differential subject difficulty make almost no difference to the ranking of schools. Overall there is a correlation of 0.998 between the original percentage of candidates achieving five A*-C grades and the estimated percentage after adjustments. Furthermore, there are only 8 schools (out of 2,928) where the difference exceeds 5 percentage points and none where it exceeds 10 percentage points. This again indicates that adjustments to grading to account for variations in subject difficulty are unlikely to have any substantial effect upon school performance measures.

**Reference**

Bramley, T. (2014) Multivariate representations of subject difficulty. *Research Matters: A Cambridge Assessment publication*, *18*, 42–47. Available online at: http://www.cambridgeassessment.org.uk/Images/174492-research-matters-18-summer-2014.pdf.

# Statistical moderation of school-based assessment in GCSEs

**Joanna Williamson** Research Division

## Introduction

School-based assessment (SBA) such as coursework is included in high-stakes qualifications around the world. In the United Kingdom (UK) for example, selected General Certificate of Secondary Education (GCSE) and General Certificate of Education (GCE) Advanced level (A level) examinations include SBA components[1] alongside examination components[2]. Moderation is required in order to address the question of comparability of SBA marks across different centres. Under current procedures for GCSEs and A levels (see Gill, 2015), moderators re-mark a sample of each centre's SBA work. The awarding body uses the relationship between the moderator mark and centre mark (in the re-marked sample) to decide what adjustment, if any, should be applied to that centre's SBA marks.

Statistical moderation is an alternative form of moderation that calibrates and/or monitors the marks of an assessment on the basis of a statistical relationship with another assessment. Its validity depends on the two assessments having a strong relationship in terms of both assessment content and candidate performance, but they need not measure precisely the same construct. In the context of SBA, the most common statistical moderation practice is to calibrate candidate marks on SBA component(s) using marks from the exam component(s) of the same overall assessment. The motivation for statistical moderation is to preserve information about candidates' SBA performance (such as their ranking within the centre) whilst acknowledging that marking may vary between centres. Statistical moderation removes the absolute meaning of SBA marks, and calibrates them to a new scale that is common to all candidates, that is, the exam component.

During recent reforms of GCSEs and A levels, the Office of Qualifications and Examinations Regulation (Ofqual) proposed the use of statistical moderation in GCSE assessment (Ofqual, 2015a). Previous research by Taylor (2005), using results data from the AQA awarding body, found that statistical moderation generally adjusted marks downward, since SBA marks for GSCE and A level were usually higher than exam marks. The study also found that many candidates would have been awarded different grades under statistical moderation, and that there was a disappointing "absence of any pattern, across different specifications" in terms of statistical moderation outcomes (Taylor, 2005, p.51). The present article outlines methods of statistical moderation that are used in jurisdictions around the world, and explores the effect of applying these methods to results data from three Oxford, Cambridge and RSA Examinations (OCR) GCSEs. This involved statistically moderating all SBA components, aggregating SBA marks with exam marks, and then calculating candidates' statistically moderated final grades from these aggregate scores. Analysis focuses on comparing the statistically moderated results to operational results (moderated under existing, non-statistical procedures) in terms of marks, grades, and the rank-order of candidates and centres.

## Methods of statistical moderation

Statistical moderation is a form of assessment linking, where "the goal is to put scores from two or more tests on the same scale – *in some sense*." (Kolen & Brennan, 2004, p.423). Given a suitable pair of assessments (e.g., SBA unit and exam unit), there exist multiple ways to statistically moderate. Table 1 shows the methods investigated in this article: the first four methods are variations of linear scaling, the next two are forms of curvilinear scaling and the final method is rank mapping. Of these, the most commonly used method is linear scaling that matches the mean and standard deviation (SD) of SBA marks within each centre to those of the exam marks (Method 2). The three simplest linear methods (1, 2 and 4) and rank mapping (Method 7) were previously investigated by Taylor (2005). Despite different statistical procedures, many of the methods share common outcomes, as summarised in Table 2.

1. Recent qualification reforms have reduced the use of SBA in GCSE assessment (Ofqual, 2015b). Of the 23 'new' GCSEs (9–1) ready for first teaching in September 2015 or 2016, 7 contain SBA components.
2. In GCSE and A level, examination components are always externally set and assessed. They are usually written exams.

