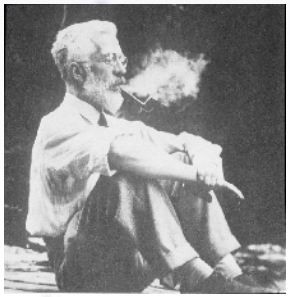
Applied Machine Learning

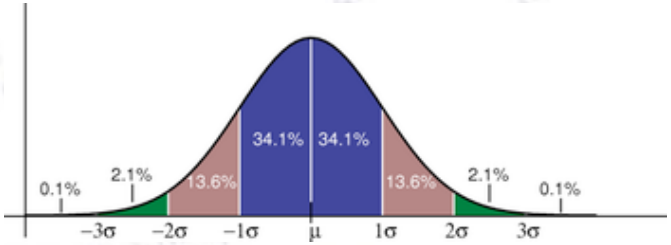
Comments on Grading








Troels C. Petersen (NBI)



"Statistics is merely a quantisation of common sense - Machine Learning is a sharpening of it!"

Teacher's Scores

The grading input data

There were 20 projects in total: 15+5 on the 12th and 13th.

We were 5 teachers, who all gave 5 grades [1,10] on the points of evaluation:

- Complexity of problem and depth of solution (incl. appendix)
- Choice of methods and arguments behind
- ML performance and own evaluation of it
- Clarity of presentation & Learning of classmates
- Implementation, technical details, optimisation, etc. (incl. appendix)

Thus each project got 25 scores from the teachers. We decided to weight teachers equally, and the five points of evaluation as: [0.3, 0.15, 0.25, 0.15, 0.15]

In addition, each project got on average ~18 scores from (68) fellow students.

The following are to show you the cross checks that we've gone through to even out differences, and evaluate as accurately as possible.

Calibrating between teachers

We do not use the grading scale in the same way and extend:

```
                          Raw data:
Teacher1 :    Mean of scores: 5.95    RMSE of scores = 0.94
Teacher2 :    Mean of scores: 6.17    RMSE of scores = 1.09
Teacher3 :    Mean of scores: 6.88    RMSE of scores = 0.84
Teacher4 :    Mean of scores: 7.06    RMSE of scores = 0.82
Teacher5 :    Mean of scores: 6.53    RMSE of scores = 0.81
```

Therefore, we calibrate the scale to have same mean and RMSE:

```
                       Calibrated data:
Teacher1 :    Mean of scores: 5.00    RMSE of scores = 1.00
Teacher2 :    Mean of scores: 5.00    RMSE of scores = 1.00
Teacher3 :    Mean of scores: 5.00    RMSE of scores = 1.00
Teacher4 :    Mean of scores: 5.00    RMSE of scores = 1.00
Teacher5 :    Mean of scores: 5.00    RMSE of scores = 1.00
```

In this way, one can more fairly compare and combine them.

```
Uncalibrated average RMSE of teacher scores for single projects: 1.42
Calibrated average RMSE of teacher scores for single projects   0.61
```

Checks between teachers

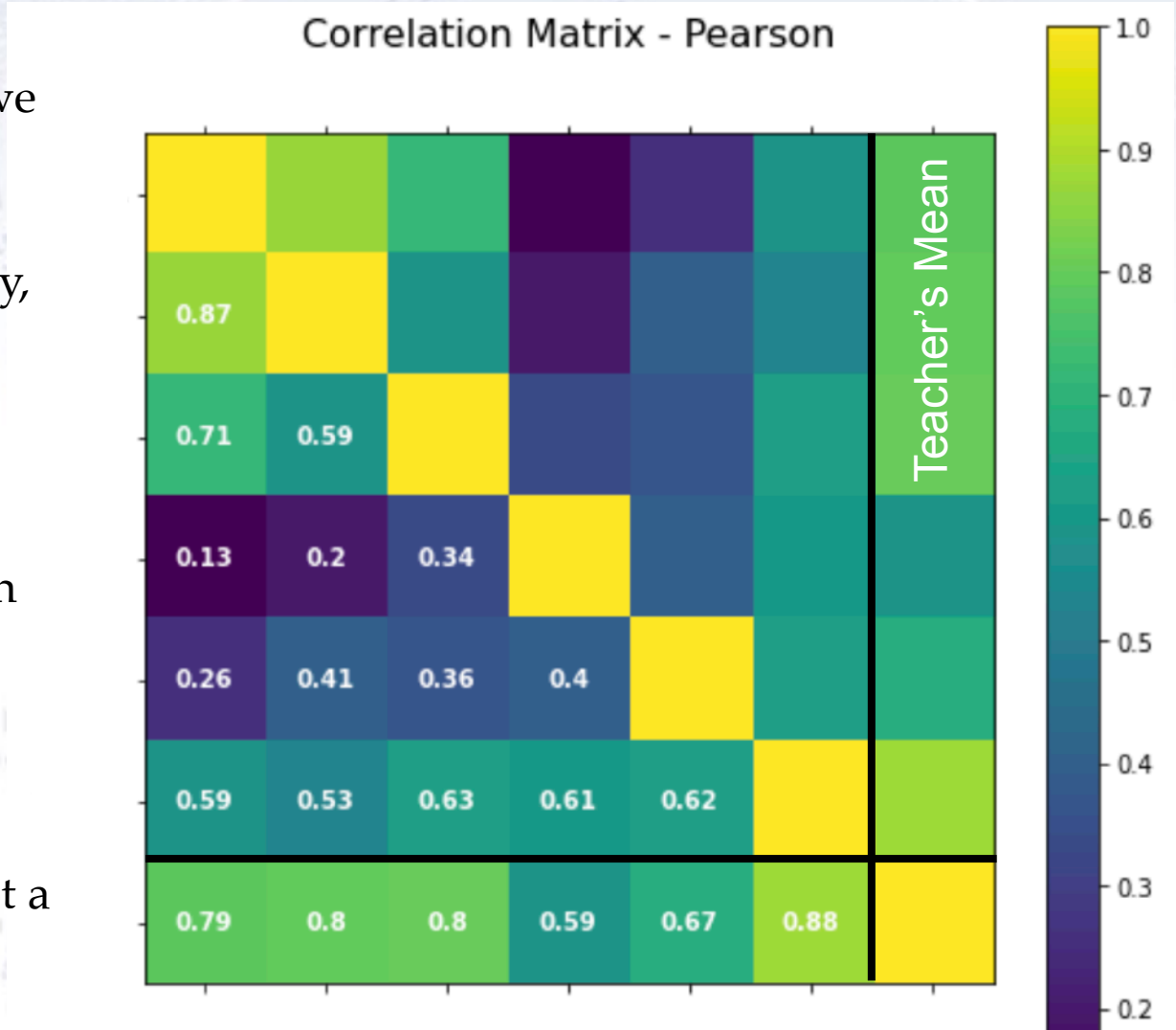
The typical standard deviation between the five teacher averages was about 0.5.

