- The acceptable tolerance for each school/course/ grade combination was two ventiles higher than its historical best and two ventiles lower than its historical worst performance[11].
- Moderation took place if the estimate was outside the tolerance range
- In addition to centre moderation to ensure consistency with that centre's historic attainment, this approach also ensures that the cumulative moderated outcomes across centres for a course are within pre-defined national tolerances using the SPDs.

To implement this moderation procedure the optimisation technique based on mixed integer linear program (or programming) (MILP) was used (SQA, 2020, p.40). MILP is part of a family of Mathematical Programming techniques that optimise (by maximising or minimising) a (linear) objective function subject to a number of constraints. Mixed integer programming adds an additional condition that some of the variables are integers. MILP has many applications such as production planning, scheduling, et cetera. (Williams, 2013). SQA defined the optimisation problem as follows: When adjustment was needed the primary linear objective function was to minimise the number of candidates moved between the grades to meet the centre constraints for each grade and A-C rate (SQA Technical Report, p. 40).

As explained previously, we have not had access to the student datasets and detailed methodology and the detailed algorithm/computer code used by SQA (nor the resources/time to undertake such an analysis in the context of a rapid review). These would be needed to conduct a comprehensive analysis of the working of the ACM, and/or examining in detail the overall suitability of using the MILP approach to moderation, as well as exploring whether some changes in the definition of the optimisation problem, including the formulation of the primary objective function, could have produced better moderation results. The datasets and codes would also be required to conduct modelling and evaluate alternative approaches. Such an analysis would be necessary to address various questions raised by our review, for example relating to evidence of unexplained variance in moderation between different schools (with some centre moderation results for some subjects being lower than they should, based on the centre's historical attainment trends), between subjects in the same schools (e.g. MSA position paper) and (anecdotally) between candidates within the same cohort. We have, however, seen local authority and school level analysis of trends in grade adjustment, suggesting a number of problems highlighted below.

Issues arising from the moderation process

The first issue is the one that has received lots of media attention: the schools in areas with higher level of socio-economic disadvantage have been downgraded more than schools in more advantageous areas. Concerns about the impact of statistical moderation on the outcomes of pupils from disadvantaged schools were voiced repeatedly before the publication of the results on 4th August. For example, a

[11] If any centre had only one or two years' attainment history on a course for which they had entries in 2020, then the historic range for that centre on that grade was extended in each direction, to provide a range of five ventile bands. The additional allowance of two ventiles in each direction is then further applied to this extended ventile range. Centres with no history (i.e. presenting entries for a course for the first time) were therefore awarded the original estimates submitted by their centres (SQA, 2020, p.36)

letter sent to the DFM in July by Johann Lamont MSP, detailing comments made by constituents – made the following points:

> The SQA is going to change pupils' grades to ensure attainment is in line with "prior attainment" of that centre. This will disproportionately punish schools in more deprived communities whilst simultaneously over rewarding schools in more affluent communities. This is because the pass rate in the former is historically lower than that of the latter. (letter from Johann Lamont MSP, copied to the DFM, 15th July, 2020)

This outcome might have been anticipated. Existing research shows that there is a large variation in the accuracy of the predicted grades between different types of schools and by student socio-economic background (Wyness, 2016). There are two reasons why the schools in areas with higher level of socio-economic disadvantage were downgraded more than schools in more advantageous areas:

1. Schools in socially and economically disadvantaged areas historically have on average lower levels of attainment than schools in advantaged areas. Therefore, standardizing in line with prior attainment of the centre disproportionately affects schools in more deprived areas. As a result, high performers at historically low attaining schools would be disproportionately affected by moderation based on historical record of the school because their grades are out with the aggregate level historical performance.
2. Pupils in poor schools are more likely to be lower attaining. Lower attaining pupils are harder to predict, and more likely to be over-predicted. Hence, moderating grades based on the actual performance of their schools would inevitably result in more downgrading for these pupils: students from the most disadvantaged backgrounds are more likely to experience moderate to severe over-prediction (from 2 to 5 grade points) than those from the most advantaged background (ibid).