**Table 1: Methods of statistical moderation**

| | *Description* | *Moderation formula* | *Examples of use* | *Advantages* | *Criticisms* |
|---|---|---|---|---|---|
| 1 | Adjusts SBA mean to match exam mean | $y_i = \overline{z} + (x_i - \overline{x})$ | South Africa | Transparency<br>Few parameters to estimate | Out-of-range marks<br>Potentially unfair when mark distributions skewed |
| 2 | Adjusts SBA mean and SD to match exam mean and SD | $y_i = \overline{z} + \frac{\sigma_z}{\sigma_x}(x_i - \overline{x})$ | West African Senior School Certificate<br>Western Australia Certificate of Education | Transparency<br>Few parameters to estimate | Out-of-range marks<br>Potentially unfair when mark distributions skewed<br>'Company you keep' factor unacceptably high |
| 3 | Adjusts SBA mean and SD, taking into account both inter- and intra- group differences | $y_i = x_{mean} + \frac{s_y}{\sigma_x}(x_i - \overline{x})$<br>$+\beta(\overline{z} - z_{mean})$ | Hong Kong Diploma of Secondary Education | Allows for global difference in SBA and exam performance | Low transparency |
| 4 | Adjusts SBA marks based on regression of exam marks onto SBA marks | $y_i = \overline{z} + \frac{\sigma_z}{\sigma_x}(x_i - \overline{x}) \cdot r$ | | | Moderated marks are 'compressed' about the mean<br>Potentially unfair when mark distributions skewed |
| 5 | Quadratic polynomial mapping, fixing max, mean and min SBA marks onto max, mean and min exam marks | $y_i = ax_i^2 + bx_i + c$ | New South Wales High School Certificate | Copes with differences in SBA/exam mark distributions<br>High-attaining candidates protected<br>No out-of-range marks | Low transparency<br>Does not preserve ratio between pairs of marks |
| 6 | Simplified equipercentile mapping, with linear interpolation | $mod(x_{max}) = z_{max}$,<br>$mod(x_{Q3}) = z_{Q3}$,<br>$mod(x_{Q2}) = z_{Q2}$,<br>$mod(x_{Q1}) = z_{Q1}$,<br>$mod(x_{min}) = z_{min}$ | Victorian Certificate of Education (Australia) | Copes with differences in SBA/exam mark distributions<br>Mark intervals somewhat preserved | Vulnerable to effects of individual marks<br>Low transparency<br>Unsuitable for small groups |
| 7 | Maps SBA marks to equivalently-ranked exam marks | $mod(x_{rank\ n}) = z_{rank\ n}$ | | | 'Company you keep' factor unacceptably high<br>Mark intervals not preserved |

- $mod(x)$ is the statistically moderated mark corresponding to raw mark $x$;
- $r$ is the within-centre correlation coefficient of SBA and exam marks;
- $s_y = \sqrt{w_x\sigma_x^2 + w_z\sigma_z^2}$, where $w_x$ and $w_z$ are weightings such that $w_x + w_z = 1$;
- $\beta$ is the (pooled) slope after regressing raw SBA marks onto exam marks in a two-level random intercept model;
- $x_i$, $y_i$, and $z_i$ are the $i^{th}$ candidate's raw SBA mark, moderated SBA mark and exam mark respectively;
- $x_{mean}$, $\overline{x}$ and $\sigma_x$ are the global mean, centre mean and centre SD of raw SBA marks;
- $y_{mean}$, $\overline{y}$ and $\sigma_y$ are the global mean, centre mean and centre SD of moderated SBA marks;
- $z_{mean}$, $\overline{z}$ and $\sigma_z$ are the global mean, centre mean and centre SD of exam marks; and
- The formulae to calculate coefficients $a, b, c$ (Method 5) are given by MacCann (1996).

**Table 2: Statistical moderation outcomes**

| *Aspect of statistical moderation* | *1* | *2* | *3* | *4* | *5* | *6* | *7* |
|---|---|---|---|---|---|---|---|
| Are moderated SBA marks distributed about centres' mean or median exam marks? | Y | Y | N | Y | Y | Y | Y |
| Do moderated SBA marks have the same mean as the centre's exam marks? | Y | Y | N | Y | Y | N | Y |
| Is a global difference between SBA and exam marks allowed for? | N | N | Y | N | N | N | N |
| Can a centre be comparatively 'better at coursework than exams' than other centres? | N | N | N | N | N | N | N |
| Do moderated marks ever fall out of range? | Y | Y | Y | Y | N | N | N |
| Is the within-centre rank order of candidates, by SBA mark, preserved? | Y | Y | Y | Y | Y | Y | Y |
| Are the intervals between candidate SBA marks preserved? | Y | Y | Y | Y | N | N | N |
| Is the within-centre rank order of candidates, by aggregated mark, preserved? | N | N | N | N | N | N | N |
| Is the rank order of centres, by mean aggregated mark, preserved? | N | N | N | N | N | N | N |

For Methods 1, 2 and 4, there exists a variant form that allows for a global difference between the level of SBA marks and exam marks, as Method 3 does already. The 'allowed difference' variant adjusts marks so that each centre's mean moderated SBA mark differs from its mean exam mark by an 'allowed difference', defined as the difference between the global SBA mark mean and global exam mark mean. To achieve this, occurrences of the centre mean exam mark ($\overline{z}$) in the mark adjustment formulae are replaced by the centre mean exam mark plus allowed difference ($\overline{z} + (x_{mean} - z_{mean})$).

For all methods of statistical moderation, the perceived fairness (and acceptability to stakeholders) is affected by validity, transparency, and assessment context, as suggested by the advantages and criticisms noted in Table 1. The list that follows on page 32 expands upon some particularly important factors:

### Anomalous or 'flop' scores

Anomalous scores can distort the mark adjustment deemed necessary for a moderation group. The difficulty is that it is impossible to be sure that a score is 'anomalous' since authentic differences in SBA and exam performance may occur for many reasons. This issue is particularly critical for methods, such as Method 6, that are highly sensitive to individual marks.

