Schola europaea

Office of the Secretary-General

European Baccalaureate Unit

Ref.: 2020-07-D-2-en-1

FINAL REPORT ON THE MODERATION METHOD FOR THE EUROPEAN BACCALAUREATE MARKS IN 2020

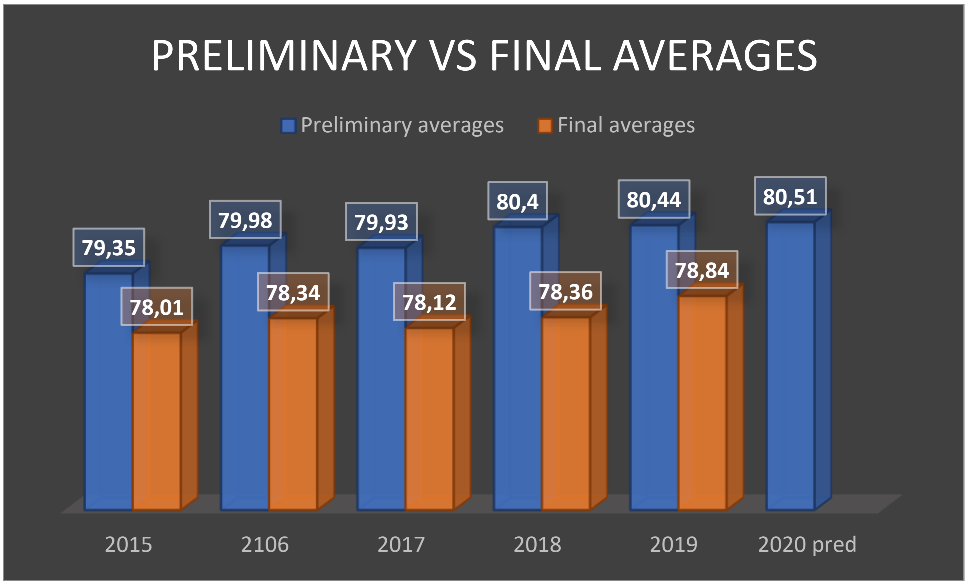
Goal:

The Board of Governors cancelled the organisation of the 2020 European Baccalaureate examinations and decided to award the European Baccalaureate Diplomas using only A1, A2 and B1 marks (preliminary marks). The distribution of results might differ significantly from previous years' final marks distributions. The European Baccalaureate Regulations foresee the possibility of applying moderation, the prerogative of the Chairman and Vice-Chairmen of the European Baccalaureate. In order to safeguard the credibility of the European Baccalaureate Diploma, it might be necessary to:

- Demonstrate that the final marks distribution, calculated using only A1, A2 and B1 in the previous years, differed significantly from the actual final marks. Also to compare this year's preliminary marks distribution with previous years' final marks distributions.
- Create a method/algorithm to moderate the 2020 final marks. The method should not negatively affect this year's students, compared with those of earlier years.

Preliminary investigations

The mean preliminary and final marks of the years 2015 to 2019 are shown in the chart below, as well as that of the final marks for 2020 calculated using A1, A2 and B1 marks only (A2 marks were not yet known for 2020 so the A1 marks were replicated as a prediction for the A2 marks; for the sake of simplicity, I will refer to these as "preliminary marks").



The overall averages fell by between 1.34 and 2.04 marks between the preliminary and the final marks, so moderation should decrease the mean of 2020 preliminary marks from 80.51 to somewhere between 78.5 and 79.2, with the maximum and minimum decrease recorded in the past years used as a reference range. The fact that the average of the final marks ranged between 78.02 and 78.84 in the past five years should also be taken into consideration.

To test whether uniform moderation (where every student's final mark is decreased by the same percentage to guarantee an ideal mean) could be used, the distributions were compared.

- First, there is no evidence for any similarities between each year's distribution of preliminary and final marks. Using a chi-squared test, the table below shows the p-values of comparing the two distributions for the years 2015 to 2019.

	2015	2016	2017	2018	2019
p	2.7×10^{-9}	7.3×10^{-12}	2.2×10^{-16}	6.0×10^{-21}	1.36×10^{-13}

- The distribution of the 2020 preliminary marks was also compared with the distributions of previous years' final marks using a chi-squared test. The table below show the p-values for its comparison with the years 2015 to 2019.

	2015	2016	2017	2018	2019
p	7.6×10^{-29}	1.1×10^{-13}	6.4×10^{-8}	0.031	0.23

Based on the results of the above two tests, it can be safely concluded that it is not enough to apply uniform moderation as the distribution of the marks also needs to be adjusted.