Therefore, if acceptable tolerance for each school/course/ grade combination is based on the school's historical performance, then given the tendency of over-predicting grades in these schools the estimates would need to be adjusted (downgraded) more to meet the acceptable tolerances. This approach, which at a centre level managed to produce plausible distributions in line with and often better than a centre's historical patterns, seemed to fare far worse at the level of subjects[12] and worse still at the level of individual pupils. Although, this year, the results of schools in areas of socio-economic deprivation were overall better compared to previous years, emerging evidence suggests that individual level injustices have happened, with 'outliers', such as high performing pupils in these low performing schools, who were arbitrarily downgraded. The evidence of the narrowing of the attainment gap between the students from the least and the most disadvantaged socio-economic backgrounds in 2020 has been praised, yet this feels like over-focusing on a wrong metric, since this aggregate trend hides the fact that high attaining students from lower socio-economic backgrounds and improving schools in disadvantaged areas were downgraded more by the moderation procedure, than their more socially and economically advantaged peers in historically better

[12] Based on the information provided by head teachers and LAs. A full analysis of the datasets would be needed to confirm this.

performing schools. More research is needed to gauge the nature and extent of these patterns.

Some schools presented data for this review (based on the analyses of the adjusted grades in relation to the teachers' estimates and historical trends) that shows that, although the 2020 grade distribution at the level of schools broadly resembled historical grade distributions, there were huge variations between the 2020 results and the historical trends for some subjects, and from the evidence presented by the head teachers, there were many candidates whose grades were moderated down in an apparently arbitrary way. Conversely, mediocre students in high performing schools may be unduly rewarded with higher than their estimated grades. While the latter problem was far less discussed in the media than the former one, we saw that many teachers felt very strongly, not only where their estimates were downgraded, but also when their estimates were upgraded in an unjustified way.

The second problem suggested by emerging evidence was that some centres were extensively moderated and ended up with attainment levels lower than they expected or had achieved in the previous four years[13]. From our conversation with the SQA technical panel, it seems that this inadvertently resulted from trying to prevent a creation of centre constraints that are too rigid and do not allow some degree of variability for centres that might perform better/worse in 2020 than in the past four years. To achieve this greater variability, the tolerance intervals estimated for every centre (for each grade/course/level), first based on the centre's performance over the years 2016-2019, were expanded both upwards and downwards. The rationale for allowing a tolerance both ways, rather than upwards only, was to avoid unnecessary upward adjustment of estimates which were lower than the historical performance. Yet, in some cases, where a centre was found to have 'overestimated' compared to the historical attainment, it was adjusted downwards towards the point lower than their historical attainment (although still within their tolerance interval). We think it plausible that in addition to what was mentioned, there were cases where centres had estimated better grades than in previous years, yet still within the tolerance range, but might have been downgraded anyway, because the national level corrections for the tolerance range were added, which might have been lower than the centres' historical attainment[14]. Of course, these are only hypotheses which cannot be tested without having access to the computer code, used for the moderation algorithms, and the data. Yet, it seems that introducing more rigid restrictions on the lower boundary of tolerances, which would not allow the centres to be moderated below their historical averages, would have solved these problems.

[13] Based on the information provided by head teachers and LAs. A full analysis of the datasets would be needed to confirm this.

[14] Distributions (SPD) to introduce adjustments to the tolerances, based on centre-level constraints. The SPDs were estimated as a proportional national attainment level for each grade on a given course. Although the SQA (2020) Technical Report says that they "sought to take the average of as many recent comparable years of attainment data as was available for the course", it subsequently clarified the SPDs were based on two year (2018 and 2019) averages only for National 5 courses, while the SPDs produced for Higher and Advanced Higher courses were predominantly based on 2019 data (p. 29). The latter means that, if the 2019 national result were particularly low for some courses (which we know has been the case for some Higher courses), that could have affected the acceptable tolerances for these courses and downgraded the centre results more than it should, based on the centre's historical attainment alone.