### Small moderation groups

The smaller the moderation group, the greater the risk of a misleading score distribution which can lead to unfair adjustments to candidate marks. Small moderation groups may necessitate adaptations to statistical procedures and/or manual intervention. As well as increasing cost and complexity, this can harm perceived fairness since different processes are applied to different centres and candidates.

### Transparency

It is usually considered important that the statistical procedures leading to moderated marks are transparent to stakeholders. A difficulty is that steps to address other concerns, such as validity, often result in more sophisticated statistical procedures (e.g., Method 3) that are less transparent.

### 'Company you keep' factor

Under statistical moderation, candidate marks are "inevitably affected" by the performance of others in their moderation group (Wilmut & Tuson, 2005, p.52). The degree to which this occurs is difficult to quantify, but a high degree is perceived as very unfair. Methods 1 and 2 are criticised on the basis that results are too strongly influenced by the moderation group. As an example, Table 3 and Figure 1 show a group of 12 candidates statistically moderated by Method 2. Two cases are shown: (1) where all candidates complete the qualification, and (2) where candidates 1–3 do not complete the qualification. The SBA and exam marks of the other candidates (4–12) remain the same, but their moderated marks differ substantially depending on whether the three lowest-attaining candidates complete the qualification or not.

### Disadvantaging particular candidates

There are concerns whenever statistical moderation appears to affect some candidates differently. Substantial changes to the relative intervals between pairs of candidate marks are perceived as unfair, for example, since it is difficult to justify candidates with very similar raw SBA marks receiving very different moderated marks. Truncation of marks (after statistical moderation results in marks out of range) also results in

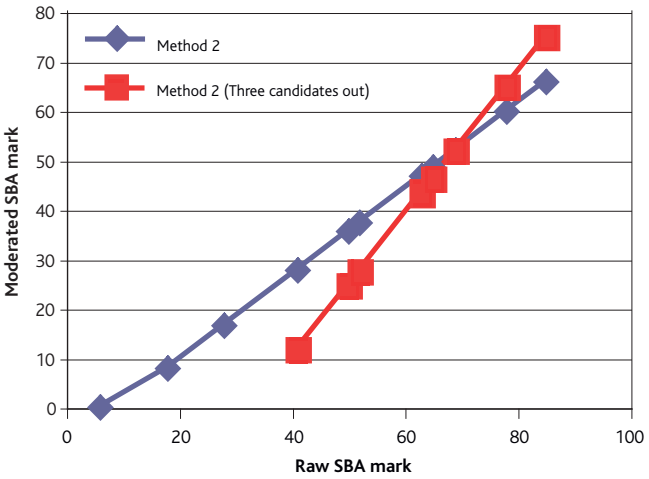


**Figure 1: Illustration of 'Company you keep' factor**

problematic results: loss of discrimination between candidates, marks of zero for valid SBA efforts, and different effective mark reductions for different candidates.

Another area of concern is large downward mark adjustments, which are perceived as especially unfair for high-attaining candidates in competitive contexts. Stanley, MacCann, Gardner, Reynolds, and Wild (2009, p.54) note that moderation using a linear method "often fails to work satisfactorily" due to negatively skewed SBA mark distributions that result from teachers "setting assessment tasks at which the students can excel" or from overly generous marking applied unevenly across the mark