It is a difficult task to evaluate! That is why we take averages.

We don't agree perfectly, but no combination of persons have less than 0.13 in correlations!

The mean is included in the last column/row, respectively.

As a cross check, I evaluated twice and got a 0.87 correlation!

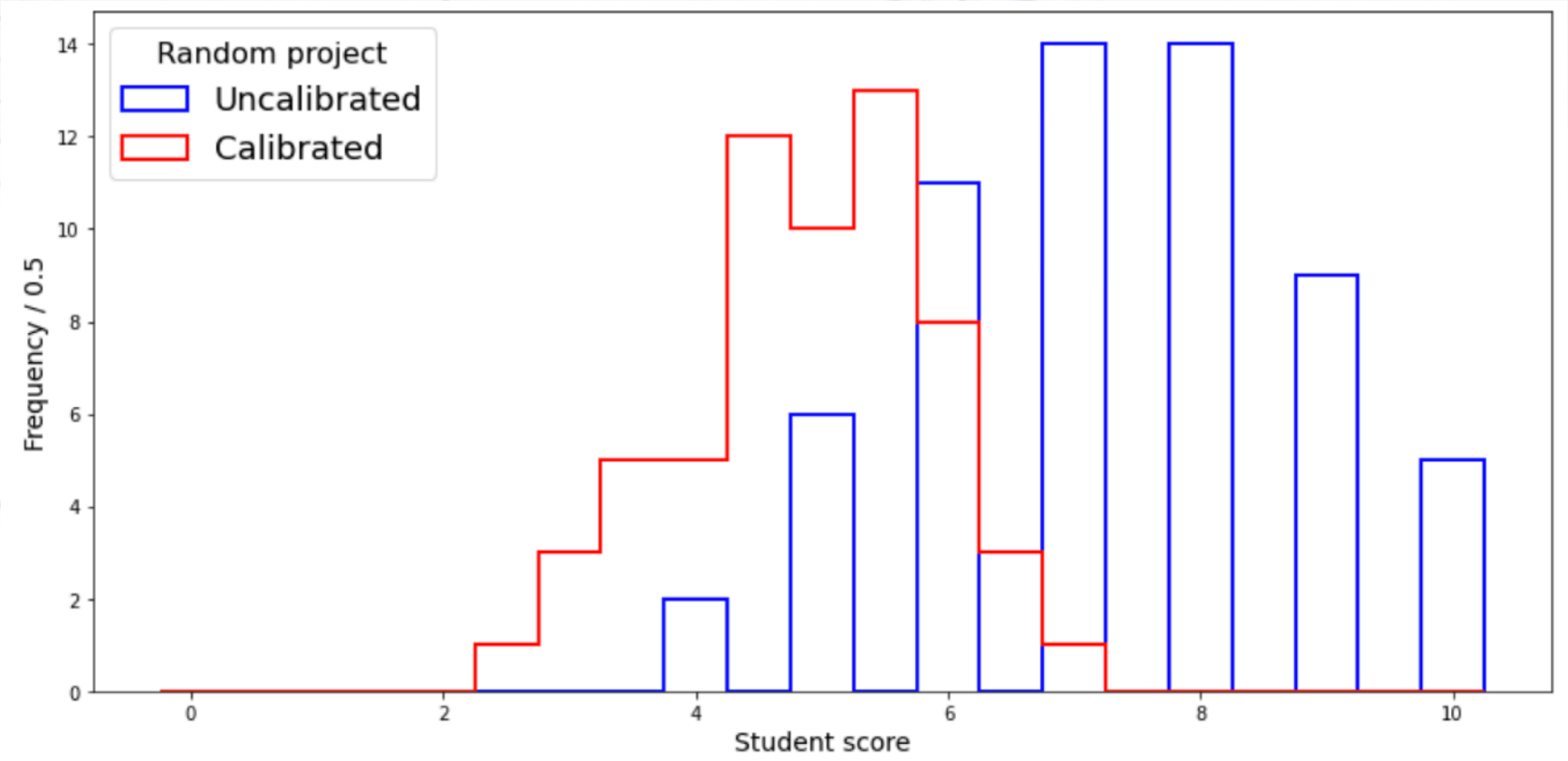


Student's Scores

Student gradings

One evaluation point was how good YOU were at evaluating others ML work.

