean difference	-5.0	-5.1	-5.1	-5.0
Mode	0	0	0	0
Standard deviation	4.5	5.2	5.8	5.3

GCSE History A: with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	12355			
Maximum mark for coursework	50			
Weighting (%)	25%			
Tolerance	3 marks			
Mean difference	0.6	0.6	0.6	0.6
Mode	0	0	0	0
Standard deviation	4.0	4.3	4.7	4.5

GCSE History B: no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	44360			
Maximum mark for coursework	50			
Weighting (%)	25%			
Tolerance	3 marks			
Mean difference	-7.3	-7.4	-7.4	-7.3
Mode	-7	-6	-7	-7
Standard deviation	3.8	4.2	4.9	4.5

GCSE History B: with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	44360			
Maximum mark for coursework	50			
Weighting (%)	25%			
Tolerance	3 marks			
Mean difference	-7.2	-7.2	-7.2	-7.2
Mode	-6	-6	0	0
Standard deviation	4.1	4.4	4.9	4.6

GCSE History B: with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	44360			
Maximum mark for coursework	50			
Weighting (%)	25%			
Tolerance	3 marks			
Mean difference	0.2	0.2	0.2	0.2
Mode	0	0	0	0
Standard deviation	3.5	3.7	4.0	3.9

GCSE Humanities: no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	12548			
Maximum mark for coursework	60			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	-3.8	-3.8	-3.9	-3.8
Mode	-1	-6	-7	-3
Standard deviation	5.0	6.2	8.1	6.5

GCSE Humanities: with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	12548			
Maximum mark for coursework	60			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	-3.2	-3.2	-3.2	-3.2
Mode	0	0	0	0
Standard deviation	5.0	5.7	6.8	5.9

GCSE Humanities: with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	12548			
Maximum mark for coursework	60			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	0.4	0.3	0.3	0.4
Mode	0	0	0	0
Standard deviation	4.6	5.1	5.9	5.3

GCSE Religious Studies A: no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	18784			
Maximum mark for coursework	83			
Weighting (%)	20%			
Tolerance	5 marks			
Mean difference	-8.0	--8.0	-8.0	-8.0
Mode	-6	-7	-6	-7
Standard deviation	6.5	7.1	8.0	7.6

GCSE Religious Studies A: with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	18784			
Maximum mark for coursework	83			
Weighting (%)	20%			
Tolerance	5 marks			
Mean difference	-7.5	-7.5	-7.6	-7.5
Mode	0	0	0	0
Standard deviation	6.9	7.3	8.0	7.7

GCSE Religious Studies A: with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	18784			
Maximum mark for coursework	83			
Weighting (%)	20%			
Tolerance	5 marks			
Mean difference	0.7	0.7	0.7	0.7
Mode	0	0	0	0
Standard deviation	6.2	6.4	6.9	6.7

APPENDIX B

SUMMARY STATISTICS FOR STATISTICAL MODERATION IN GCSE BUSINESS STUDIES AND DESIGN & TECHNOLOGY

These are tiered specifications

GCSE Business Studies A Foundation tier: no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	8926			
Maximum mark for coursework	63			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	-3.9	-3.9	-3.9	-3.9
Mode	-4	-2	-2	-2
Standard deviation	5.8	7.0	9.3	7.4

GCSE Business Studies A Foundation tier: with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	8926			
Maximum mark for coursework	63			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	-3.5	-3.5	-3.5	-3.5
Mode	0	0	0	0
Standard deviation	5.8	6.6	8.0	6.8

GCSE Business Studies A Foundation tier: with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	8926			
Maximum mark for coursework	63			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	0.4	0.4	0.4	0.4
Mode	0	0	0	0
Standard deviation	5.5	6.0	7.1	6.2

GCSE Business Studies A Higher tier: no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	10000			
Maximum mark for coursework	63			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	-12.4	-12.4	-12.4	-12.3
Mode	-13	-13	-14	-16
Standard deviation	6.4	7.4	9.2	7.8