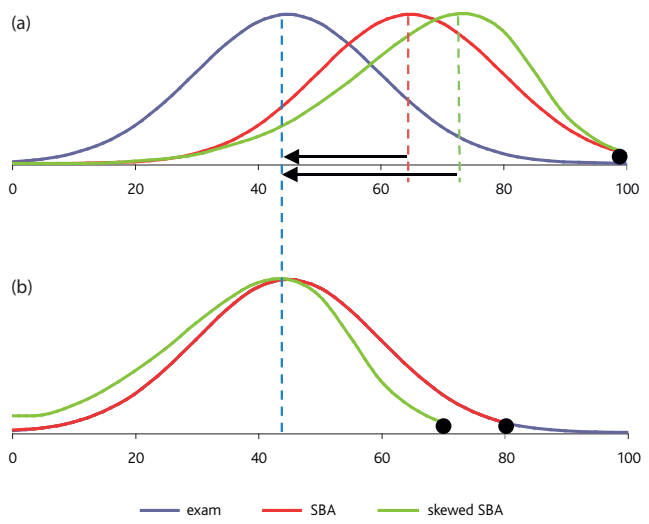


**Figure 2: Example mark distributions (a) before moderation and (b) after moderation (dotted lines indicate the mean)**

**Table 3: Example candidate data, moderated by Method 2**

| | | *Candidate No.:* | | | | | | | | | | | | *Mean* | *SD* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *1* | *2* | *3* | *4* | *5* | *6* | *7* | *8* | *9* | *10* | *11* | *12* | | |
| (1) | Raw SBA mark | 6 | 18 | 28 | 41 | 50 | 52 | 63 | 65 | 65 | 69 | 78 | 85 | *52* | *23.2* |
| | Exam mark | 9 | 28 | 15 | 10 | 32 | 38 | 23 | 70 | 51 | 45 | 65 | 58 | *37* | *20.2* |
| | Moderated SBA mark | 0 | 8 | 16 | 28 | 36 | 37 | 47 | 49 | 49 | 52 | 60 | 66 | | |
| (2) | Raw SBA mark | - | - | - | 41 | 50 | 52 | 63 | 65 | 65 | 69 | 78 | 85 | *63* | *13.0* |
| | Exam mark | - | - | - | 10 | 32 | 38 | 23 | 70 | 51 | 45 | 65 | 58 | *44* | *18.7* |
| | Moderated SBA mark | - | - | - | 12 | 25 | 28 | 43 | 46 | 46 | 52 | 65 | 75 | | |

range. The SBA mark distribution in these cases will have an 'inflated' centre mean, and the mark adjustment resulting from Methods 1 and 2 will also therefore be inflated (Figure 2a). High-attaining candidates from such a centre will consequently receive lower moderated marks than high-attaining candidates from a centre with a non-skewed distribution (Figure 2b). If the centre's downward mark adjustment was compensating for SBA mark inflation that the higher-attaining candidates within the centre did not benefit from, this appears unfair. A related aspect of perceived fairness is the difficulty in justifying a large mark reduction applied to an 'almost perfect' mark, compared with applying the same reduction to a low- or mid-level mark.

If the standard deviation of a centre's exam marks is lower than the standard deviation of the SBA marks, then Method 2 will also compress SBA marks towards the mean. Where the overall mark adjustment is downward, higher-attaining candidates will therefore receive a larger mark reduction than lower-attaining candidates. This effect is not unique to Method 2, but is mentioned here since it can exacerbate the problem of large mark reductions for high-attaining candidates.

To illustrate the effect of the different methods, Figure 4 and Figure 5, show the effects of statistically moderating an SBA unit for one centre. Twenty-nine candidates took GCSE X at this centre in June 2015. Their raw SBA marks (mean 67.4) and exam marks (mean 39.7) are plotted in Figure 3.

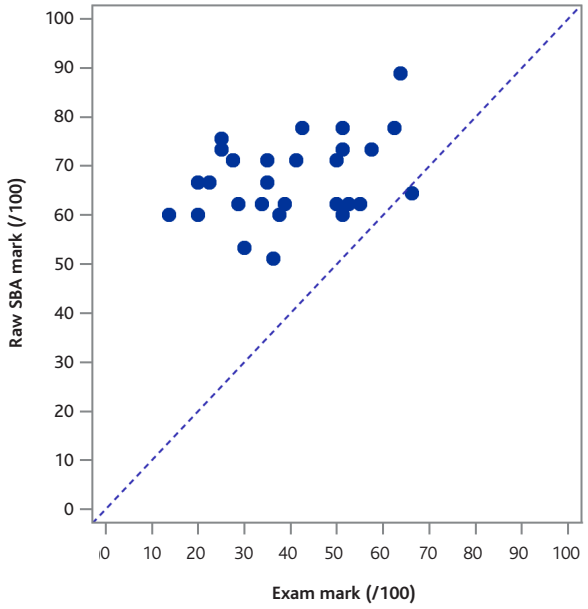


**Figure 3: Raw SBA marks against exam marks (r=0.33)**

Figure 4 plots the moderated SBA marks resulting from each method of statistical moderation against the candidates' raw SBA marks, and Figure 5 compares the mark distributions resulting from each method. Method 3, the only method not to distribute moderated SBA marks about the mean or median exam mark, is clearly differentiated from the other methods. The highly reduced spread of marks resulting from Method 4 is also very noticeable.

## Method

The statistical moderation methods described in Table 1, plus the allowed difference variants, were applied to June 2015 results data from three

