

1 **TEACHING DATA SCIENCE TO STUDENTS IN BIOLOGY**
2 **USING R, RSTUDIO AND LEARNR: ANALYSIS OF THREE**
3 **YEARS DATA**

GUYLIANN ENGELS^{*}, PHILIPPE GROSJEAN

Numerical Ecology Department, Complexys and InforTech Institutes, University of Mons
Avenue du Champ de Mars, 8, 7000 Mons, Belgium

FREDERIQUE ARTUS

Pedagogical Support and Quality Assurance Department, University of Mons
Place du Parc, 20, 700 Mons, Belgium

ABSTRACT. We examine the impact of implementing active pedagogical methodologies in three successive data science courses for a biology curriculum at the University of Mons, Belgium. Blended learning and flipped classroom approaches were adopted, with an emphasis on project-based biological data analysis. Four successive types of exercises of increasing difficulties were proposed to the students. Tutorials written with the R package learnr were identified as a critical step to transition between theory and the application of the concepts. The cognitive workload needed to complete the learnr tutorials was measured for the three courses and it was only lower for the last course, suggesting students needed a long time to get used to their software environment (R, RStudio and git). Data relative to students' activity, collected primarily from the ongoing assessment, were also used to establish student profiles according to their learning strategies. Several suboptimal strategies were observed and discussed. Finally, the timing of students contributions, and the intensity of teacher-learner interactions related to these contributions were analyzed before, during and after the mandatory distance learning due to the COVID-19 lockdown. A lag phase was visible at the beginning of the first lockdown, but the students' work was not markedly affected during the second lockdown period which lasted much longer.

4 1. **Introduction.** In a context where there is an exponentially growing mass of
5 data [31], a reproducibility crisis in Science [4], and a progressive adoption of Open
6 Science practices [5], statistics is broadened to a wider discipline called Data Science
7 [13]. For the Data Science Association, “the Data Science means the scientific study
8 of the creation, validation and transformation of data to create meaning” ([http:](http://www.datascienceassn.org/code-of-conduct.html)
9 [://www.datascienceassn.org/code-of-conduct.html](http://www.datascienceassn.org/code-of-conduct.html)). These changes have also
10 led to the emergence of data science programs in universities and other higher
11 education institutions [15, 12]. One example is the Harvard Data Science initiative
12 (<https://datascience.harvard.edu/about>) launched in 2017. With a broader
13 approach, also comes a broader audience. Such data science courses are not limited
14 to computer scientists, mathematicians or statisticians. Students in humanities,

2020 *Mathematics Subject Classification.* Primary: XXXXX; Secondary: YYYYY..

Key words and phrases. Data Science Teaching, Flipped Classroom, R, RStudio, git.

^{*} Corresponding author: Guyliann Engels.

15 social sciences, and natural sciences also attend them (for instance, the data science
 16 training at Duke University [12]). The focus of such courses is for students to
 17 develop the ability to deal with real datasets in all their complexities, to be able
 18 to conduct reproducible analyses, and to interpret these analyses in the light of
 19 knowledge in their field of expertise.

20 These data science courses pose several pedagogical challenges because numerous
 21 and unfamiliar concepts must be acquired by a heterogeneous class population [22].
 22 Learning objectives span a large range of cognitive abilities and, in these courses,
 23 the intended learning outcomes aim to develop high-level cognitive process abilities,
 24 such as conceptual, procedural, and even metacognitive knowledge [27]. To meet
 25 such learning objectives, active learning methods are useful so that students can
 26 better acquire these high-level cognitive skills [18]. Advances in educational psy-
 27 chology, reviewed by Kirchner and Hendrick [25], give a scientific background to
 28 understand why a pedagogical practice does or does not work. Many teaching and
 29 learning frameworks involving numerical tools turn to a blended learning scenario,
 30 including remote activities to be done before and after in-class work, individual
 31 and group problem-solving, peer instruction and ongoing assessment. The flipped
 32 classroom approach also involves a mix between at home and in class activities, but
 33 learning occurs before the classes that are dedicated to discussions and problem
 34 solving. This allows students to be active in their learning, which has the benefit
 35 of improving student competences [18]. Moreover, this approach enables students
 36 to work at their own pace. Their diverse learning strategies are respected as they
 37 are actors in their own learning process [39]. Such frameworks are open learning
 38 centered and are supported by a varied and rich pedagogical environment [9].