GCSE Business Studies A Higher tier: with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	10000			
Maximum mark for coursework	63			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	-12.2	-12.3	-12.3	-12.2
Mode	-13	0	0	0
Standard deviation	6.6	7.5	9.1	7.8

GCSE Business Studies A Higher tier: with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	10000			
Maximum mark for coursework	63			
Weighting (%)	25%			
Tolerance	4 marks			
Mean difference	0.5	0.6	0.6	0.6
Mode	0	0	0	0
Standard deviation	6.0	6.7	7.7	6.9

GCSE Design & Technology (Food) Foundation tier: no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	37765			
Maximum mark for coursework	95			
Weighting (%)	60%			
Tolerance	6 marks			
Mean difference	-12.4	-12.4	-12.4	-12.4
Mode	-14	-13	-18	-12
Standard deviation	8.4	9.6	12.5	10.3

GCSE Design & Technology (Food) Foundation tier: with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	37765			
Maximum mark for coursework	95			
Weighting (%)	60%			
Tolerance	6 marks			
Mean difference	-11.9	-12.0	-12.0	-11.9
Mode	0	0	0	0
Standard deviation	8.9	9.7	12.0	10.2

GCSE Design & Technology (Food) Foundation tier: with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	37765			
Maximum mark for coursework	95			
Weighting (%)	60%			
Tolerance	6 marks			
Mean difference	0.5	0.5	0.5	0.5
Mode	0	0	0	0
Standard deviation	8.0	8.5	10.0	8.9

GCSE Design & Technology (Food) Higher tier: no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	30190			
Maximum mark for coursework	95			
Weighting (%)	60%			
Tolerance	6 marks			
Mean difference	-14.4	-14.4	-14.4	-14.4
Mode	-15	-15	-17	-15
Standard deviation	6.8	7.5	9.1	8.0

GCSE Design & Technology (Food) Higher tier: with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	30190			
Maximum mark for coursework	95			
Weighting (%)	60%			
Tolerance	6 marks			
Mean difference	-14.1	-14.1	-14.1	-14.0
Mode	0	0	0	0
Standard deviation	7.4	7.9	9.2	8.3

GCSE Design & Technology (Food) Higher tier: with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	30190			
Maximum mark for coursework	95			
Weighting (%)	60%			
Tolerance	6 marks			
Mean difference	0.3	0.4	0.5	0.6
Mode	0	0	0	0
Standard deviation	6.1	6.5	7.2	6.9

APPENDIX C

SUMMARY STATISTICS FOR STATISTICAL MODERATION IN GCE SPECIFICATIONS

The specifications included in this appendix are GCE Biology Specification A (AS), GCE Business Studies, GCE French, GCE Geography Specification A, GCE Geography Specification B and GCE Psychology Specification A.

GCE Biology A (AS): no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	13413			
Max. mark for coursework unit	30			
Weighting (%)	30% of the total AS assessment			
Tolerance	2 marks			
Mean difference	-9.0	-8.9	-8.9	8.9
Mode	-9	-9	-9	-9
Standard deviation	3.0	4.0	3.2	4.2

GCE Biology A (AS): with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	13413			
Max. mark for coursework unit	30			
Weighting (%)	30% of the total AS assessment			
Tolerance	2 marks			
Mean difference	-8.9	-8.8	-8.9	-8.9
Mode	-9	-9	-9	-9
Standard deviation	3.1	4.1	3.3	4.3

GCE Biology A (AS): with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	13413			
Max. mark for coursework unit	30			
Weighting (%)	30% of the total AS assessment			
Tolerance	2 marks			
Mean difference	-0.1	-0.2	-0.1	0.2
Mode	0	0	0	0
Standard deviation	2.5	2.7	2.6	3.3

GCE Business Studies: no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	7323			
Max. mark for coursework unit	84			
Weighting (%)	15% of the total A level assessment			
Tolerance	6 marks			
Mean difference	0.0	0.1	0.1	0.1
Mode	1	-1	0	0
Standard deviation	6.9	7.4	8.8	8.0