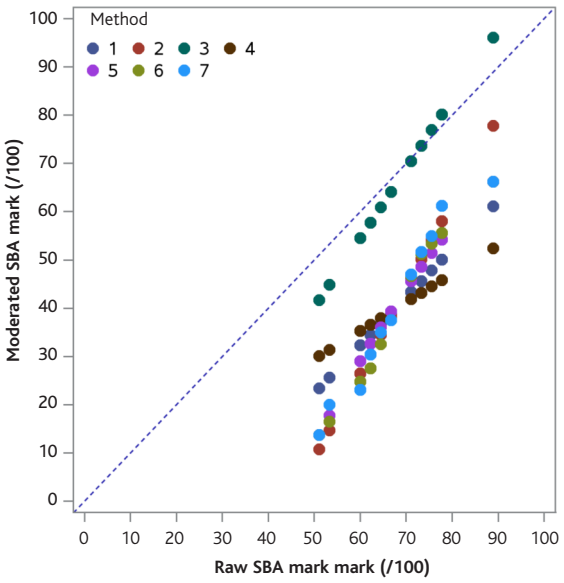


**Figure 4: Moderated SBA marks against raw SBA marks**

OCR GCSEs. Centres with fewer than six candidates[3] were excluded, as were centres where SBA and exam marks had zero or negative correlation[4]. For each specification, marks were first converted onto a scale of 0–100, and then all SBA components were statistically moderated by the corresponding exam unit (or a linear combination of the exam units if the specification had multiple). Statistically moderated SBA marks were truncated to the allowed mark range (if outside this), rounded to the nearest whole number, and combined with the exam marks using the weightings implied by Uniform Mark Scale (UMS) allocations. From these aggregated marks, a statistically moderated final grade was calculated for each candidate. New grade boundaries were calculated for each method, such that each statistically moderated grade distribution matched that of June 2015. The present study differs in this respect from that of Taylor (2005), which calculated statistically moderated grades using operational grade boundaries.

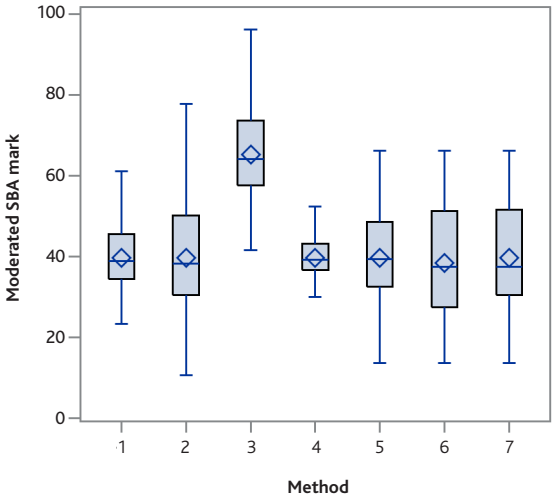


**Figure 5: Statistically moderated SBA mark distributions**

3. The smallest definition of acceptable group size found in the literature.
4. This excluded 1.6% of GCSE X candidates, 3.9% of GCSE Y candidates, and 4.1% of GCSE Z candidates.

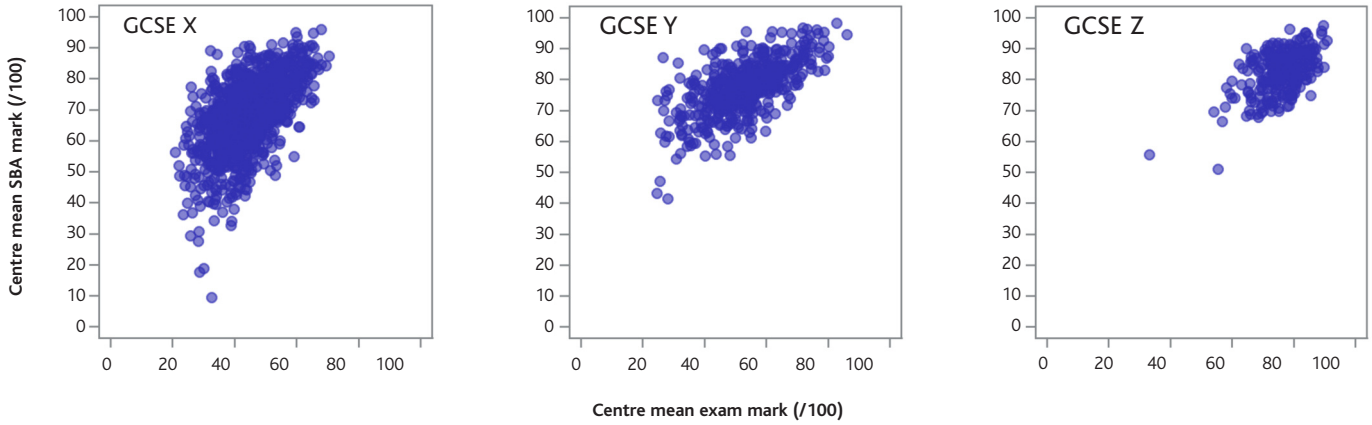


**Figure 6: Scatter plots of mean SBA marks against mean exam marks, by specification**

## Data

The chosen GCSE specifications each had at least one exam unit and at least one SBA unit, had overlap between the assessment objectives of SBA unit(s) and exam unit(s), and were awarded to at least 8,000 candidates in June 2015. Table 4 summarises the characteristics of component units for each specification. Only one SBA unit is shown per specification, since for the specifications with additional SBA units, the characteristics of additional units differed very little from those shown.

**Table 4: Summary of component characteristics (0–100 mark scales)**

| | *GCSE X* | *GCSE Y* | *GCSE Z* |
|---|---|---|---|
| Difference between raw SBA mean and exam mean | -24.2 marks | -24.7 marks | -6.9 marks |
| Variability of SBA and exam mark difference[5] | 9.6 marks | 9.1 marks | 6.7 marks |
| Level of spread in SBA marks compared with exam marks | SD ~3 marks higher | SD ~3 marks lower | SD ~2 marks lower |
| Shape of mark distributions | Strong negative skew (SBA) vs. negligible skew (exam) | Strong negative skew (SBA) vs. negligible skew (exam) | Both highly negatively skewed |
| Mean within-centre correlation of SBA and exam marks | r = 0.62 | r = 0.58 | r = 0.46 |

Figure 6 compares each centre's mean SBA mark with its mean exam mark. For GCSE X and GCSE Y, the difference between centres' mean SBA mark and mean exam mark was highly variable, particularly for centres with low mean exam marks.