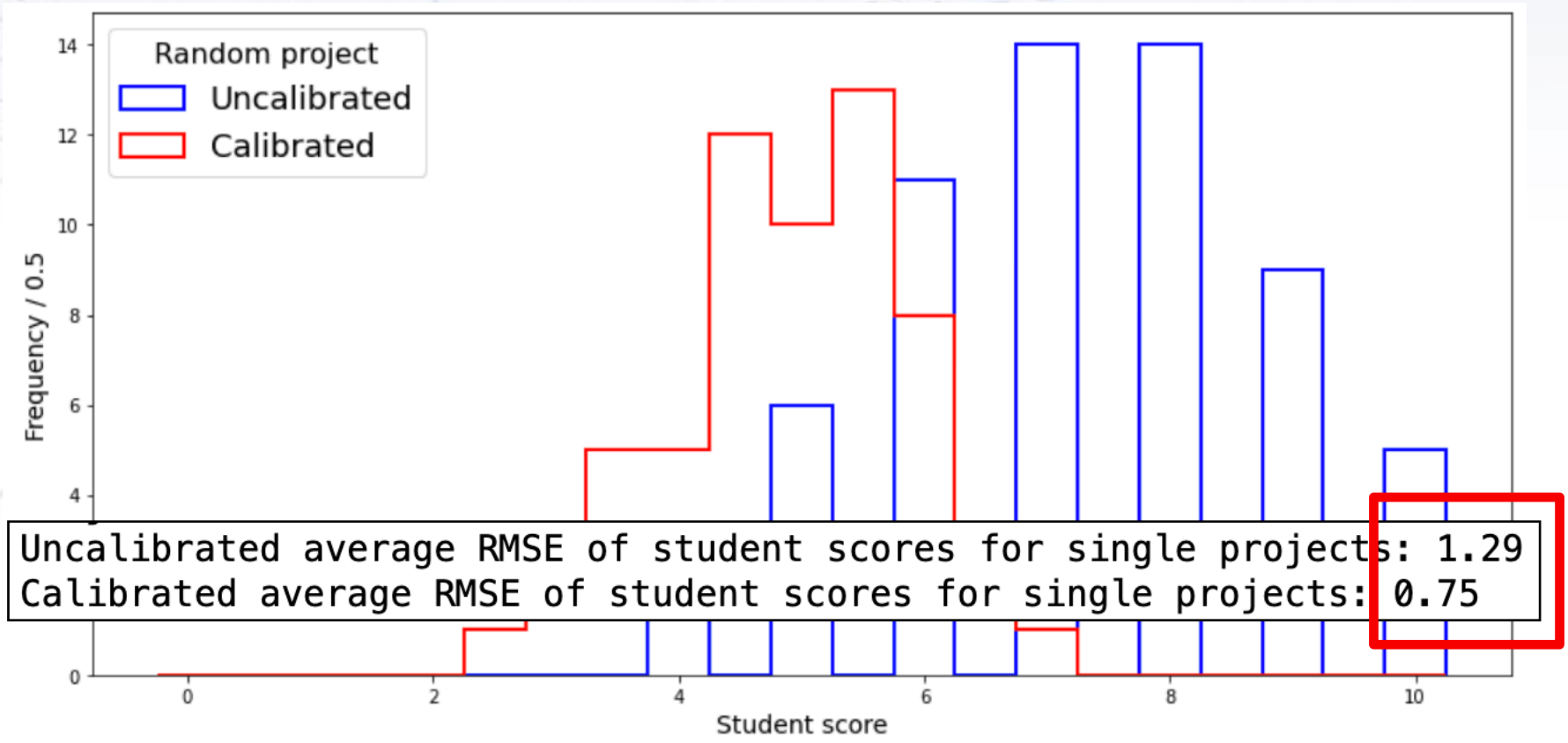
Students (also) don't agree - in fact less! Below is a project's distribution of scores.



Student gradings

One evaluation point was how good YOU were at evaluating others ML work.

Students (also) don't agree - in fact less! Below is a project's distribution of scores.