My first attempts were to use back and forth normalisation with a Box-Cox transformation. Apart from being perfectly capable of achieving a desired mean and standard deviation, it also significantly improved the p-value in the chi-squared tests in most years. This improvement, however, was still not good enough to declare a "really good fit"; not to mention that its functioning would not be transparent for most people involved.

I opted therefore for a different, possibly more broadly intelligible approach.

Initial proposal

Step 1: Determine the desired distribution: students' results have been grouped in cohorts each corresponding to a range of 5 marks (except for the first one): 0 to 59.99, 60 to 64.99, 65 to 69.99, etc. In order to fix the distribution, we can propose possible "ranges" for the percentage (or number) of students in each cohort based on the evidence from the previous years. The final decision about what distribution to adopt (i.e. which percentage to use for each cohort) would then lie with the body responsible for moderation of the results. Once this decision is made, we would reach the desired number of students with the marks 0 to 59.99, 60 to 64.99, etc.

Using the past five years' final marks, the ranges for the different cohorts using the number of students for 2020 are shown in the table below.

	min.	max.
0-59.99	38	51
60-64.99	137	144
65-69.99	240	296
70-74.99	340	377
75-79.99	394	429
80-84.99	431	488
85-89.99	342	414
90-94.99	170	227
95-100	18	23

Step 2: It is requested that it be ensured that no student, with an overall result of at least 60 as preliminary mark, would eventually fail. Therefore, once the results are known, the number of students in the two lowest cohorts may need to be adjusted: if the number of students with a preliminary mark below 60 (f) is smaller than the "ideal" number of students who should end up with a mark below 60 (i, determined in step 1), we will have to use f to determine the number of students who are failing. Consequently, it is necessary to add the difference ($i - f$) to the desired number of students in the 60 to 64.99 cohort.

Step 3: Take the preliminary marks and determine the percentage grades corresponding to the limits of the different cohorts. For example, if there are 30 students projected to have a result below 60, consider the result of the 30th student as the highest possible grade which will eventually fall into this cohort (in the case of a tie, we can stop before the 30th student to be as lenient as possible). The next mark will be the lowest in the 60 to 64.99 cohort. Denote the mean of the two numbers by L_{60}. Repeat this process with 65, 70, etc.

Step 4: Use linear interpolation to calculate the moderated final marks. Individual preliminary results will be referred to as p in the formulae below.

We would thus derive from this the desired distribution of the students' final marks (apart from small differences arising from potential ties).

It would still be necessary to check that the overall mean would fall into the expected range.

Denoting the preliminary marks by p and the corresponding moderated final mark by $f(p)$, this means that the following formula is to be applied:

$$f(p) = \begin{cases} 95 + \dfrac{(p_{max} - 95)(p - L_{95})}{p_{max} - L_{95}}, & x \geq L_{95} \\ 90 + \dfrac{4.99(p - L_{90})}{L_{95} - L_{90}}, & L_{90} \leq x < L_{95} \\ 85 + \dfrac{4.99(p - L_{85})}{L_{90} - L_{85}}, & L_{85} \leq x < L_{90} \\ 80 + \dfrac{4.99(p - L_{80})}{L_{85} - L_{80}}, & L_{80} \leq x < L_{85} \\ 75 + \dfrac{4.99(p - L_{75})}{L_{80} - L_{75}}, & L_{75} \leq x < L_{80} \\ 70 + \dfrac{4.99(p - L_{70})}{L_{75} - L_{70}}, & L_{70} \leq x < L_{75} \\ 65 + \dfrac{4.99(p - L_{65})}{L_{70} - L_{65}}, & L_{65} \leq x < L_{70} \\ 60 + \dfrac{4.99(p - L_{60})}{L_{65} - L_{60}}, & L_{60} \leq x < L_{65} \\ x, & 0 \leq x < L_{60} \end{cases}$$

where p_{max} is the highest preliminary mark, while L_n is the lower limit of the cohort with a moderated final mark between n and $n + 5$ (that is, if $L_n \leq p < L_{n+5}$, $n \leq f(p) < n + 5$).

As mentioned previously, the calculations supporting the above proposal were based upon a projected preliminary mark, which only took the A1 and B1 marks into consideration as the A2 marks had not been available at the time.