39 The data transformation part of the process in data science is a challenge for
 40 students with little or no background at all in computing sciences. Students that
 41 do not master one or more computer languages enter an unfamiliar world and have
 42 to deal with many exotic concepts, techniques and tools. Version control systems
 43 like git, and their Internet hosting counterparts like GitHub, Gitlab or Bitbucket
 44 are also tools that are taught and used in data science courses [17, 24]. The use
 45 of document formats that dissociate content from presentation, namely LaTeX,
 46 Jupyter Notebook, R Markdown and Quarto to cite but a few, also contribute to
 47 the large number of potentially new tools learners have to discover [6]. On the
 48 other hand, a student in computing science already masters one or more computing
 49 languages. They are acquainted with version control systems, with databases and
 50 with the way data are manipulated and represented on a computer. Yet, the same
 51 students from computing science could have difficulties grasping the context of
 52 the study related to foreign disciplines. A student in mathematics or statistics is
 53 familiar with various concepts that underpin the techniques used to analyze the
 54 data. Students in biology, medicine, psychology, social sciences, economics, etc.
 55 obviously have very different *a priori* knowledge. The gap between knowledge and
 56 learning outcomes generates anxiety (see for instance [33]). The course must thus
 57 be organized in a way that learners progress little by little to avoid exposure to too
 58 many intimidating concepts and tools at once, taking into account their respective
 59 *a priori* knowledge and their initial gaps.

60 Suitable computer hardware and software environments are required to apply the
 61 concepts learned in data science courses. Different approaches range from using soft-
 62 ware accessed from a server [40] (RStudio Cloud (<https://rstudio.cloud/> [36]),
 63 Chromebook data science (<http://jhudatasience.org/chromebookdatasience/>))

64 to local installation on the student's computers. The former requires infrastructure
65 to run the software on a server, and that software is only accessible to the students
66 during the course. The latter raises problems of licensing for proprietary software
67 as well as installation and configuration issues. An intermediary solution uses pre-
68 configured virtual machines, or containers (e.g. Docker) [11, 8]. Such a solution is
69 the most flexible one because it can be deployed almost anywhere (in the computer
70 lab, at home, on a laptop, etc.). As practical applications are important to learn
71 data science [28], the correct choice of software is critical. Early exposure to the
72 tools the students are most susceptible to use later in their work is desirable. This
73 was highlighted by Auker and Barthelmess [3] for instance, for ecologists and by
74 Alvarenga da Silva and Sampaio Moura [2] for physicians.

75 Recently, data science has also been used to analyze the effect of various pedagog-
76 ical practices on the outcome of these courses thanks to learning analytics [16]. A
77 vast amount of data can be collected about students' activities, and the analysis of
78 these data allows the comparison of the impact of different pedagogical approaches,
79 or to quantify and document the impact of such changes in the courses [34].

80 At the University of Mons in Belgium (UMONS), we reworked our biostatistics
81 courses in the biology curriculum in 2018. A series of Data Science courses were
82 introduced, both for our undergraduate and graduate students. The goal of these
83 courses was to train biological data scientists capable of extracting meaningful in-
84 formation from raw biological data. They must be able to do so in a reproducible
85 way and with the correct application of statistical tools and an appropriate critical
86 mind. The goal was thus vastly broaden and it not longer solely targeted skills in
87 biostatistics.

88 The learning outcomes defined for these Biological Data Science courses were
89 thus to be able to analyze most recurrent biological data in practice and to present
90 the results clearly and accurately in a scientific report. In order to achieve these
91 learning outcomes the students had to master skills in biostatistics and scientific
92 writing, and they had to become proficient in the use of computer tools like R,
93 RStudio, git, GitHub and R Markdown. They also had to develop a critical mind
94 in statistical thinking. All of these learning outcomes are described in the students'
95 study programme at UMONS (see for instance for the academic year 2020-2021
96 [41, 42, 43]). A preconfigured VirtualBox machine with R, RStudio, R Markdown,
97 git, and a series of preinstalled R packages was used ([https://www.sciviews.org/
98 software/svbox/](https://www.sciviews.org/software/svbox/)) as a convenient means to deploy the same software environment
99 both on the university computers and on the students' own laptops.

100 As our courses were reworked, we also decided to use flipped classroom and pro-
101 gressive adoption of suitable pedagogical practices in that particular context with a
102 cyclical approach that consisted of stating goals, building pedagogical material with
103 a large emphasis on numerical tools and collection of students' activities, and finally,
104 analyzing the data collected. This approach allowed us to enhance our teaching ac-
105 tivities the following academic year with improved pedagogical techniques. This
106 approach is known as the educational data mining knowledge discovery cycle [34].
107 Here, we present the main results spanning three successive academic years from
108 2018-2019 to 2020-2021, including two particular periods where distance learning
109 was forced due to the COVID-19 pandemic lockdown. In this paper, we will focus
110 on the following three questions:

- 111 • Transition from theory to practice is critical and tutorials built with learnr
112 (<https://rstudio.github.io/learnr/>) are capstones in our courses. What

113 cognitive workload and perceived workload do these tutorials represent for
114 students?

