Final exam *versus* projects

1 Introduction

As a complement to the article, we were interested in the comparison between summative evaluations and data science projects. This analysis is complementary to the article. This analysis, we will focus on the following question.

Traditionally, evaluative exams are organized at the end of an academic term. In the case of our data science courses, does a final exam accurately assess the expected learning outcomes as good as the ongoing assessment of students' project where they analyze biological data in practice?

2 Methods

Evaluation grids containing rubrics were used to grade students' projects. These grades were then compared to the results obtained in the traditional exams administered at the end of the term. Pearson correlation coefficients were used to quantify the correlation between the grades in the projects and at the final exams.

3 Results

An exam at the end of an academic term is a common practice as summative assessment. So, we compare grades our students got from such an exam with the score they obtained directly in their projects in 2018-2019 and 2019-2020. The final exam was written in learnr and it mixed a few questions about the theory with many partly solved data analyses that students had to explain, criticize and finalize during the exam session on the computer. The exam was thus largely focused on practical analyses in line with the expected outcomes. Projects represent the reference here as they closely match expected learning outcomes (practical analysis of real-life cases). In this context, the correlation between grades at final exams and in projects is used to estimate if final exams also match expected learning outcomes. Final exams are easier to grade in a shorter time than the projects, but we are not sure whether they evaluate correctly the practical ability students develop during these courses. Stress related to examination and questions out of the broader context that students can only grasp within their projects (e.g., bibliographic research, lengthy exploration of the dataset, group discussions, …) are not feasible during the final exam. This could possibly bias the evaluation.

```
read("../data/sdd_eval.csv") %>.%
  mutate(.,
    course_year = paste0(course, " (", acad_year, ")"),
    course_year = as.factor(course_year)) %>.%
  group_by(., course_year) %>.%
  rstatix::cor_test(., exam, result,alternative = "two.sided", method = "pearson")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   user = col_character(),
##   acad_year = col_character(),
##   course = col_character(),
##   result = col_double(),
##   exam = col_double()
```

```
## )
```

```
read("../data/sdd_eval.csv") %>.%
  mutate(., course_year = paste0(course, " (", acad_year, ")")) %>.%
  chart(., exam ~ result | course_year) +
  geom_vline(xintercept = 5, alpha = 0.3) +
  geom_hline(yintercept = 5, alpha = 0.3) +
  geom_jitter(alpha = 1, width = 0.05, height = 0.05, show.legend = FALSE) +
  ylim(c(0,10)) +
  xlim(c(0,10)) +
  labs(y = "Final exam grade (/10)", x = "Project grade (/10)") +
  theme(aspect.ratio = 1) +
  ggpubr::stat_cor(aes(label = ..r.label..),
    color = "black", geom = "label", label.y = 3.7)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   user = col_character(),
##   acad_year = col_character(),
##   course = col_character(),
##   result = col_double(),
##   exam = col_double()
## )
```

```
## Warning: Removed 1 rows containing non-finite values (stat_cor).
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```

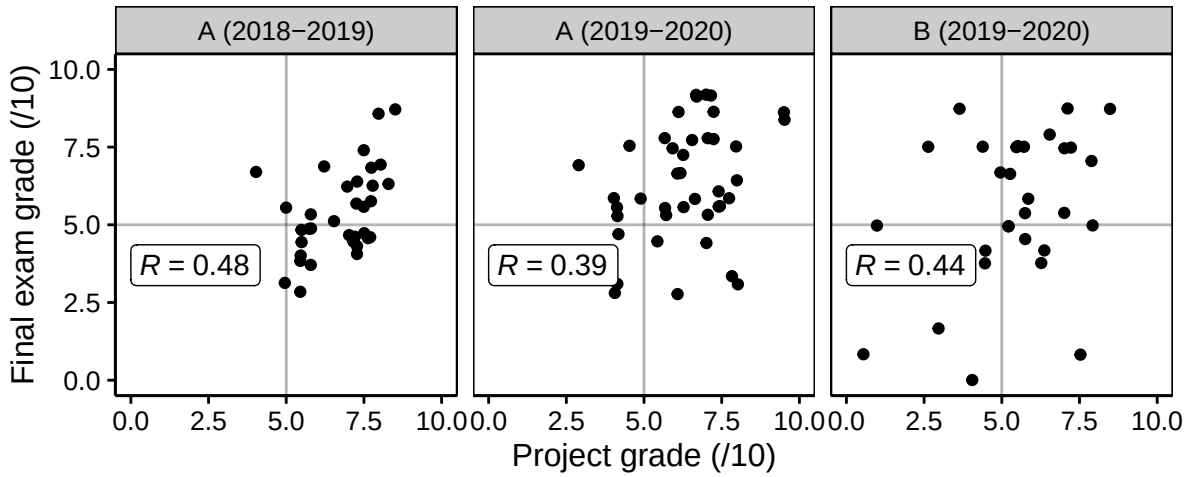

Figure 1: Grades obtained at the final exam versus grades obtained for the projects for courses A and B during two years (course B was still in its old form in 2018-2019 and is thus not represented). Course A (2018-2019) and course B (2019-2020) present the same cohort of students, while course A (2019-2020) is a different cohort. R is the Pearson correlation coefficient.