GCE Business Studies: with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	7323			
Max. mark for coursework unit	84			
Weighting (%)	15% of the total A level assessment			
Tolerance	6 marks			
Mean difference	0.5	0.6	0.6	0.6
Mode	0	0	0	0
Standard deviation	6.7	6.9	7.2	7.1

GCE Business Studies: with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	7323			
Max. mark for coursework unit	84			
Weighting (%)	15% of the total A level assessment			
Tolerance	6 marks			
Mean difference	2.6	2.6	2.6	2.6
Mode	0	0	0	0
Standard deviation	6.6	6.8	7.2	7.0

GCE French: no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	1395			
Max. mark for coursework unit	60			
Weighting (%)	15% of the total A level assessment			
Tolerance	4 marks			
Mean difference	-0.3	-0.1	0.0	-0.2
Mode	1	0	0	-2
Standard deviation	6.0	6.4	6.6	6.8

GCE French: with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	1395			
Max. mark for coursework unit	60			
Weighting (%)	15% of the total A level assessment			
Tolerance	4 marks			
Mean difference	0.0	0.2	0.2	0.0
Mode	0	0	0	0
Standard deviation	6.0	6.1	6.2	6.4

GCE French: with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	1395			
Max. mark for coursework unit	60			
Weighting (%)	15% of the total A level assessment			
Tolerance	4 marks			
Mean difference	1.8	1.9	1.9	1.8
Mode	0	0	0	0
Standard deviation	5.9	5.9	6.1	6.3

GCE Geography A: no tolerance, no allowed difference[11]

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	2699			
Max. mark for coursework unit	100			
Weighting (%)	20% of the total A level assessment			
Tolerance	(6 marks)			
Mean difference	-18.8	-18.5	-18.5	-18.6
Mode	-21	-15	-14	-19
Standard deviation	12.2	13.3	15.4	13.9

GCE Geography A: with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	2699			
Max. mark for coursework unit	100			
Weighting (%)	20% of the total A level assessment			
Tolerance	(6 marks)			
Mean difference	-18.7	-18.4	-18.4	-18.5
Mode	0	0	0	0
Standard deviation	12.3	13.2	15.1	13.8

GCE Geography A: with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	2699			
Max. mark for coursework unit	100			
Weighting (%)	20% of the total A level assessment			
Tolerance	(6 marks)			
Mean difference	-1.9	-1.6	-1.3	0.0
Mode	0	0	0	0
Standard deviation	12.3	12.9	14.0	13.0

[11] As noted in the main text (section 5.5), coursework in this specification is *externally*-assessed. Therefore, there should be no moderation adjustments at all. The tolerance recorded in the tables has been calculated on the same basis as for internally-assessed coursework.

GCE Geography B: no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	1082			
Max. mark for coursework unit	100			
Weighting (%)	15% of the total A level assessment			
Tolerance	6 marks			
Mean difference	-2.3	-2.4	-2.4	-2.2
Mode	0	1	-9	-2
Standard deviation	8.0	10.1	12.2	10.7

GCE Geography B: with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	1082			
Max. mark for coursework unit	100			
Weighting (%)	15% of the total A level assessment			
Tolerance	6 marks			
Mean difference	-1.8	-1.8	-1.8	-1.7
Mode	0	0	0	0
Standard deviation	7.9	9.0	10.1	9.2

GCE Geography B: with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	1082			
Max. mark for coursework unit	100			
Weighting (%)	15% of the total A level assessment			
Tolerance	6 marks			
Mean difference	-0.1	-0.2	-0.2	-0.1
Mode	0	0	0	0
Standard deviation	7.7	8.9	9.7	9.1

GCE Psychology A (moderating instrument consisting of all written units): no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	24054			
Max. mark for coursework unit	60			
Weighting (%)	15% of the total A level assessment			
Tolerance	4 marks			
Mean difference	-8.7	-8.6	-8.6	-8.6
Mode	-8	-9	-9	-7
Standard deviation	4.8	5.7	5.2	6.0