Another potential source of the problem might be the way in which the optimisation problem was defined. A linear program contains two elements: a cost/optimal function and a set of constraints. The constraints must be met at all costs. The cost function, on the other hand should be minimised provided that no constraint is violated. The solution tells us the optimal value of each unknown, such that the cost function is minimised and each constraint is met. If the priority is to prevent an extreme grade movement this should have been set as a constraint. The cost function, on the other hand, should depend on the difference from the historical attainment patterns, since this is what the approach sought to minimise. Yet, SQA did this the other way around. They set the cost function depending on grades movement and assumed that giving high penalties to particular types of movement (e.g. three or more grades) would prevent these movements; but this did not always seem to be the case. High cost is unlikely, but it is still possible. Only if it was given as a constraint could it never happen.

The third problem lies in small numbers of entries in many courses[15] and the resulting problem of over-moderation of some courses[16], which were big enough to be included in the moderation procedure, but still far too small to obtain reliable statistical estimates. Thus, relatively small numbers of candidates distributed across many centres means it is challenging to make statistically significant decisions across centres and nationally in some low-uptake subjects (ibid).Yet, the SQA states that, for the approach adopted in the moderation process for setting centre constraints, sample sizes are not critical (SQA, 2020). SQA believed that the problem of year on year variability of the outcomes for small centres was solved by setting the tolerance range for each grade/course/centre as the minimum to maximum attainment of the centre on this grade for this course, for years 2016-2019, plus additional tolerances to allow for year-on-year changes in the centre performance. Yet, it seems that the latter still did not solve the problem of year-on year variability. This is to a large extent because teachers' estimates for small uptake courses are less based on the historical patterns and more on teachers' knowledge of the pupils whose grades were estimated this year (after all teachers know that six students who they taught last year might be very different to six students they teach this year).

The fourth problem – downgrades of more than one grade or from pass to fail – has been referred to as the waterfall effect (which was downplayed by SQA in reporting of the national trends). What we have seen in the local authority data analysis looked more like an avalanche effect– the smallest number of entries moved from A to B, then larger numbers from B to C and still larger from C to D. One local authority specifically mentioned that the largest number of moderations were for grade C[17]. We posit various reasons for this. First, when adjustment was required, entire bands were moved up or down.

---

[15] In 2019 for National 5, at least half of class entry sizes were made up of 19 or fewer candidates; for Higher, at least half of class entry sizes were made up of 14 or fewer candidates; and for Advanced Higher, as least half of class entry sizes were made up of four or fewer candidates (SQA 2020).
[16] Based on the moderation vs result data provided by teachers.
[17] Based on the information provided by head teachers and LAs. A full analysis of the datasets would be needed to confirm this.

> … where it was necessary for entries in a refined band to be moved into another refined band in another grade, those entries previously in the recipient refined band were displaced, rather than the two groups of entries merging. (SQA, 2020, p.34).

This had a knock on effect on the lower band's entries, which were respectively moved further down (when A grades become B grades, the lowest band of B grades may have to become C grade, etc.). Thus, for entries in a refined band (e.g. band 5 in grade A) to be moved into next refined band (e.g. band 6 in grade B) those entries previously in the band 6 were displaced, rather than the two groups of entries merging (SQA 2020). That might result in too many entries being moved down.

We think that one way to avoid this was to use ranking of students within the grade bands (submitted by centres) and to move a minimum amount of lowest ranked entries from the bottom of a higher band to the top of the next band when required, and merge them with the entries which are already within this band, with moving the lowest ranked entries from this band to the next one only if the total number of estimates within a refined lower band exceeded the centre's historical proportion with tolerances.