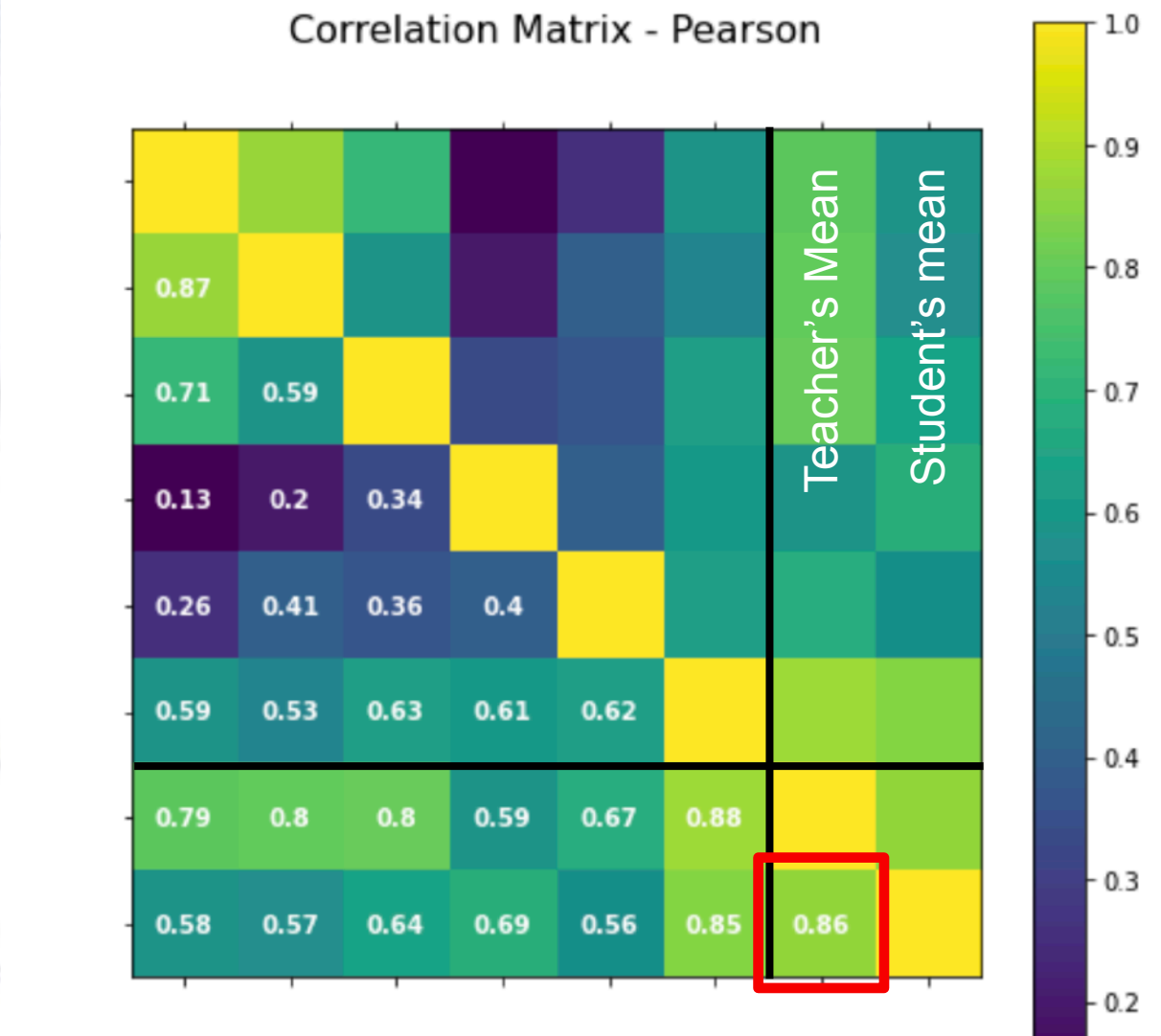
Teachers vs. student average

Considering the average of the student evaluations, we can compare….

I'm happy to see, that teachers to a large extend agree with students.

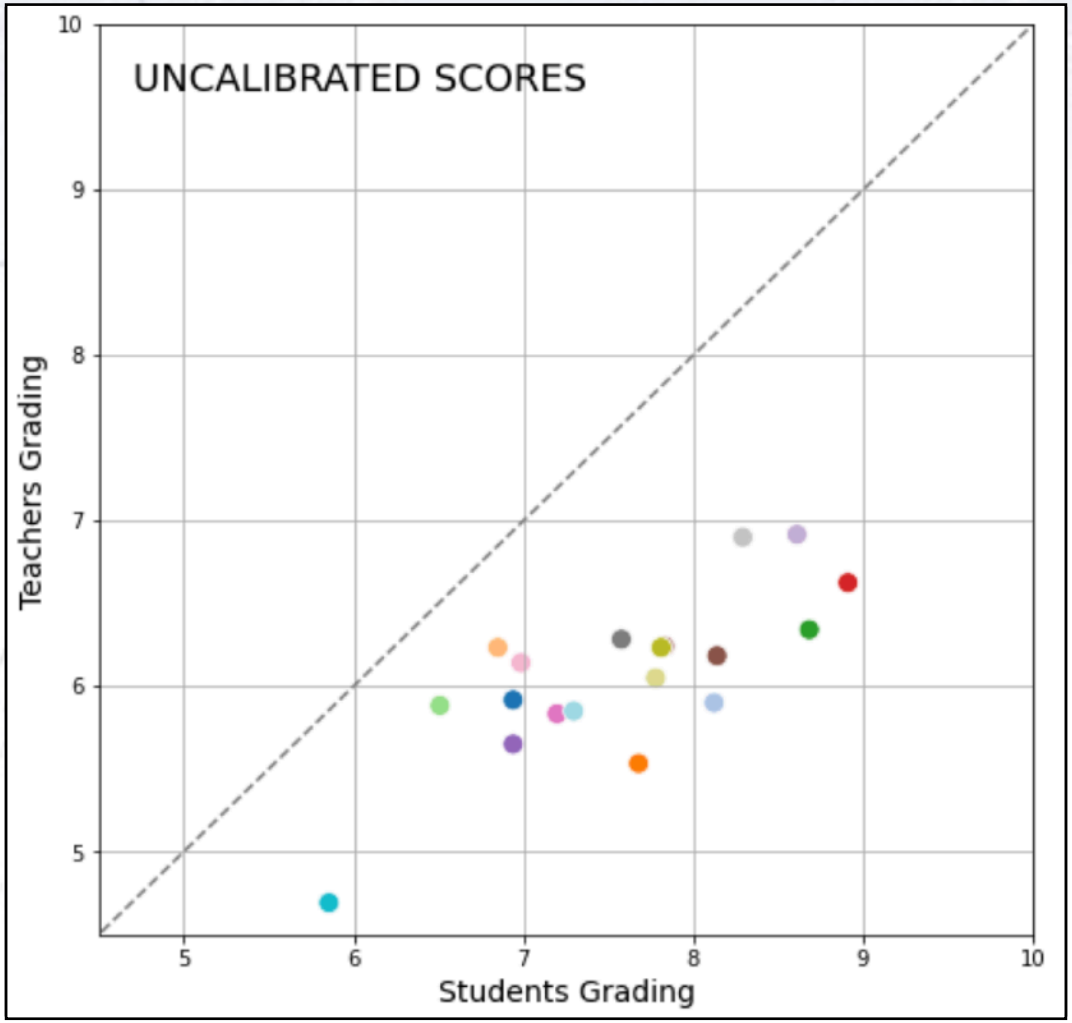
The correlation between the teacher and the student averages is 86%.

All teachers correlate significantly (and positively) with the student average.