GCE Psychology A (moderating instrument consisting of all written units): with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	24054			
Max. mark for coursework unit	60			
Weighting (%)	15% of the total A level assessment			
Tolerance	4 marks			
Mean difference	-8.4	-8.4	-8.4	-8.4
Mode	0	0	0	0
Standard deviation	5.1	5.9	5.5	6.2

GCE Psychology A (moderating instrument consisting of all written units): with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	24054			
Max. mark for coursework unit	60			
Weighting (%)	15% of the total A level assessment			
Tolerance	4 marks			
Mean difference	0.8	0.8	0.8	0.9
Mode	0	0	0	0
Standard deviation	4.4	4.6	4.5	4.8

GCE Psychology A (moderating instrument consisting of Unit 5 only): no tolerance, no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	24054			
Max. mark for coursework unit	60			
Weighting (%)	15% of the total A level assessment			
Tolerance	4 marks			
Mean difference	-14.7	-14.5	-14.6	-14.6
Mode	-9	-16	-11	-18
Standard deviation	7.3	9.4	7.8	10.0

GCE Psychology A (moderating instrument consisting of Unit 5 only): with tolerance applied but no allowed difference

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	24054			
Max. mark for coursework unit	60			
Weighting (%)	15% of the total A level assessment			
Tolerance	4 marks			
Mean difference	-14.6	-14.4	-14.5	-14.5
Mode	-9	0	0	0
Standard deviation	7.4	9.5	7.9	10.0

GCE Psychology A (moderating instrument consisting of Unit 5 only): with tolerance and allowed difference applied

	Method (i)	Method (ii)	Method (iii)	Method (iv)
No. of candidates	24054			
Max. mark for coursework unit	60			
Weighting (%)	15% of the total A level assessment			
Tolerance	4 marks			
Mean difference	0.4	-0.1	0.4	0.8
Mode	0	0	0	0
Standard deviation	6.7	7.2	6.9	8.7