A related problem is that students in lower grade bands paid a price for overestimation in higher grade bands. The following example illustrates this. The total number of estimates within grade A exceeded the centre's historical proportion with tolerances, while the total number of estimates within the grade B corresponds with the centre's historical proportion with tolerances. Yet, when one moves entries out the grade A and down to grade B, then as the result of this is to move the lowest band(s) of B grades into C grade bands, and so on. As a result, although the original number of entries achieving grades B and C were within the tolerance interval, the students would be downgraded (including from pass to fail) because their teachers 'overestimated' their higher performing classmates.

The potential inequity here lies in the arbitrary nature of the approach; its inability to deal with cohort by cohort variation and particularly its effects on individuals. The use of an appeals system is a technical solution that fails to appreciate the impact on individuals and subsequently on public opinion. As stated by CYPCS in their position paper to the review:

> However as a method it appears to have ignored the fact that each statistical point on the graph is an individual young person whose work, effort and attainment have been moderated based on factors entirely outwith their control and which have no bearing on their individual abilities. It succeeds in creating an overall perception of fairness but fails to deliver actual fairness for individuals. (CYPCS position paper)

Email correspondence between the SQA and the government suggests that this issue and its explosive implications for public opinion appear to have not been fully grasped by SQA, other than through its recourse to appeals, until the EQIA was finalised in July, nor by the government until after the results and EQIA were seen at the end of July. Even at this late stage, the focus seemed to rest on presenting a positive picture (the attainment gap had closed in general terms) rather than seeking a fuller understanding of the nuances in the data.

> The DFM has asked that we do lots of digging in the stats to show how young people from deprived backgrounds have not been disadvantaged by the results. (Government email, 6th August)

We concur with SQA's position that it was not possible, to engage in dialogue at a centre level. We do, however agree with many stakeholders that the following would have been possible:

- Analysis of data to identify anomalies, drawing on government and local government expertise in statistics.
- Dialogue with local authorities to discuss and moderate in a qualitative sense (for example engaging with the rationales for cohort variance collected by local authorities.

After examining this evidence, we believe that – despite the constraints of time and resources – more systematic engagement between SQA and different stakeholders in a process of co-construction might have resulted in developing a moderation system that was more equitable to individual candidates. This could have mitigated the impact that the publication of the results had on young people, their families, teachers and general public. It is a view reflected in the evidence submitted by stakeholders, for example:

> A stronger commitment to genuine partnership working may well have headed off the subsequent debacle. It would certainly have eliminated the bulk of individual discrepancies (EIS position paper)

Post Certification Review and Appeals

Perceived strengths	Perceived weaknesses
• The original PCR process was technically appropriate with clear guidance, based on a review of individual candidate evidence. • PCR was free-of-charge and thus there were no cost disincentives for centres. • The priority 'fast-track' PCR process was designed to address the needs of students whose university offers were dependent on their grades.	• PCR was perceived widely as an appeals process, rather than an integral part of the awarding process. • This was exacerbated by SQA not publishing details of the statistical moderation process and its likely implications. • While technically appropriate, the PCR took insufficient account of equity, especially the impact of the process on individuals. • The revised appeals process following the decision to revert to teacher grades narrowed the grounds for appeal, with subsequent problems for schools and young people. • Appeals can only be initiated by centres, with no right of appeal for young people.

Second, SQA has repeatedly emphasised to us that many centre estimates were inaccurate; and yet, the system put in place by SQA denies students an avenue to appeal against inaccurate estimates.

A related issue raised by some stakeholders, especially young people, is the view that the appeals process continues to deny young people the option to personally instigate appeals. Only a school can lodge an appeal. According to CYPCS,

> Being denied a direct right of appeal, where they believe they have experienced discrimination, breaches not only the young person's right to an effective remedy under Article 13 and the prohibition on discrimination in Article 14 of the ECHR, and Article 2 of the UNCRC and in the case of disabled young people Article 23 of the UNCRC. (CYPCS position paper).

We suggest that, following the announcement by the First Minister on 1st September 2020 that the UNCRC will be incorporated as far as possible into Scottish law, the time has come to review the rights and role of young people in the examinations appeals process.