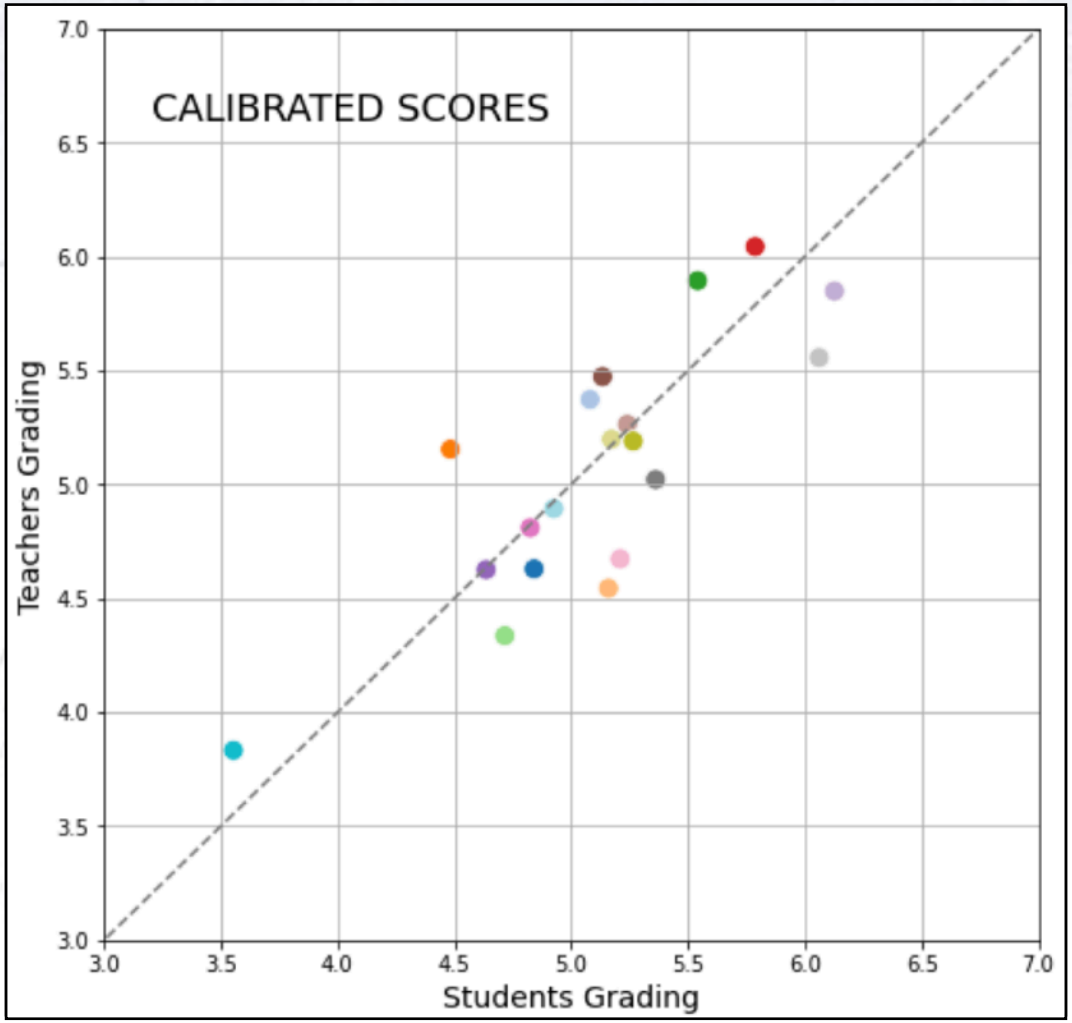
Teachers vs. student average

The correlation can also be seen for the single projects. Generally, teachers used the scale more than the students (average). The correlation is very clear (83%).



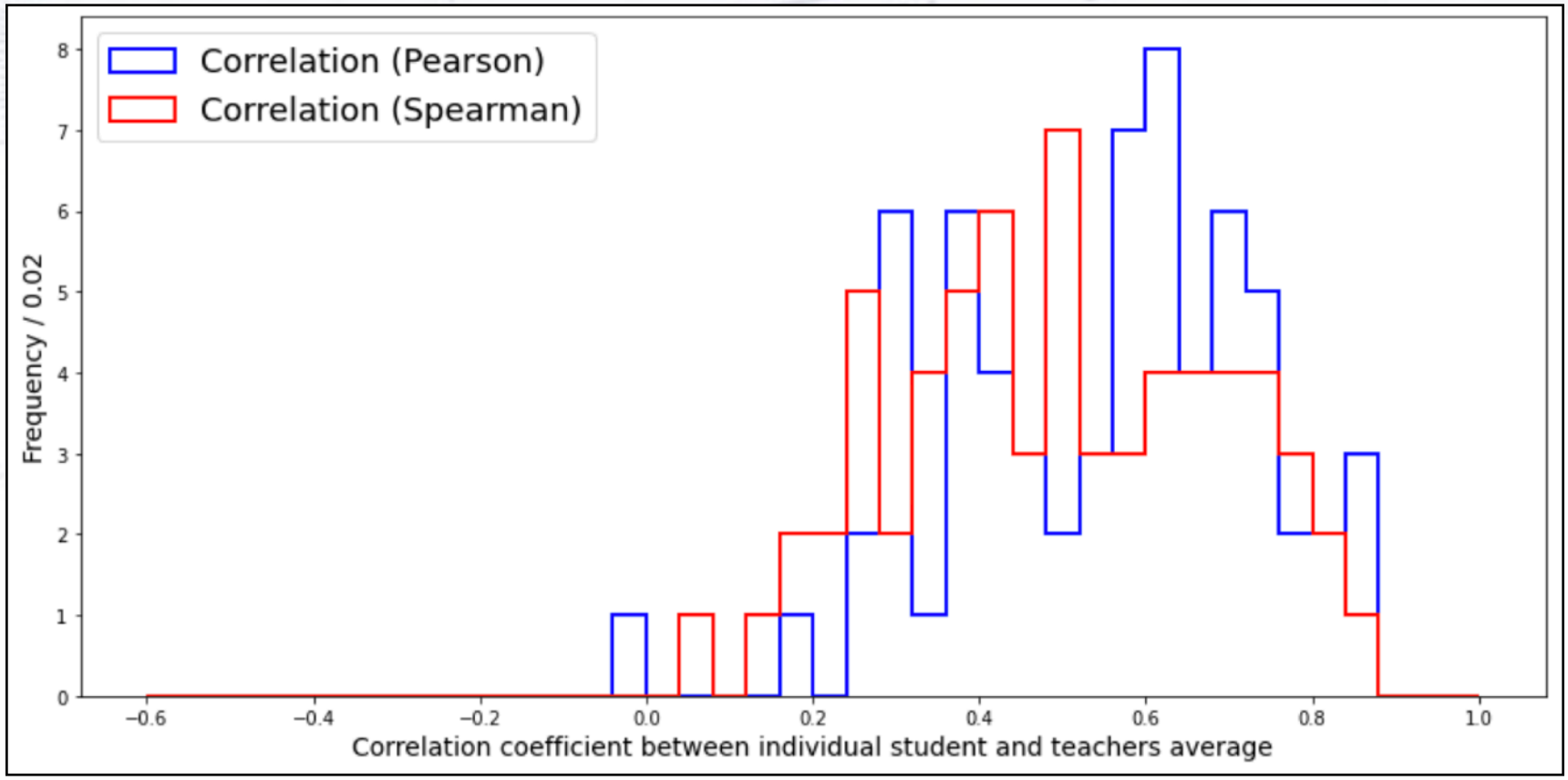
Teachers vs. student average

The correlation can also be seen for the single projects. Generally, teachers used the scale more than the students (average). The correlation is very clear (86%).