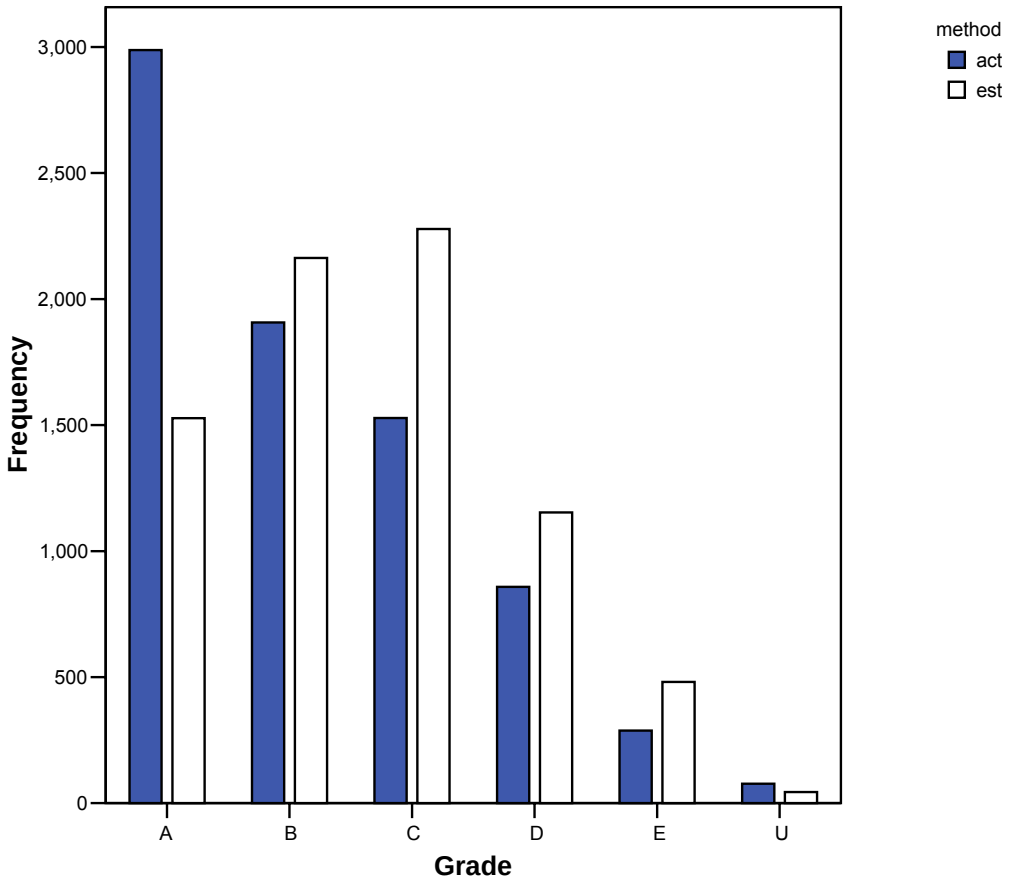
APPENDIX D

ACCURACY OF CENTRES' ESTIMATED GRADES PER UNIT IN GCE BUSINESS STUDIES

Correlations between actual grades and estimated grades were found to be similar for all units, ranging from 0.41 for Unit 5 to 0.49 for Unit 1.

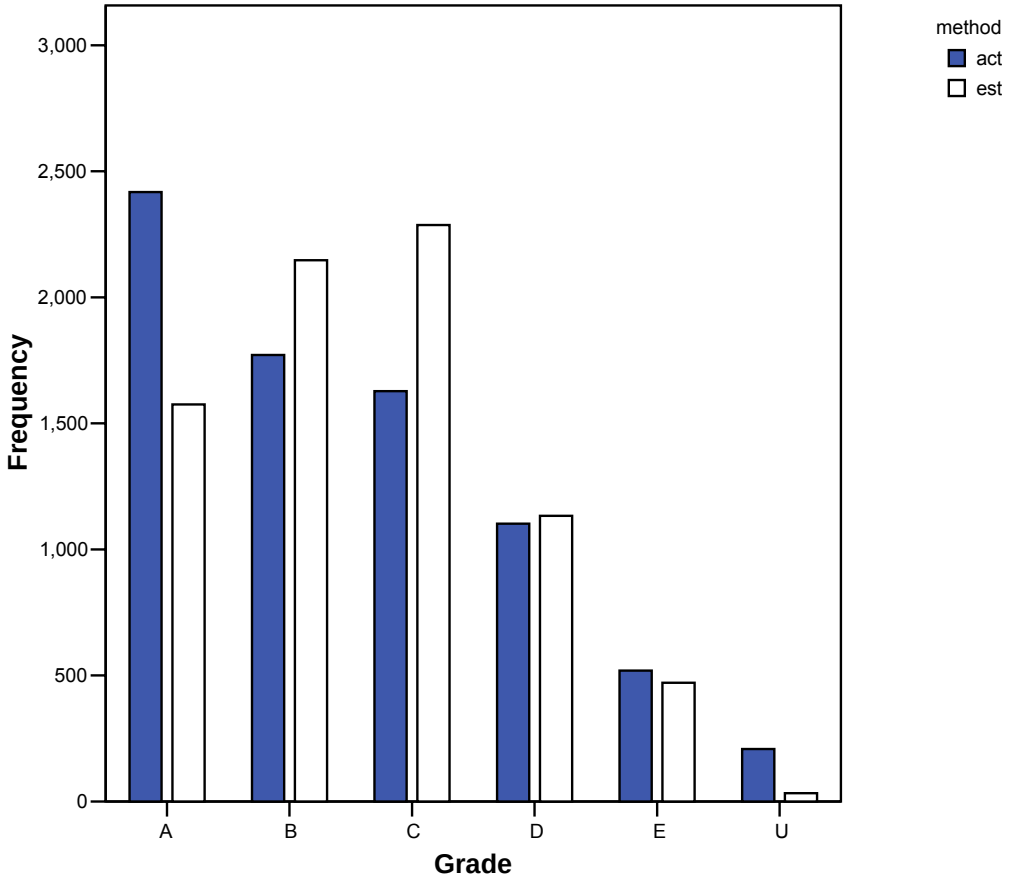
Actual grades and estimated grades for Unit 1