Equalities

Perceived strengths	Perceived weaknesses
• The principle of 'fairness to all learners' was clearly stated as underpinning the ACM. • EQIA and CRIA documents were produced by SQA. • There was a clear focus on bias in assessment, and well-received training on unconscious bias.	• EQIA and CRIA documents were produced very late in the process, with only limited evidence that equalities issues had been fully considered at the development stage of the ACM. • SQA does not routinely collect equality data about candidates. • SQA's position that it does not have a sound legal basis for routinely collecting information about protected characteristics appeared to impede analysis of data in relation to equalities issues. • The nuanced impact of the ACM in relation to equalities seems to have been obscured by a debate as to whether the ACM advantaged or disadvantaged cohorts in low SES centres.
Overall assessment	
There need for more systematic and robust systems in future to address equalities issues, particularly in relation to the collection and analysis of data, and in the central role of equalities impact assessment in the design and implementation of awarding systems.	

It is clear that equalities issues were considered at various stages of the process of developing and implementing the ACM. We have, for example, seen evidence of discussion relating to bias in the estimation process as early as March (followed by the well-received unconscious bias training), and (following an offer of support from EHRC on 9th April), ongoing dialogue between SQA and various organisations such as the Scottish Youth Parliament and EHRC regarding equalities issues. A primary focus on equalities work seems to have been in the area of bias in assessment, with less focus on how the moderation process itself might produce inequity. For example, the following extract from a presentation to the SQA Board suggests that a focus on bias may even have prevented analyses related to identifying equalities issues, by anonymising data.

> Measures built in to moderation and validation process e.g. all data anonymised for analysis, analysis at aggregate level. (presentation to the Board, 9th July)

Moreover, we found only limited evidence that equalities issues were systematically considered or built into the development of the ACM from the outset, other than the sorts of instances related above and through general commitments to and acknowledgement of equalities issues. Concerns about the absence of an Equalities Impact Assessment (EQIA) were raised as early as May by the Scottish Parliament Education and Skills Committee and Equalities and Human Rights Commission. At this point, the DFM stated it was a matter for the SQA (email correspondence). There is little evidence that this was undertaken comprehensively until July, after results were finalised. SQA (in its technical report published in August) described equality impact assessment as being developed 'in parallel with' the development of the ACM, rather than it being an integral part of the process. A meeting note on 11 July indicated that 'SQA have committed to completing and publishing an EQIA to support the certification model, but have not given an indication of a likely date yet' (Scottish Government 2020 Awarding Presentation to the Deputy First Minister, 11th July). The EQIA and accompanying Children's Rights Impact Assessment (CRIA) documents and associated processes for their development attracted considerable criticism from interested stakeholders. .

> The draft CRIA was not considered by the SQA Board until 30th July and the published document does not address the full range of rights engaged or properly assess the impact of decisions. This meant that the predictable negative impacts of the alternative certification model were not identified and no mitigations were put in place. In particular, the application of a statistical modelling approach at school level resulted in clear and obvious unfairness and disadvantage for many young people. The CRIA should have identified this. (CYPCS position paper)
>
> From the start of this process the NASUWT also pressed for the SQA to publish the details of any equality impact assessment, particularly in respect of the extent to which equalities issues were taken into effective consideration throughout the design and implementation of the moderation process for 2019/20. It is very difficult to understand how decisions were being taken in the absence of any completed equality assessment and the late arrival of the EIA only served to further undermine teachers' confidence in the process.