The comparison of the grades obtained by each student for projects, and final exams shows only a weak correlation between these two types of evaluations (Fig. 1). However, all three comparisons are significantly correlated at the 5% alpha level (Pearson correlation test, A (2018-2019), r = 0.48, t = 3.09, df = 32, p-value = 0.0041 ; A (2019-2020), r = 0.39, t = 2.71, df = 42, p-value = 0.0096 ; B (2019-2020), r = 0.44, t = 2.67, df = 30, p-value = 0.012). In 2018-2019, only one student failed in the projects, while almost one third of them failed their final exam. As the flipped classroom approach allows more time to work in-class on practical

applications, to ask questions and to discuss results, we hypothesize that the very low failure rate in the projects could be explained by better environmental conditions for the students to express their practical data analysis skills. However, in the exam conditions, a larger part of these students produced much less convincing results.

Of course, this discrepancy could also be due to a different level of difficulty between the projects and the final exam. In 2019-2020, we raised the difficulty for the projects, resulting in a more widespread distribution of the results, but with a similar pattern showing only a weak correlation between the two evaluation methods. The same conclusion can be drawn for course B, with several students failing in one of the two evaluations, but not in the other one.

These results suggest that a summative assessment (even including a series of questions involving writing R code to analyze data in the final exam) does not reflect properly the ability of the students to correctly process and analyze biological data, as in their projects, at least in the particular context of our Biological Data Science courses at UMONS. Due to these observations, the final examination was abandoned for the academic year 2020-2021, and it was replaced by an ongoing assessment of the students' activities across all four level exercises. These activities are analyzed in the following section.

4 Discussion & conclusion

The task that better evaluates their practical skills in biological data analysis is the group project because students have to demonstrate what they can do in a situation with a complex dataset and minimal instructions. It meets more closely with the courses' learning outcomes as they have to figure out a suitable question, analyze the data to answer that question, present their results in a report and discuss what they found with a critical mind. This task is complex, especially for undergraduate students. This is why we organize it in groups of two to four students, depending on the complexity of the problem. Here, they mobilize their collective intelligence to get results as many of them would be hardly capable of achieving the task alone. The tracking of individual activity in the project thanks to git allows for figuring out clearly what the contribution of each student was in the group and to score their individual contributions suitably.

We observed a low correlation between performances in projects and in grades obtained at final exams (Fig. 1). This led us to stop using final exams at the benefit of an ongoing assessment with emphasis into projects, despite a final exam is easier and faster to implement. The elimination of the final exams led to the setting up of a procedure to monitor and score the students' activity in all the exercises by automatic scoring online exercises and manual grading of all the projects against evaluation grids. This is time-consuming despite the partially automatic scoring lowers the charge. It would be interesting to investigate whether tools derived from learning analytics and computing science could second teachers in this work. For instance, reproducible research is one of the competences our students have to develop. It implies that their R Markdown documents should compile into reports in HTML or PDF format without any error. This criterion could be checked automatically.