		Estimated grade						Total
		A	B	C	D	E	U	
Actual grade	A	1171	985	631	157	41	3	2988
	B	267	676	623	259	80	2	1907
	C	70	364	613	333	138	10	1528
	D	13	108	304	284	134	15	858
	E	5	25	81	104	66	7	288
	U	1	5	26	16	22	7	77
Total		1527	2163	2278	1153	481	44	7646



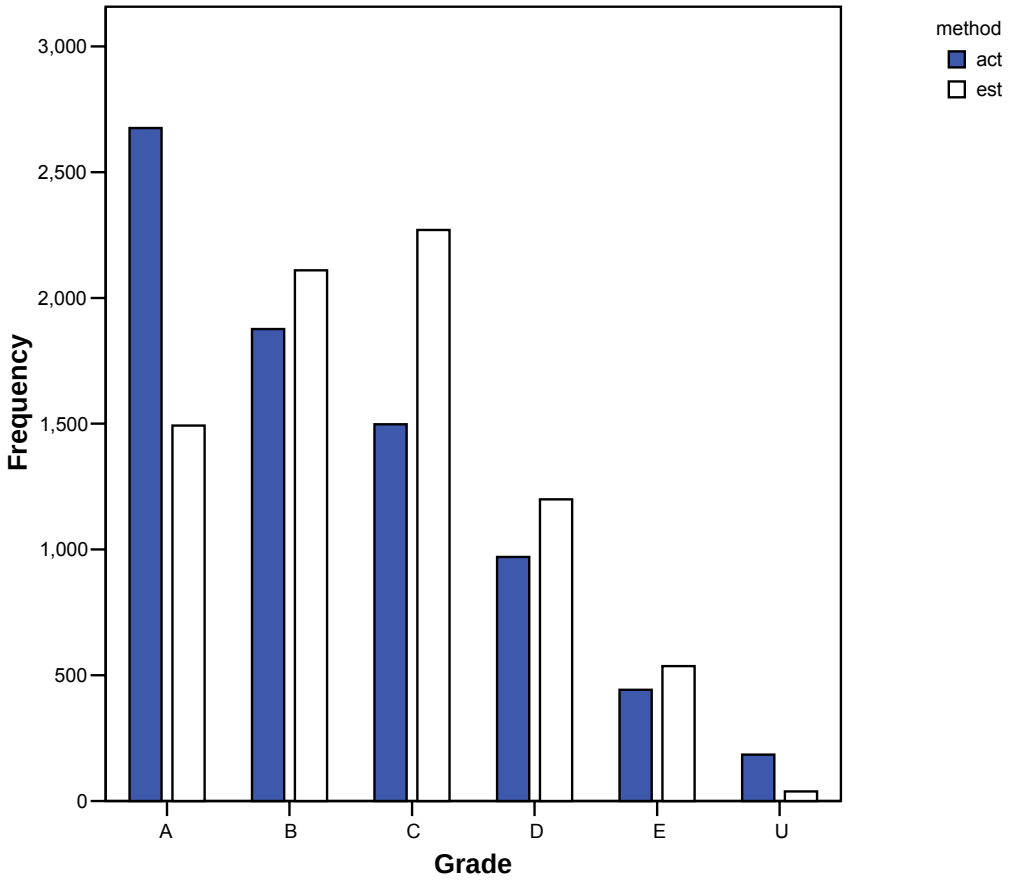
Actual grades and estimated grades for Unit 2

		Estimated Grade						Total
		A	B	C	D	E	U	
Actual Grade	A	1002	819	435	115	43	4	2418
	B	348	584	565	209	60	5	1771
	C	164	432	644	292	92	4	1628
	D	44	208	416	292	135	7	1102
	E	15	80	167	159	93	5	519
	U	2	24	60	66	48	8	208
Total		1575	2147	2287	1133	471	33	7646