> (NASUWT evidence to the Scottish Parliament Education and Skills Committee, 7th August 2020)

The EHRC has also been critical of SQA for shortcomings in its treatment of equalities issues, while acknowledging the constraints on this:

> SQA did act upon much of the information we provided. However, their effectiveness in meeting their duties was hampered by a lack of embedded structures and practice, which would have allowed them to fully consider equality in the development of the ACM. They were constrained in what they could do not only because of the very tight timescales they were working to but because:
>
> - There was limited existing knowledge and expertise in meeting the PSED, which meant awareness of equality and an understanding of their statutory equality duties were not built into their decision-making structures;
> - They do not routinely collect equality evidence, including equality data about candidates and the views and lived experiences of people with protected characteristics; and
> - There was no systematic process to ensure such equality evidence and data was used to inform decision-making. (EHRC position paper)

A lack of access to equalities data is evident in correspondence between SQA and the government in July 2020 – 'a request to perform analysis to support an Equalities Impact Assessment they are performing on their Alternative Certification Model' (email from government official to John Swinney, 24 July. SQA requested government assistance to analyse attainment patterns using protected characteristic data. SQA do not have any records of the individual data for pupils apart from grades and estimates (and postcode).

> SQA do not hold equalities data and therefore cannot examine the 2020 approach for impact on protected characteristics. (Note attached to internal government email dated 3rd August)

Two alternative approaches to this analysis were not subsequently possible: SQAs view was that they could not take receipt of equalities data from government in the absence of a 'legal basis on which to hold and process pupil characteristic data'; and the government deemed that it could not undertake the analysis prior to results day as this might be seen as unwarranted interference in the workings of an independent exams regulator.

> This means that for the analysis to proceed we would have to take receipt of SQA grade data. We would not otherwise receive the pre-moderation data and there could be some concern about us having access to this given the independent role of the SQA in using this data to award qualifications. However these concerns are somewhat reduced as (i) we would not be in a position to take receipt of the data from SQA until after results day on 4th August and (ii) the relevant documentation would make it clear that the data was shared only for the purposes of this analysis and that it would be deleted

> immediately upon its completion. We have also consulted with [Redact s30(c)] who have advised that there is no legal impediment to proceeding with the analysis. (email from government official to John Swinney, 24 July)

Our data (interviews with teachers and parents) suggested that some protected groups were disadvantaged more than others, for example children with learning difficulties, and yet the full extent of this was unknown at the time due to a lack of analysis by SQA and the government. More research is necessary to explore these patterns.

The circumstances outlined above seem to have led to a situation where some of the impacts of the moderation model were not fully anticipated or mitigated. We have, for example, found little evidence in email communications between or public statements by SQA and the government that the equity nuances had been anticipated or publicly acknowledged (even fully understood) prior to the furore that erupted after the publication of results. Emails (for example those sent internally on 4th August) suggest a government priority to defend the position that the system is fair on low SES students, in the face of accusations that low SES centres were more likely to have had awards downgraded (e.g. emails about suggested lines of argument to justify the position). Within this dichotomous argument, some implications were clearly grasped (e.g. general pattern of rising attainment in low SES schools[21]), but the focus on this, combined with a lack of systematic statistical analysis at a fine grained level, seems to have obscured other effects (e.g. reported negative effects on high performing students in low performing schools[22]).

Another equity issue lies in variation in the evidence used to underpin estimation by centres. Although estimates were largely based on the evidence submitted prior to the closedown, there is evidence that, in some centres, later evidence was taken into account, which to cite one respondent was 'incredibly unfair' (local authority panel). Moreover, the evidence for appeals was considered up to 29 May (teacher panels) – this created an issue of inequity since there was a huge variation in the ability of young people to work from home and submit additional evidence (and there was a variation between schools in the amount of available support, virtual teaching, etc.). According to one Director of Education, there needed to be a clear statement that evidence should not be generated after lockdown – this caused ambiguity and unfairness – but neither SQA nor the government provided such a statement.

Communication and transparency

Perceived strengths	Perceived weaknesses
• Extensive approach to communication developed by SQA. • Some guidance was clear and well-received. • Some evidence that SQA is developing approaches to working with young people.	• Unclear and inconsistent approaches to communication. • An apparent reluctance by SQA to share some information, widely seen as a lack of transparency.

[22] As stated earlier in the report, more analysis of data is needed to explore these emerging patterns.