Using the actual preliminary marks (A1+A2+B1 marks)

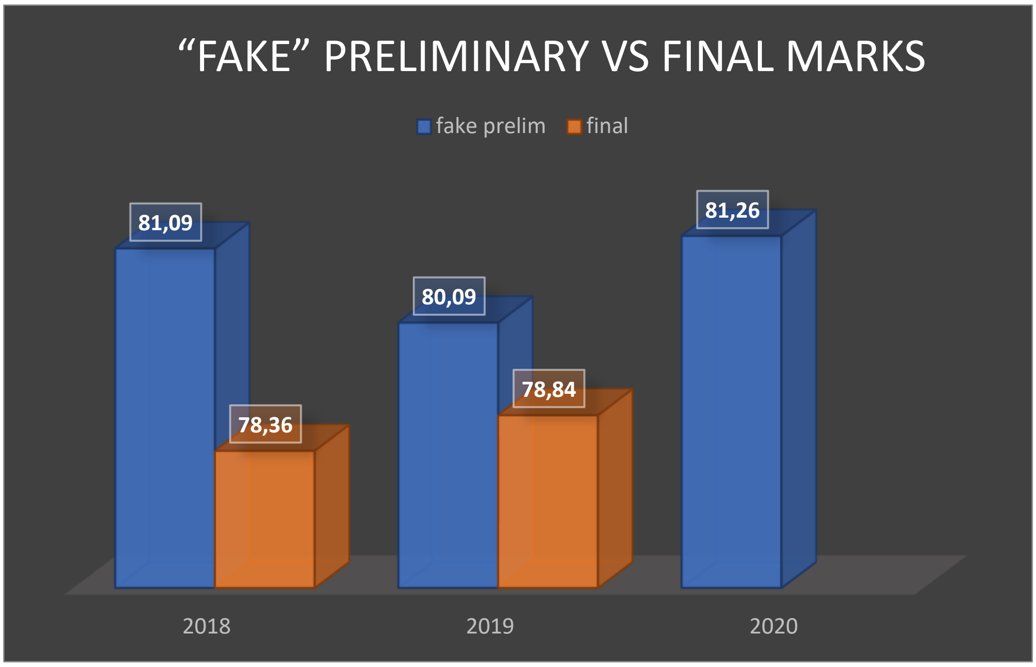
Having received the actual preliminary marks (based on the A1, A2 and B1 marks), performing respective chi-squared tests confirmed that their distribution is statistically significantly different from the distribution of the final marks of past years (while the final marks of past years were statistically very similar), as illustrated in the table below.

	2015	2016	2017	2018	2019	2020 prel.
2015	1	0.999	0.997	0.999	0.97	0.011
2016		1	1	1	1	0.008
2017			1	0.999	0.999	0.04
2018				1	1	0.062
2019					1	0.00000043
2020 prel.						1

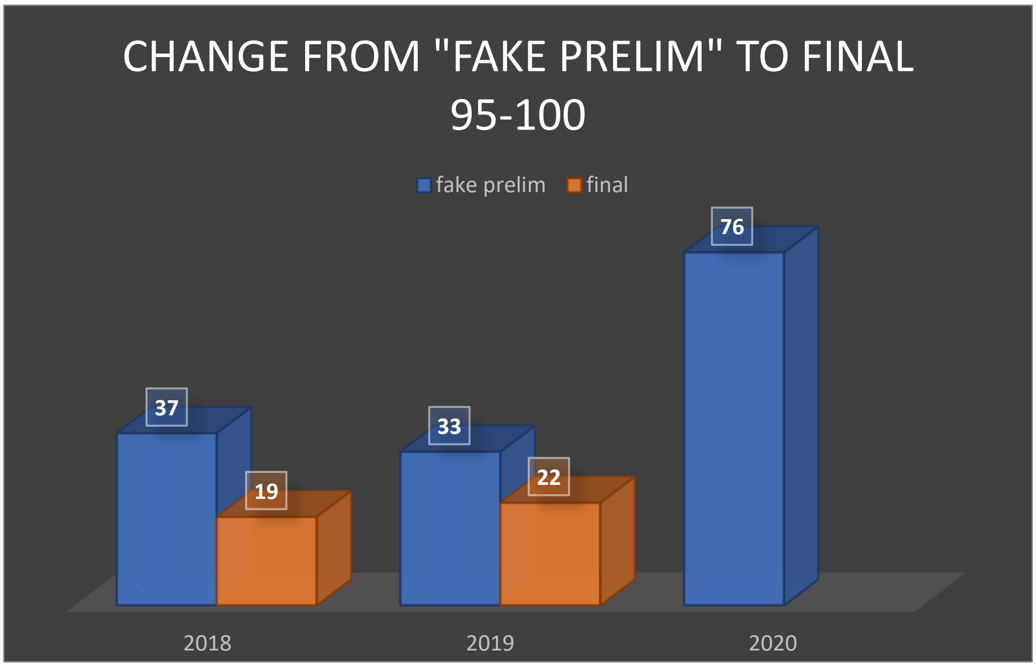
In line with the decision of the Board of Governors, this justifies the application of moderation. It was felt advisable, however, as we will see, to modify the formerly proposed moderation method to students' advantage.