Student evaluations

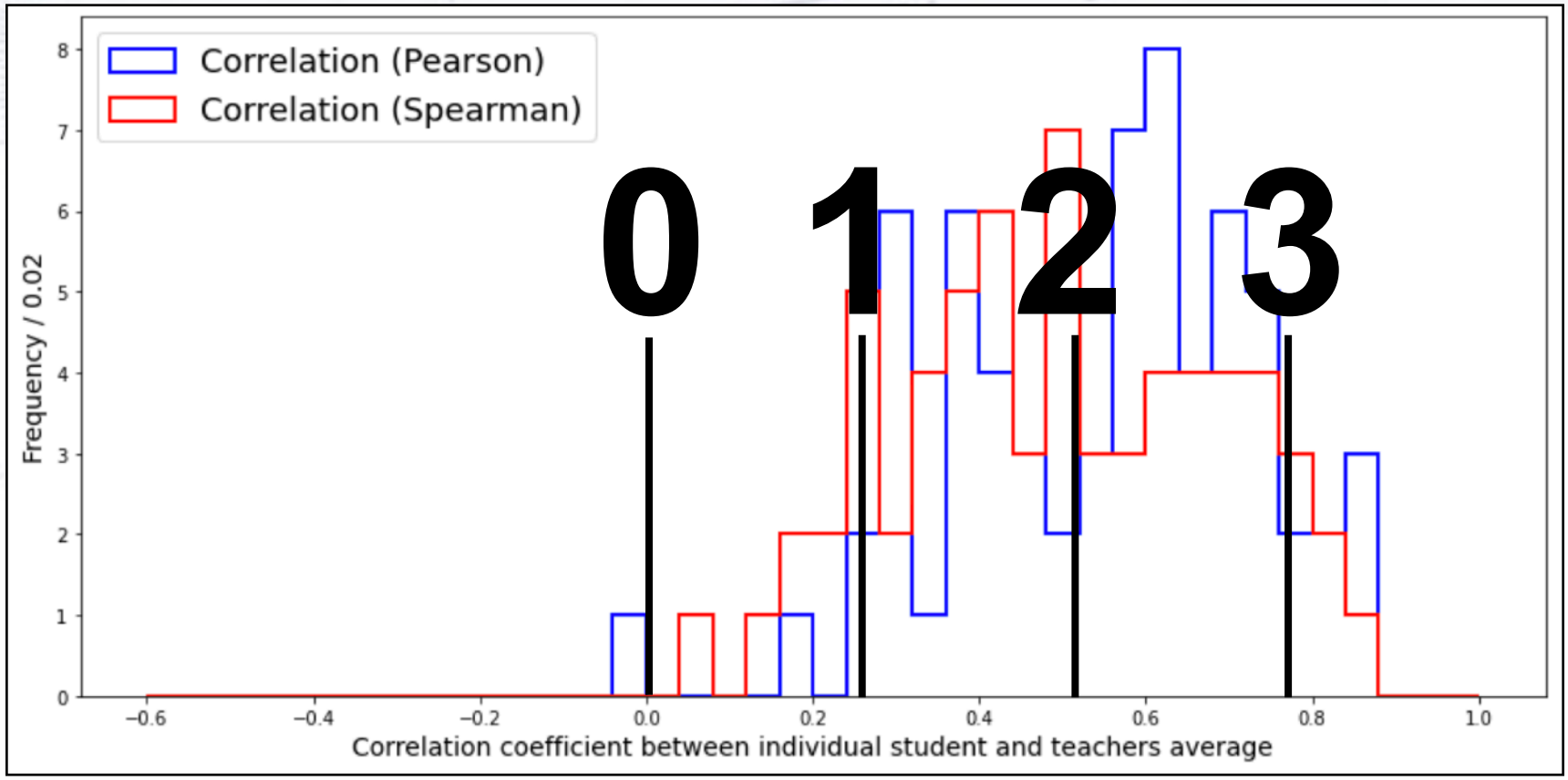
The individual student evaluations were scored by considering their correlation with the teacher's average. Almost all correlations were positive, and generally around 0.4-0.7.



Student evaluations

The individual student evaluations were scored by considering their correlation with the teacher's average. Almost all correlations were positive, and generally around 0.4-0.7.