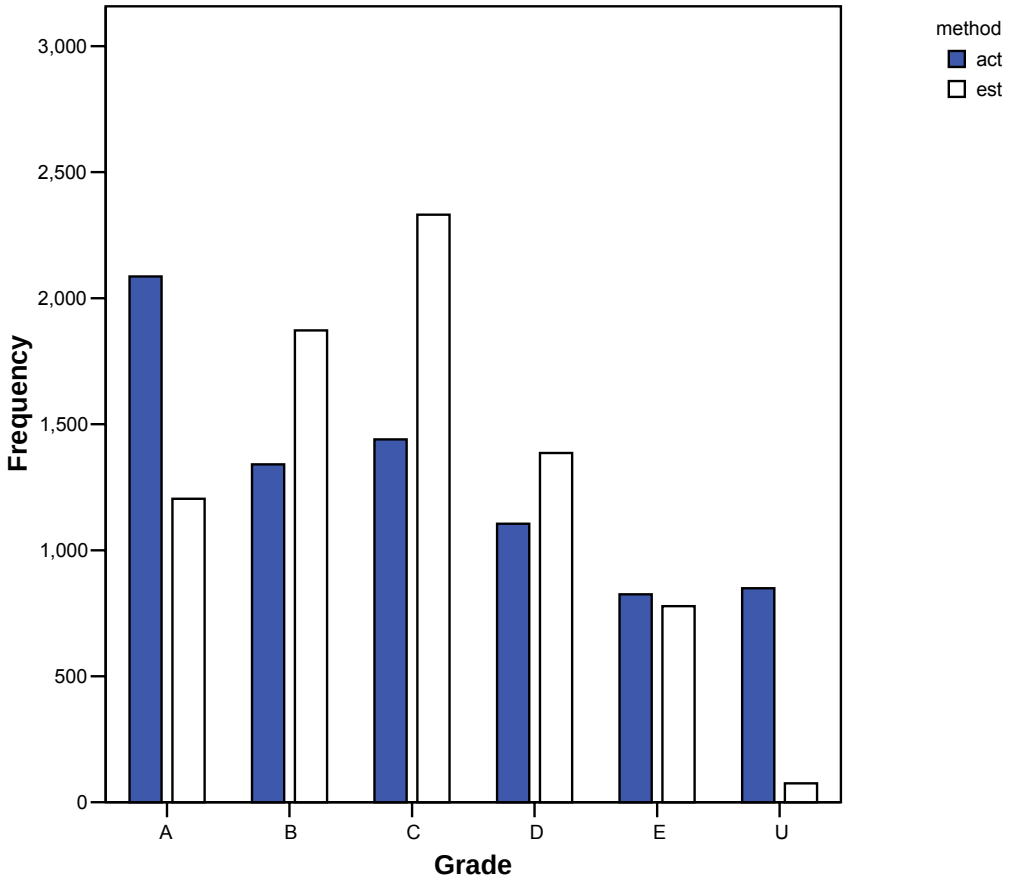
Actual grades and estimated grades for Unit 3

		Estimated grade						Total
		A	B	C	D	E	U	
Actual grade	A	1023	913	527	170	39	4	2676
	B	329	589	600	263	93	2	1876
	C	103	381	565	311	129	9	1498
	D	30	161	378	250	141	10	970
	E	7	46	148	150	85	6	442
	U	1	20	52	55	49	7	184
Total		1493	2110	2270	1199	536	38	7646