A considerable difference in both the mean and the distribution of these marks and those of both the preliminary and the final marks of previous years was detected. It then became reasonable, in order to provide fairer and more balanced moderation in the light of the new data available, to compare the results with the results of the past years calculated in the same way (A1+A2+B1 duplicated; this calculation will be referred to as "fake preliminary results" below). This would make it possible to produce a fair comparison of the performance of the different populations. Since the time period between the arrival of the A2 marks from the schools and the meeting of the Board of Inspectors was less than 72 hours (including a whole weekend), this could only be done for the past two years. The findings and their consequences for the proposed moderation, as well as some further considerations, will follow:

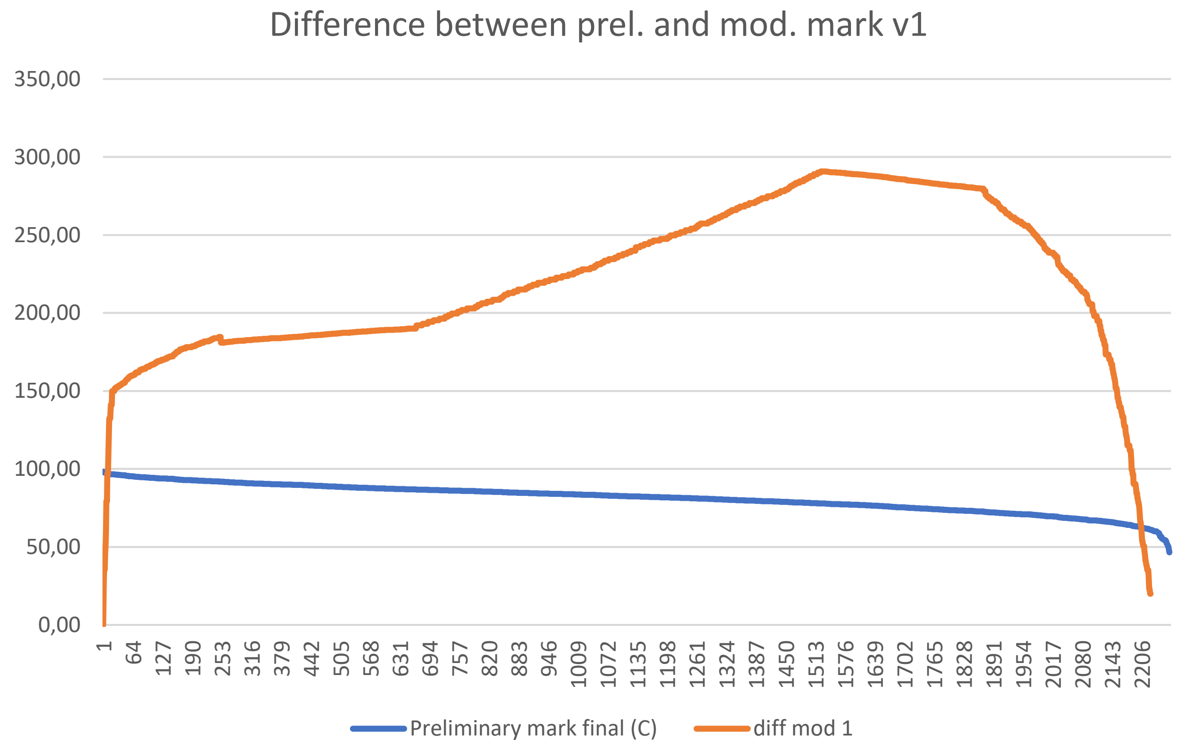
1) The difference between the average of "fake preliminary" and that of the final marks ranged between 1.25 and 2.73 (the final mark average being lower). If this is also taken into consideration when determining the range in which the mean of the final marks should fall, the aforementioned range is extended, and now the final average should be between **78.01** and **80.01**.



2) Comparing the distribution of the "fake preliminary" marks of this year and the past two years, one striking difference is the percentage of students in the highest cluster (95-100). While the highest value in previous years was **1.75%**, this year it is **3.36%**, which implies a large group of top students. To be able to reach a fair distribution this year, the change in the number of students in the top cohort from the "fake preliminary" to the final marks was looked at. If the range of percentage change is preserved, the number of students in the top cohort can be expected to be between 39 and 51 (compared with 18 to 23 if calculated from the final marks of previous years).