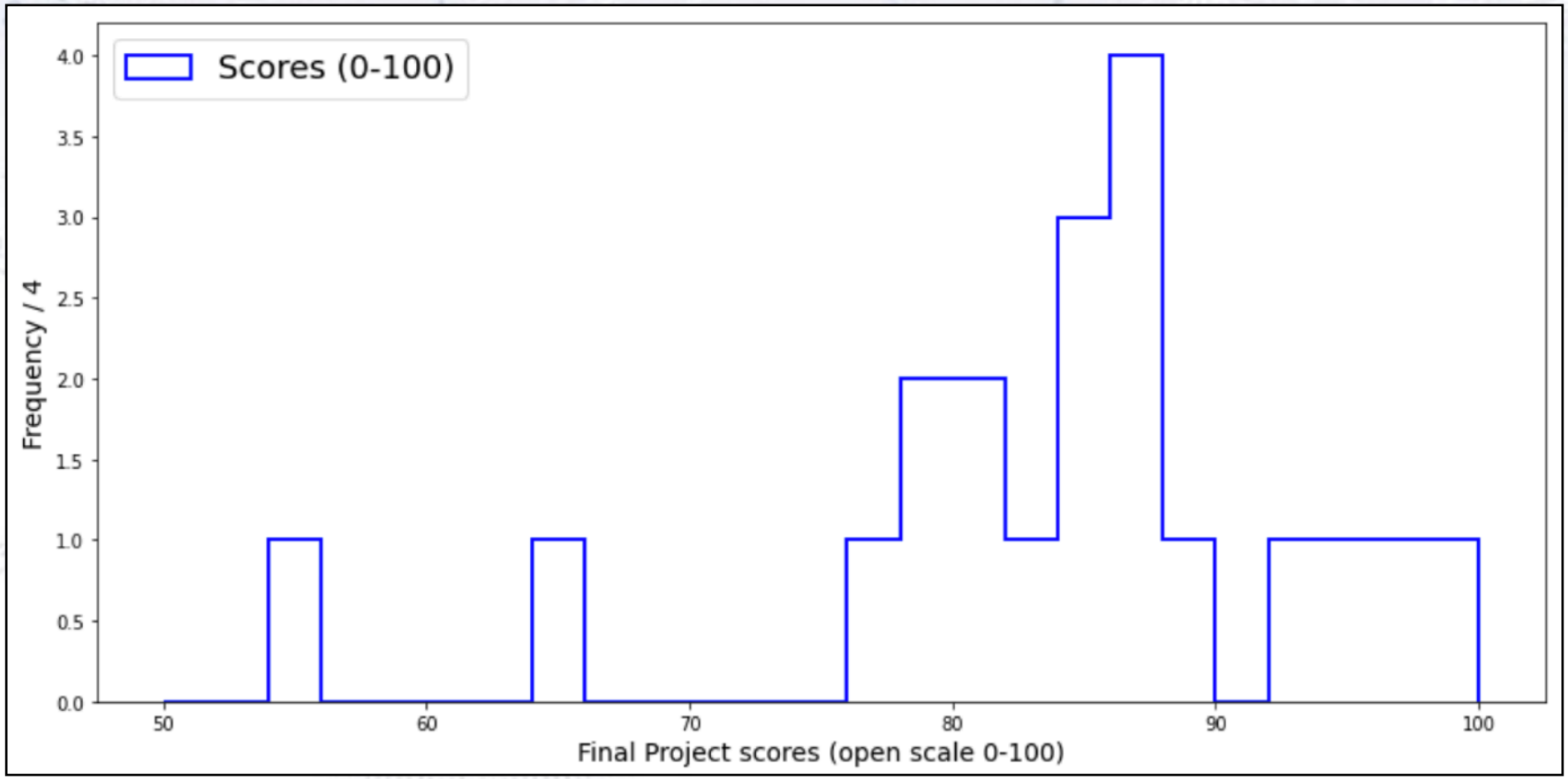
The points given is four times the (Pearson) correlation coefficient.



Final Project Scores

Final project scores

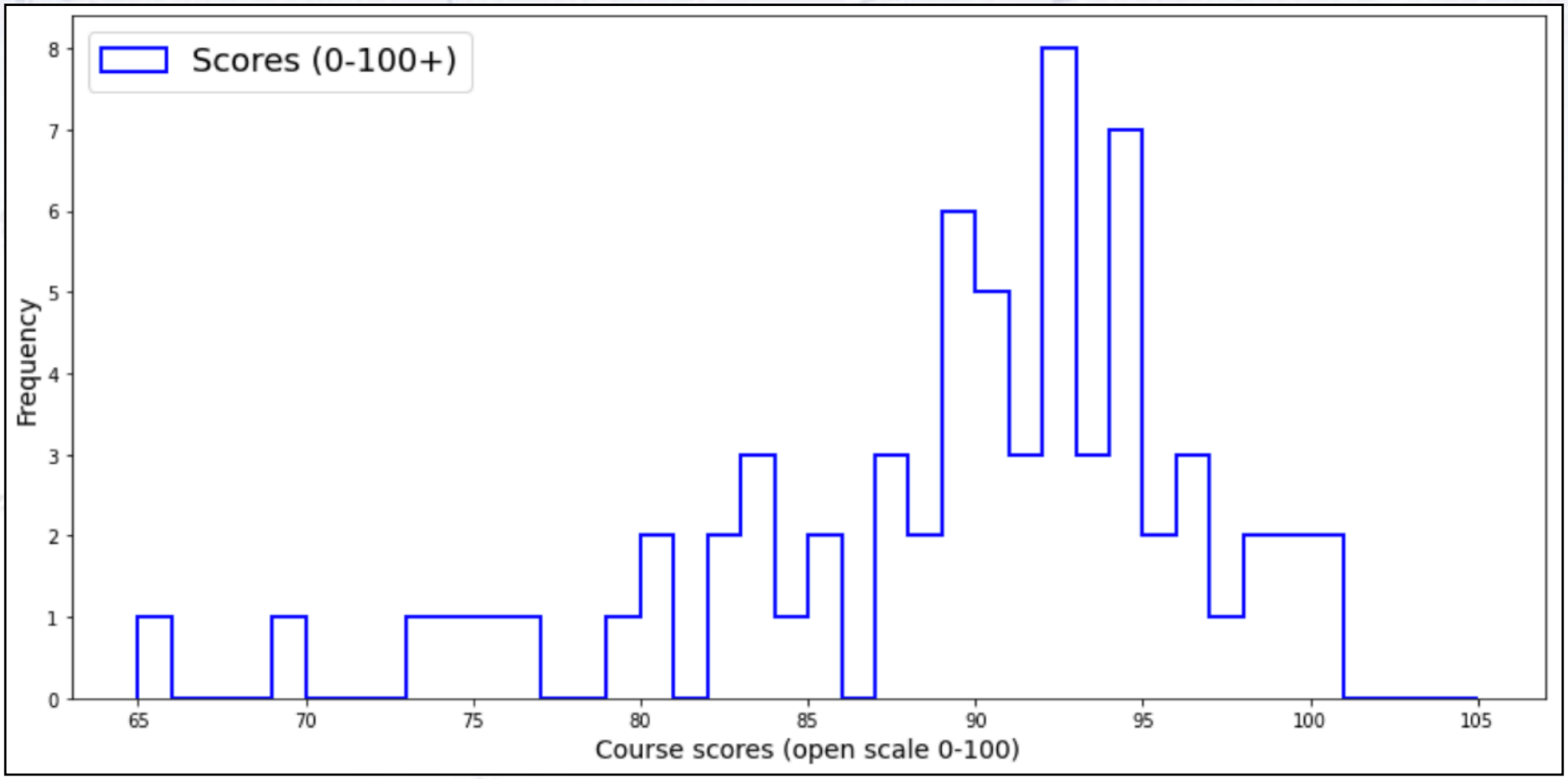
The scores obtained by taking the teachers (calibrated) average multiplied by 14 and then added 14. The final distribution of final project scores is shown below.