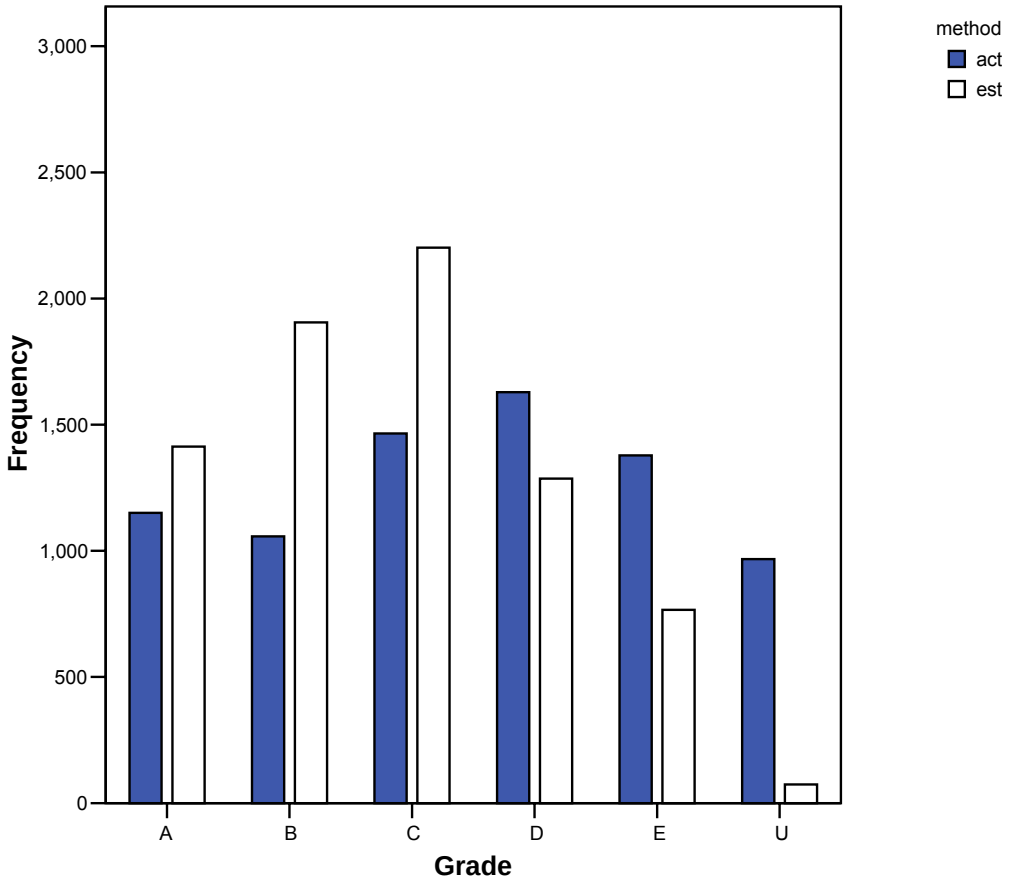
Actual grades and estimated grades for Unit 4

		Estimated grade						Total
		A	B	C	D	E	U	
Actual grade	A	757	704	460	132	31	2	2086
	B	231	424	446	176	61	3	1341
	C	127	348	543	305	112	5	1440
	D	52	218	401	263	159	12	1105
	E	27	100	262	264	161	11	825
	U	10	78	219	246	254	42	849
Total		1204	1872	2331	1386	778	75	7646



Actual grades and estimated grades for Unit 5

		Estimated Grade						Total
		A	B	C	D	E	U	
Actual Grade	A	495	364	197	65	28	1	1150
	B	281	334	286	118	37	1	1057
	C	301	447	430	199	83	5	1465
	D	211	382	569	287	165	15	1629
	E	94	270	446	347	202	19	1378
	U	31	108	274	270	251	33	967
Total		1413	1905	2202	1286	766	74	7646



Actual grades and estimated grades for Unit 6

		Estimated grade						Total
		A	B	C	D	E	U	
Actual grade	A	488	304	210	60	21	1	1084
	B	300	333	262	96	41	1	1033
	C	278	423	420	175	75	6	1377
	D	189	374	461	253	136	11	1424
	E	76	250	413	273	164	8	1184
	U	37	178	477	415	398	39	1544
Total		1368	1862	2243	1272	835	66	7646