3) If we simulate moderation based on the original proposal, we also find that there is an extremely sharp decline in the results of students in the top cohort. This is illustrated in the graph below. (100.00 corresponds to losing 1 mark.)



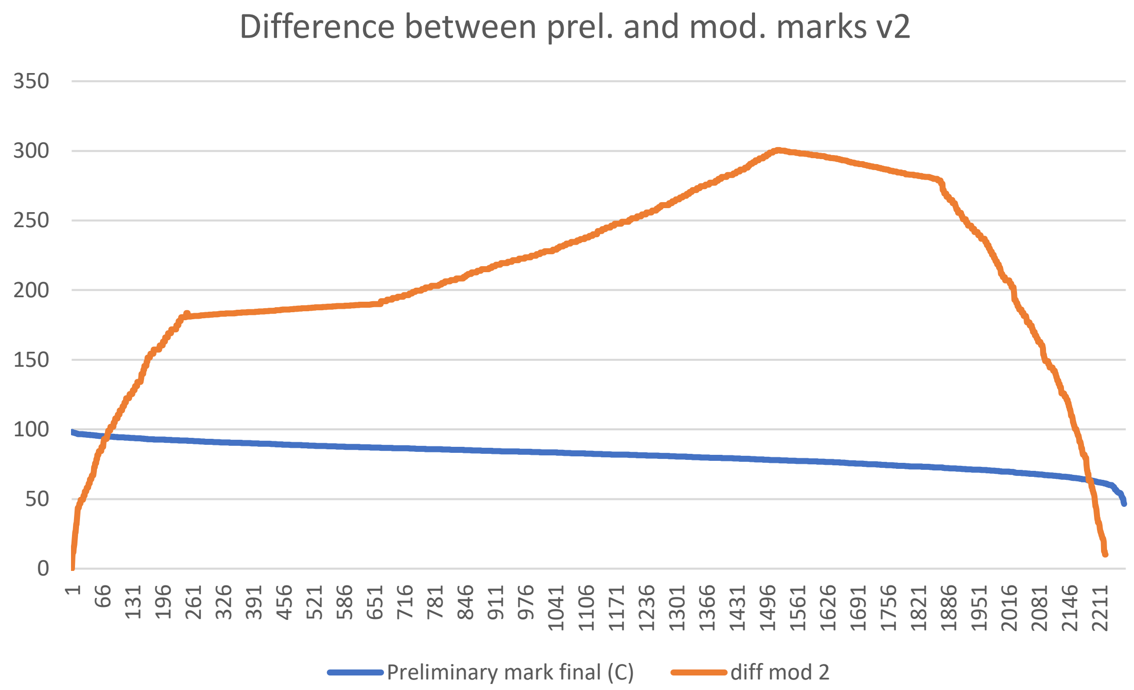
This is the result of using linear interpolation for a much smaller number of students than in other cohorts. It is desirable to avoid this and provide a smoother decrease. In order to do so, the cohorts need to be redefined so that the different cohorts are of comparable size. (Note that the bottom cohort, that is the number of students who are below 60, is fixed.) A better distribution is achieved if, instead of constant 5-mark cohorts, we use the following (the minimum and maximum number of students, determined by the results of the past years, in each cohort is given):

	min.		max.
0-59.99	38		51
60-69.99	382		440
70-74.99	340		377
75-79.99	394		429
80-84.99	431		488
85-89.99	342		414
90-100	190		300

Note that this will imply adaptation of the moderation formula, which becomes

$$f(p)=\begin{cases} 90+\dfrac{(p_{max}-90)(p-L_{90})}{p_{max}-L_{90}}, & x\geq L_{90} \\ 85+\dfrac{4.99(p-L_{85})}{L_{90}-L_{85}}, & L_{85}\leq x<L_{90} \\ 80+\dfrac{4.99(p-L_{80})}{L_{85}-L_{80}}, & L_{80}\leq x<L_{85} \\ 75+\dfrac{4.99(p-L_{75})}{L_{80}-L_{75}}, & L_{75}\leq x<L_{80} \\ 70+\dfrac{4.99(p-L_{70})}{L_{75}-L_{70}}, & L_{70}\leq x<L_{75} \\ 60+\dfrac{4.99(p-L_{60})}{L_{70}-L_{60}}, & L_{60}\leq x<L_{70} \\ x, \quad 0\leq x<L_{60} \end{cases}$$

4) If a lenient version of the above moderation (meaning that the maximum number of students is used for top cohorts and the minimum number for bottom cohorts) is applied, the differences between the preliminary and the moderated marks are shown in the graph below.