## Findings and discussion

### Candidate marks and grades

The SBA units of GCSEs X, Y and Z were each statistically moderated ten times, using each method in turn. Figure 7 summarises the resulting mark adjustments for the three SBA units shown in Table 4. For Method 3 and the allowed difference variants, the mean mark adjustment is close to zero. For all other methods, the mean mark adjustment reflects the difference between the mean SBA mark and the mean exam mark, hence a reduction of about 24 marks for GCSE X and GCSE Y, and a reduction of about 7 marks for GCSE Z. The variability of mark adjustments reflects the variability of SBA and exam mark levels, as shown in Figure 6, and therefore is substantially lower for GCSE Z than for the other two specifications.

Across all three specifications, the method resulting in the lowest standard deviation of mark adjustments is Method 3, the Hong Kong linear scaling method. The lower levels of mark and grade changes under

5. Standard deviation of the difference between centres' mean SBA mark and mean exam mark.

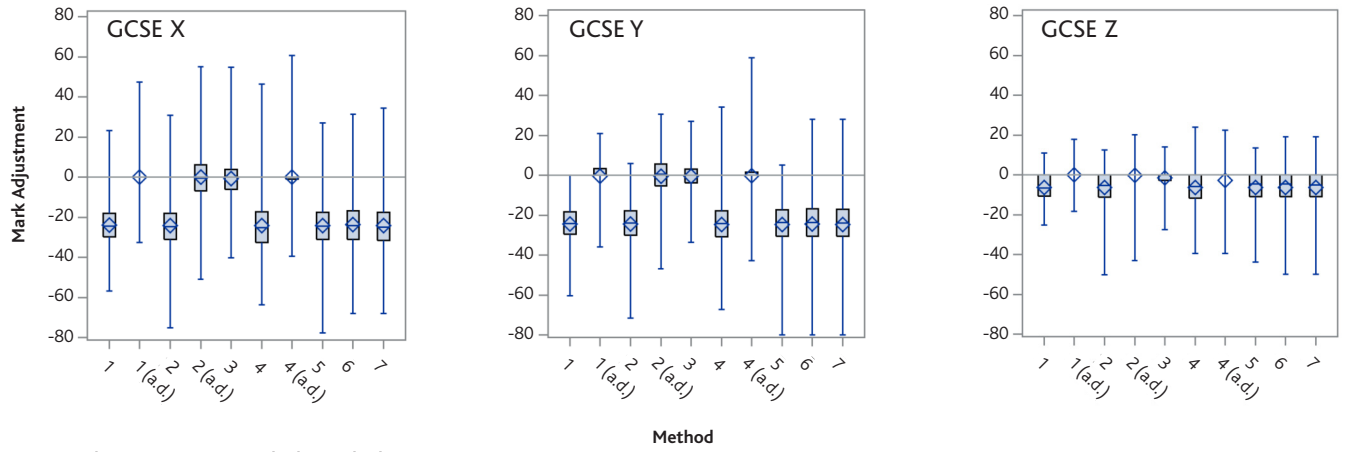


**Figure 7: Adjustments to raw marks, by method**

Method 3 can be attributed to factors accounted for by the Method 3 moderation formula that other methods in this study do not address. Method 3 does not assume that a centre's mean SBA mark will equal its mean exam mark, and or even the mean plus an allowed difference. Rather, the formula accounts for the reality of regression to the mean and does not 'expect' centres to over-/under-perform equally on different assessments. In addition, the formula adjusts the spread of SBA marks by taking into account the weighted average of spread in the SBA and exam units, so that the spread of moderated marks more closely resembles that of the original SBA marks than under most other methods. These aspects to the moderation formula minimise the overall changes to candidate marks.

The mark adjustments shown here are far larger than the typical mark adjustments made under current moderation practice (see Gill, 2015). As a result of this discrepancy, candidates' statistically moderated marks differed substantially from their operationally moderated marks[6]. Table 5 summarises the differences for the SBA units described in Table 4 and shows that they are almost as large, and variable, as the raw mark adjustments. For some methods (only among Method 3 and allowed difference variants), the mean difference between statistically and operationally moderated marks is positive, indicating that statistical moderation resulted in marks on average *higher* than operationally moderated marks.