Final Course Scores

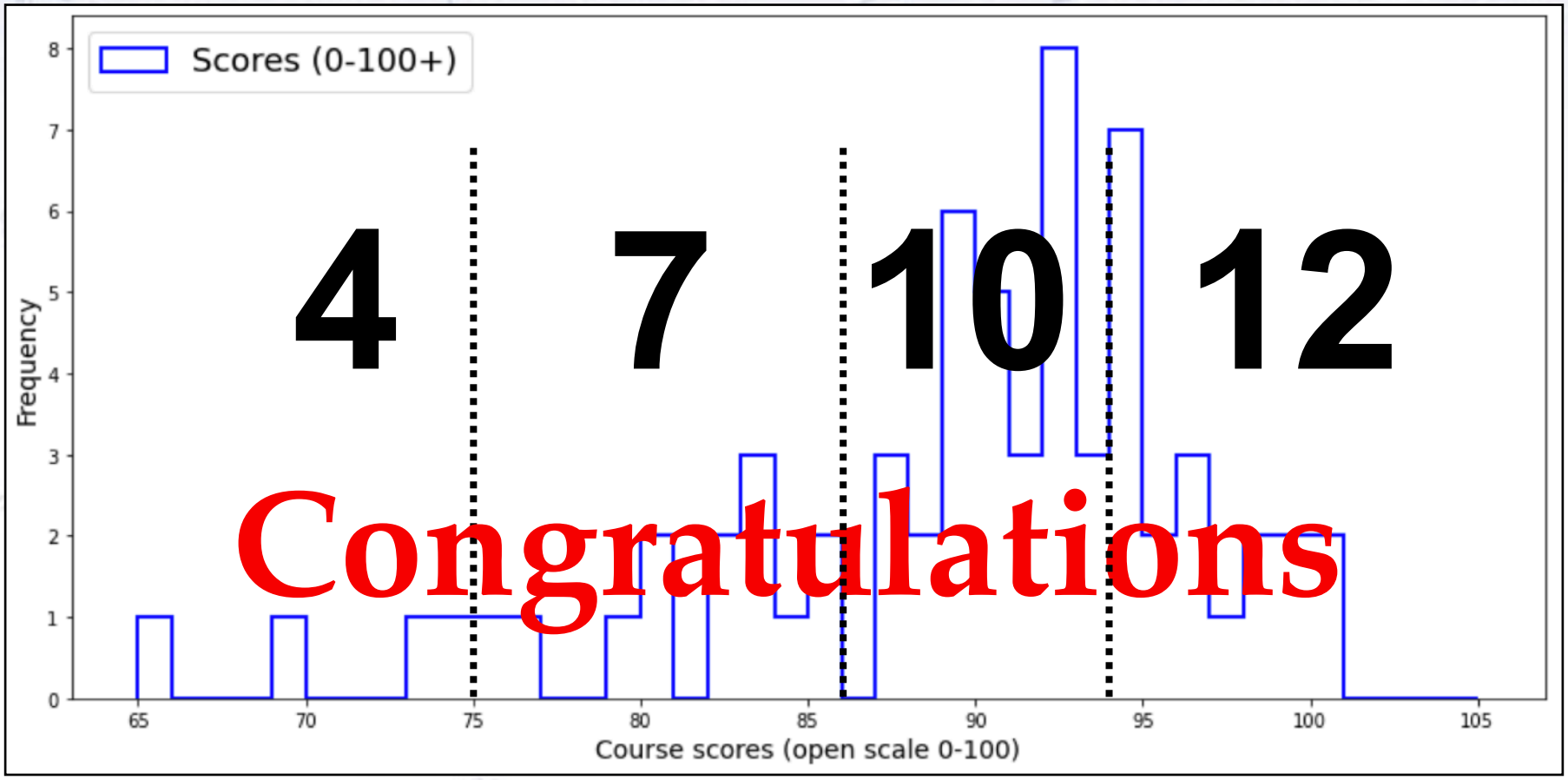
Final course scores

The scores from the **initial project and the final project** (and the ML scoring) were put together as prescribed, and produced the following scores/grades:



Final course scores

The scores from the **initial project and the final project** (and the ML scoring) were put together as prescribed, and produced the following scores/grades:



Our Impressions

You all did very well, which is also reflected in the grade average given!

Three specific impressions:

1. Generally, we felt, that **everybody could actually get ML to work efficiently** and solve problems with it. Super. We hope that this was your impression too.

2. Specifically, you all seem to be able to use BDT/NN on structured data, but also **CNN on images** and **LSTM/GRU on time series**. That is fantastic to see!

3. Many of you have also **worked hard on preprocessing data**, and realised that this is often a tough process. This is very much the case in the real world.

Summary Quote:

"Amazing, what everyone have become capable of in mere 8 weeks."

[Overheard in the exam break]

Thanks to all of you for the past 8 weeks and all your wonderful hard work

AppML Class of 2024