While the sudden drop in the results of the top cohort has been avoided, it can be observed that the results of individual students are decreased by up to 3 marks from the "fake preliminary" to the final results. The average student loses about 1.5 marks, between preliminary and final mark, based on evidence from the previous years. In order to avoid negatively affecting this year's students, compared with those of earlier years, it is reasonable to set a ceiling on the number of reduced marks at 1.5 marks, in line with the "average" student of the past years.

This means that the formula to be used needs to be adjusted again. Using the lenient version of the moderation above and applying the 1.5-mark ceiling to the reduction of any individual mark results in the following formulae:

$$f(p) = \begin{cases} i(p), & p - i(p) < 1.5 \\ p - 1.5 & \text{otherwise} \end{cases}$$

where

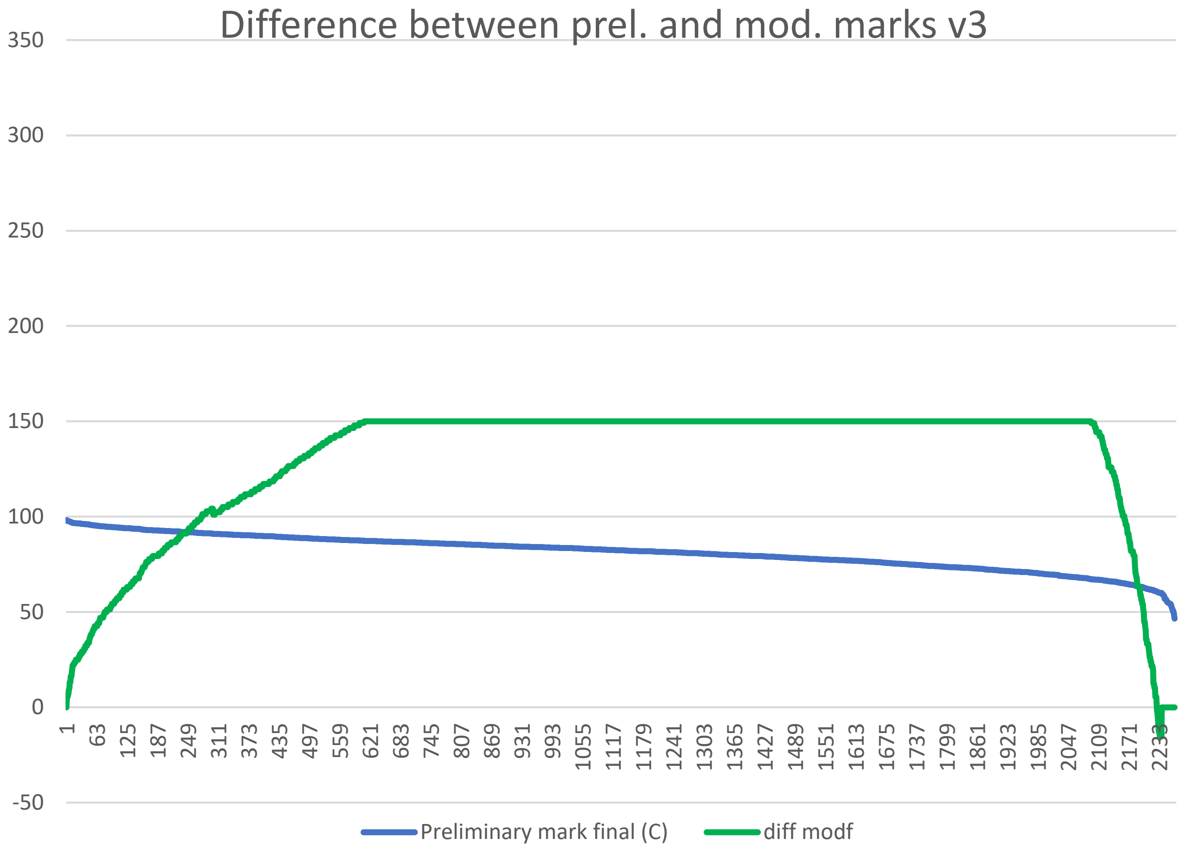
$$
i(x) = \begin{cases} 90 + \dfrac{8.2(x - 91.05)}{7.15}, & x \geq 91.05 \\ 85 + \dfrac{4.99(x - 86.6)}{4.4}, & 86.6 \leq x < 91.05 \\ 80 + \dfrac{4.99(x - 82.4)}{4.2}, & 82.4 \leq x < 86.6 \\ 75 + \dfrac{4.99(x - 78)}{4.4}, & 78 \leq x < 82.4 \\ 70 + \dfrac{4.99(x - 72.8)}{5.2}, & 72.8 \leq x < 78 \\ 60 + \dfrac{9.99(x - 59.8)}{13}, & 60 \leq x < 72.8 \\ x, & 0 \leq x < 59.8 \end{cases}
$$

(As before p is the preliminary mark and $f(p)$ is the corresponding moderated final mark.)