**Table 5: Differences between statistically moderated and operationally moderated marks**

| *Method* | *GCSE X* | | *GCSE Y* | | *GCSE Z* | |
|---|---|---|---|---|---|---|
| | *Mean* | *SD* | *Mean* | *SD* | *Mean* | *SD* |
| 1 | -22.6 | 10.26 | -22.64 | 9.06 | -5.51 | 6.6 |
| 1 (a.d.) | 1.36 | 9.62 | 1.48 | 8.42 | 0.81 | 5.9 |
| 2 | -22.77 | 11.24 | -22.67 | 10.1 | -5.51 | 7.38 |
| 2 (a.d.) | 1.26 | 11.15 | 1.31 | 9.51 | 0.79 | 6.44 |
| 3 | 0.67 | 9.14 | 1.5 | 7.14 | -0.65 | 4.11 |
| 4 | -22.82 | 12.69 | -22.68 | 10.24 | -5.51 | 7.08 |
| 4 (a.d.) | 1.45 | 11.22 | 1.54 | 9.23 | -1.75 | 7.21 |
| 5 | -22.87 | 11.73 | -22.68 | 10.74 | -5.49 | 7.71 |
| 6 | -22.25 | 11.98 | -22.43 | 10.78 | -5.49 | 7.7 |
| 7 | -22.82 | 12.07 | -22.68 | 10.77 | -5.51 | 7.57 |

Because grade boundaries were recalculated for each statistically moderated mark distribution, a large mean difference between statistically and operationally moderated marks did not itself cause differences between statistically moderated and operational candidate grades. If mark differences had been uniform across centres and candidates, then overall candidate rank orders and consequently grades would have matched those of June 2015 (with lowered grade boundaries). In practice, however, differences between statistically moderated and operationally moderated marks varied substantially across centres and candidates, as already noted. The overall rank orders of candidates after statistical moderation were therefore substantially different to the June 2015 rank orders, leading to differences between statistically moderated and operational grades.

6. The June 2015 SBA marks after current moderation practices, but before conversion to UMS marks.

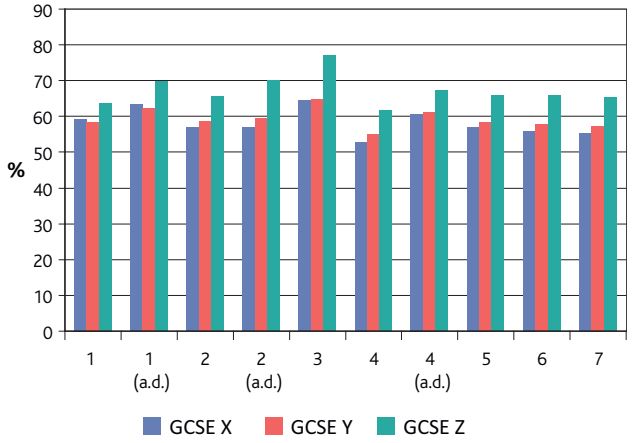


**Figure 8: Percentage of candidates awarded the same grade as the June 2015 grade, by method, by GCSE**

Figure 8 shows the proportions of candidates whose statistically moderated grade matched their June 2015 grade, for each method and specification. For all three specifications, Method 3 resulted in the highest proportion of candidates retaining their grade, and Method 4 resulted in the lowest proportion of candidates retaining their grade, reflecting the results of Table 5. The majority of statistically moderated grades were within one grade of candidates' June 2015 grade. The distribution of grade differences for Method 3 (Figure 9) shows the typical spread.

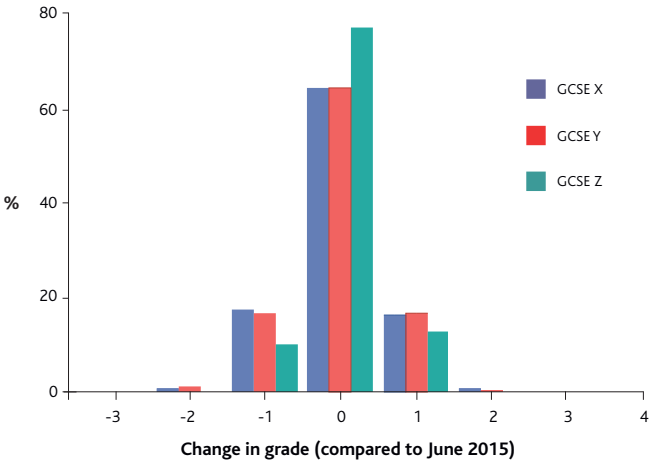


**Figure 9: Percentage of candidates per grade change, by specification, for Method 3**

A higher proportion of GCSE Z candidates retained their June 2015 grade than in the other two specifications, under all methods. This reflects the lower variability in differences between statistically moderated and operationally moderated marks for GCSE Z (Table 5), which itself reflects the lower variability in SBA and exam mark levels for GCSE Z (Figure 6). It is important that the variability in SBA and exam mark levels for GCSE Z was not only low in absolute terms, but low in relation to the mark width of individual grades. For GCSE Y and GCSE Z, variability in SBA and exam mark levels was higher in relation to the mark width of individual grades, and grade changes were thus more likely to occur.