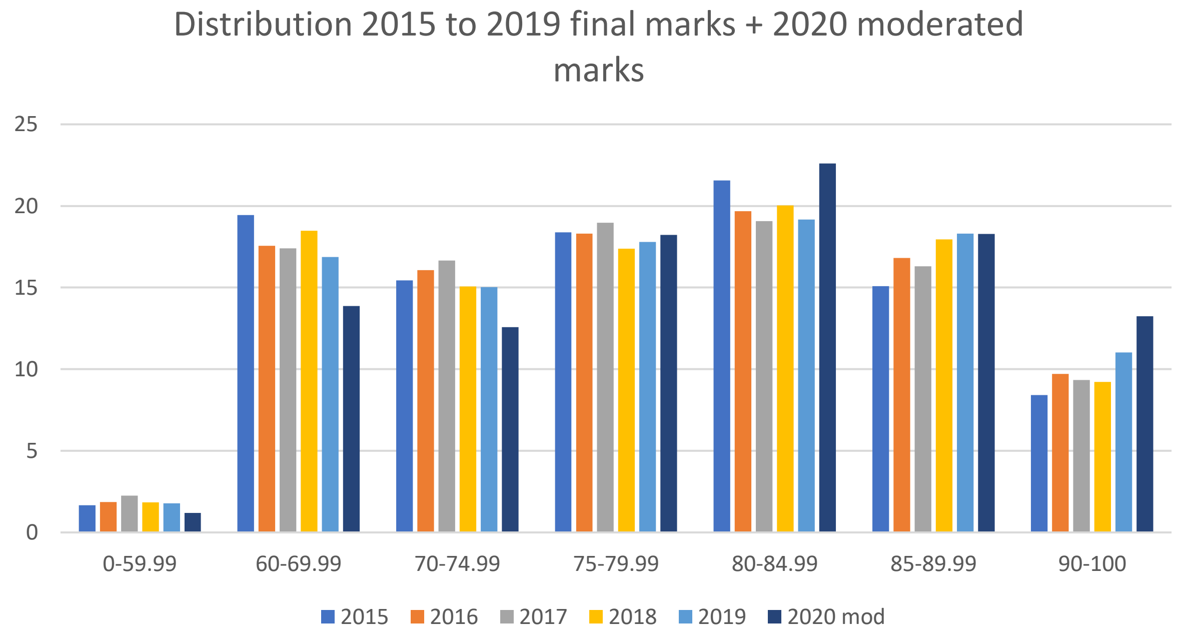
This will obviously improve the overall results as well as the individual results for the vast majority of students, but we will still remain within our predefined ranges: the mean result will be 79.96 (range: 78.01 to 81.01 so it is very close to the higher extreme of the range). The number of students in each cohort will also be in or near the target range:

	min.	mod.	max.
0-59.99	38	27	51
60-69.99	382	375	440
70-74.99	340	346	377
75-79.99	394	386	429
80-84.99	431	430	488
85-89.99	342	401	414
90-100	190	300	300

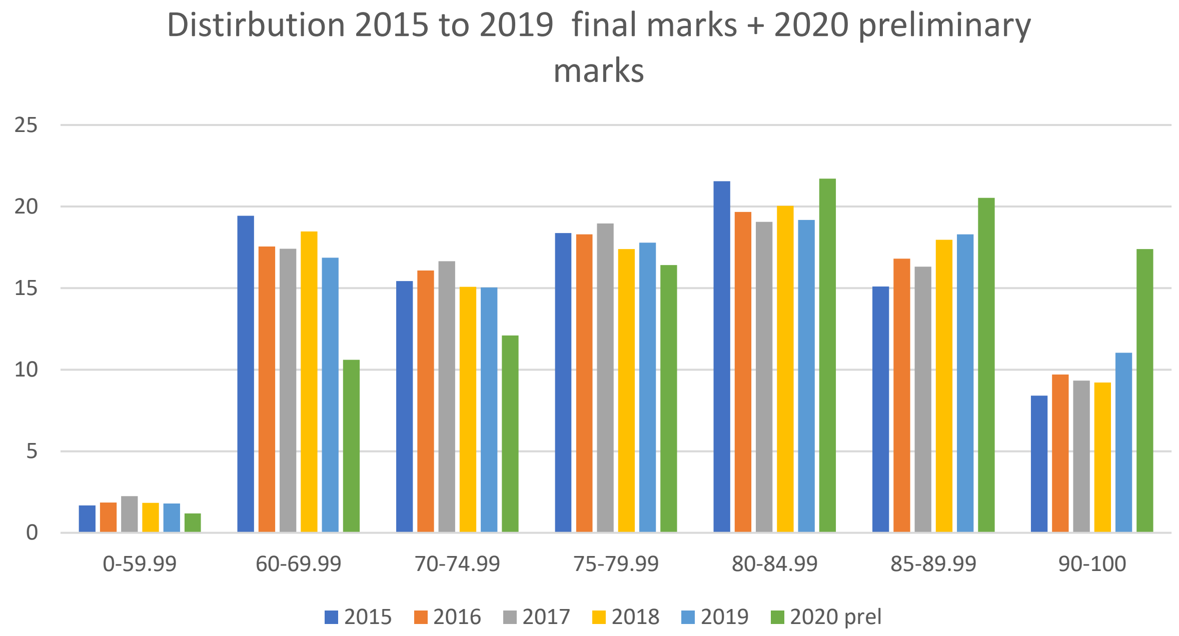
If we look at what the graph of the differences between the preliminary and moderated marks looks like, it can be concluded that we have successfully reduced moderation’s negative effects upon individual students.