## Rank order

None of the statistical moderation methods altered the within-centre rank ordering of candidates by SBA mark, but all methods changed the rank order by aggregated mark. Statistical moderation also resulted in a

different rank order of centres (by mean aggregated mark) compared with the June 2015 rank order. For each of the methods in this study, the final rank order of centres is fundamentally determined by exam performance. Considering moderation formulae alone, the final rank order of centres is *entirely* determined by exam performance. In practice, however, factors beyond the basic formulae, such as rounding to integer marks and truncating scores to the allowed mark range, led to the rank order of centres differing between methods.

## Conclusions

This study set out to find methods of statistical moderation used to moderate SBA and to investigate the outcomes of applying these methods to OCR GCSEs. The study identified and explored seven methods: four variations of linear scaling (Methods 1 to 4), two forms of curvilinear scaling (Methods 5 and 6) and finally rank mapping (Method 7).

Statistically moderated marks for the three GCSEs considered were generally lower than both raw marks and operationally moderated marks, in line with Taylor's (2005) findings. In terms of changes to candidate grades, this study agrees that "there were … large numbers of candidates who would change grade" (Taylor, 2005, p.51), even though the present study recalculated grade boundaries in order to preserve overall grade distributions. The high frequency of grade changes reflects high variability in the level of SBA marks compared with exam marks, as illustrated by the scatter plots of Figure 6. For GCSE X and GCSE Y, this variability was particularly large in comparison with the mark widths of grades, and so mark adjustments led to frequent grade changes.

Method 3, the Hong Kong linear scaling method, consistently resulted in lower levels of change to candidate results than other methods, and this is well accounted for by mathematical features of the mark adjustment formula. This formula minimised the overall changes to candidate marks whilst, like the other statistical moderation methods in this study, ensuring that the overall ranking of centres was determined by exam performance rather than SBA performance. It is important to note that the resulting levels of mark and grade changes were still high for both GCSE X and GCSE Y, with results very different from operational results. In terms of appropriateness for GCSE assessment, the level of transparency of Method 3 is also a potential concern, since the moderation procedure uses a more complex formula than the other methods. In Hong Kong, the complete statistical procedures and formulae are published for the public[7], but it is not clear whether a statistical procedure of this complexity would be fully understood by all stakeholders in GCSE assessment.

Under all methods, mark and grade changes for GCSE Z were smaller than those for GCSE X and GCSE Y, and these differences can be linked to clear differences in the original mark distributions: specifically, the SBA and exam mark distributions of GCSE Z had similar shape, and much lower variation in mark levels, than those of GCSE X and GCSE Y. In contrast to Taylor (2005, p.51), who concluded that there was "an absence of any pattern, across different specifications, with respect to the sizes of the adjustments arising from statistical moderation", the present study found that the magnitudes of mark adjustments and levels of grade change appeared to relate fairly directly to the characteristics of the mark distributions of the individual specifications considered.

Overall, the findings support Taylor's conclusion that "the outcomes appear to be very different (at least at candidate level) from those obtained under the current system of moderation by inspection" (2005, p.51). The present study cannot say which of the statistically moderated or operational marks is more 'correct', but clearly demonstrates that the marks resulting from statistical moderation procedures are very different to the marks awarded under current procedures. Careful work would be required in order to explain and justify statistical moderation procedures, if mark adjustments of the level seen in this study were to be accepted. In particular, it would be important for stakeholders to understand that statistically moderated marks carry relative rather than absolute meaning, and in this respect are fundamentally different to moderated marks under current procedures.

**References**


Gill, T. (2015). The moderation of coursework and controlled assessment: A summary. *Research Matters: A Cambridge Assessment publication*, *19*, 26–31. Available online at: http://www.cambridgeassessment.org.uk/Images/202665-research-matters-19-winter-2015.pdf

Hong Kong Examinations and Assessment Authority. (2010). *Moderation of School-based Assessment Scores in the HKDSE*. Hong Kong: HKEEA. Retrieved from http://www.hkeaa.edu.hk/DocLibrary/Media/Leaflets/HKDSE-SBA-ModerationBooklet_r.pdf

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices*. New York: Springer.

MacCann, R. G. (1996). *The Moderation of Higher School Certificate Assessments using a Quadratic Polynomial Transformation: a Technical Paper*. Sydney: New South Wales Board of Studies.

Ofqual. (2015a). *GCSE Computer Science: Consultation on Conditions and Guidance*. Coventry: Ofqual. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/404598/2015-02-07-gcse-computer-science-consultation-on-conditions-and-guidance.pdf

Ofqual. (2015b). *Guidance: Summary of changes to GCSEs from 2015. Coventry: Ofqual*. Retrieved from https://www.gov.uk/government/publications/gcse-changes-a-summary

Stanley, G., MacCann, R., Gardner, J., Reynolds, L., & Wild, I. (2009). *Review of Teacher Assessment: Evidence of What Works Best and Issues for Development*. London: Qualifications and Curriculum Authority.

Taylor, M. (2005). *Teacher Moderation Systems*. London: Qualifications and Curriculum Authority.

Wilmut, J., & Tuson, J. (2005). *Statistical Moderation of Teacher Assessments*. London: Qualifications and Curriculum Authority.


7. See Hong Kong Examinations and Assessment Authority (2010)