Comparison of the distribution of final marks in the past five years with that of this year's final marks shows that there is still a difference that can arise from the specificities of this year's population of students.



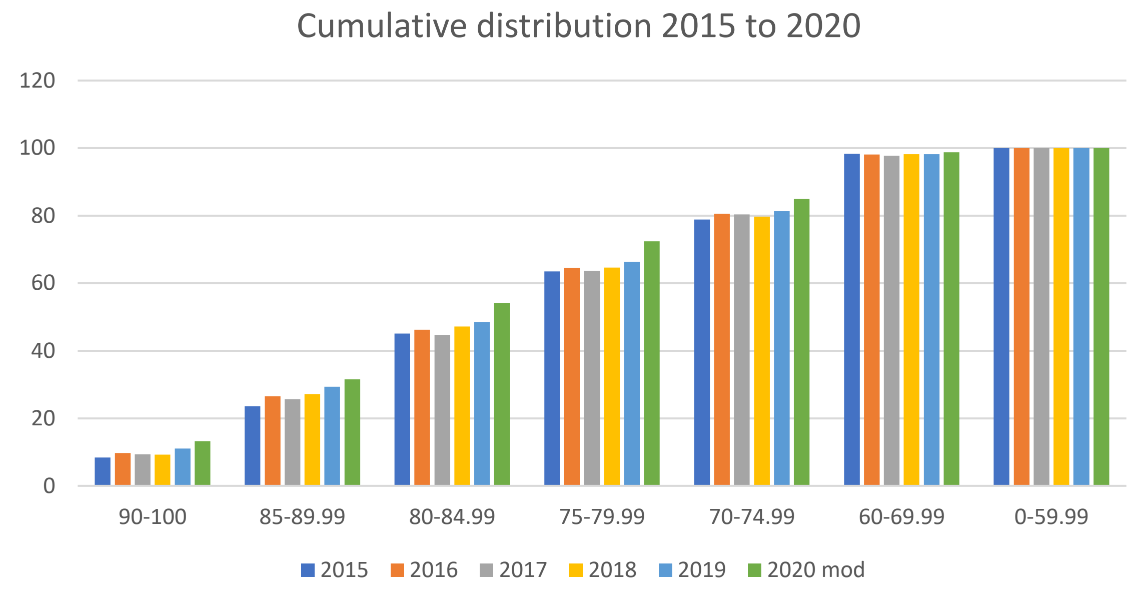
At the same time the striking difference that characterised the distribution of this year's preliminary marks when compared with the final marks of the last five years has been reasonably moderated, preserving the credibility of the European Baccalaureate.



The visual impression is confirmed repeating the chi-squared test with the moderated marks: its results show that the statistical difference between the distributions of the final marks of the past years and this year's moderated marks has been considerably reduced.

	2015	2016	2017	2018	2019	2020 mod.
2015	1	0.999	0.997	0.999	0.97	0.45
2016		1	1	1	1	0.727
2017			1	0.999	0.999	0.571
2018				1	1	0.686
2019					1	0.9
2020 mod						1

The comparison of cumulative graphs of the past five years and of this year helps in understanding that, reasonably, this year's students have not been negatively affected, compared with those of earlier years.


Cumulative distribution 2015 to 2020
120
100
80
60
40
20
0
90-100
85-89.99
80-84.99
75-79.99
70-74.99
60-69.99
0-59.99
2015
2016
2017
2018
2019
2020 mod