- 115 • How could we use learning analytics to spot suboptimal learning strategies and
116 discriminate different student profiles in our biological data science courses?
- 117 • Did the quick shift from face-to-face to distance learning imposed by the
118 COVID-19 lockdown periods affect our students' production and did it require
119 increased exchanges with the teaching staff to support it?

120 **2. Methods.** This study focuses on three successive courses of increasing difficulty,
121 referred to here as A, B and C and designed as a continuum. These courses were
122 part of the core curriculum and were thus mandatory for all students enrolled on
123 the Bachelor, or Master in Biology at UMONS.

- 124 • Course A was about data preparation, description and visualization. It also
125 introduced inference with most common hypothesis tests in biology (t test,
126 ANOVA, etc.). It was taught in the second year of the Bachelor. Course A
127 was set up to assume only background knowledge acquired by the students
128 during their first year of the same Bachelor in Biology at UMONS.
- 129 • Course B taught data modeling (linear, generalized linear and nonlinear mod-
130 els) and multivariate analyses (PCA, MDS, clustering, etc.) to students en-
131 rolled in the third year of the Bachelor in Biology at UMONS. Course A was
132 a prerequisite to Course B (all students following Course B have previously
133 passed Course A).
- 134 • Course C was taught to all Master students in the Biology section (first year
135 of the Master). This course focused on machine learning, time series analysis
136 and the analysis and visualization of georeferenced data. These students had
137 either passed both Courses A and B at UMONS, or they demonstrated similar
138 knowledge. Only a very few number of students (only one student in 2020-
139 2021 out of a total of 26) came from a different Bachelor and thus had different
140 courses, not including A and B, in their curriculum.

141 The course material was available online (<https://wp.sciviews.org>) and was
142 centralized on a Wordpress site. Students had to login with their GitHub ac-
143 count and their academic data were collected from the UMONS Moodle server
144 (<https://moodle.umons.ac.be>). The courses were broken down into modules that
145 amounted of roughly 15h of work each. There were two in-class sessions of 2h and
146 4h per module (outside the lockdown periods, of course). There was roughly 3h of
147 preparation at home before each session, and 3h of work to complete one module.
148 The main activities in the class were analysis of actual data (projects). Students
149 also asked questions and followed brief lectures (15 minutes) on selected topics in
150 the class. They had to propose and vote for the topics to be covered during these
151 short lectures. Finally, we encouraged students to help each other and to explain
152 what they understood to their colleagues. Indeed, students' questions were some-
153 times redirected by the teacher to other students that had already mastered the
154 topic. Teachers rarely answered questions directly. When it was possible, they
155 rather proposed new tracks or ideas to investigate and helped learners to find the
156 solution themselves. Students who went through the activities before the others
157 were also encouraged to help their slower colleagues.

158 Regarding the timing, one module was taught every other week so that students
159 had enough time to prepare the material at home before the in-class session, and
160 after it, to finalize their projects for the module. As a term is made of 14 weeks, we

161 did not teach more than six modules in a course unit to avoid teaching too much in
162 a short time. After reading the theory, students were exposed to exercises of four
163 increasing levels of difficulty. They thus had to apply the concepts repeatedly but in
164 different contexts, which broke any monotony and maintained a stimulating rhythm
165 all along their progression. They had to learn the principles in the online book
166 (<https://wp.sciviews.org>) and self-assessed their comprehension of the concepts
167 using H5P (<https://h5p.org>) exercises (Level 1 difficulty). These exercises were
168 simple questions (TRUE/FALSE, multiple choice, etc.). Learnr tutorials (<https://rstudio.github.io/learnr/>)
169 were used for the Level 2 exercises. They were
170 gently introduced the students to the R code required for the analyses and guided
171 them step by step through their first data analysis. These tutorials were thus the
172 entry point to the practice.

173 Most of the practical work was dedicated to GitHub projects (Level 3 and Level
174 4 difficulties). At this stage, the use of R instructions was not sufficient to complete
175 the exercises. Students had to also become acquainted with git, GitHub, R Mark-
176 down and RStudio to manage the projects. They also had to interpret the results
177 they obtained. The individual projects (Level 3) contained guidance on how to per-
178 form the different steps of the analyses. The group projects (Level 4, groups of two
179 or four students) did not contain such guidance. At Level 3, the goals were clearly
180 specified in the projects. At Level 4, students had to imagine suitable biological
181 and statistical questions that could be answered by analyzing the data proposed in
182 the project. Working on these projects represented both the core of in-class activ-
183 ities and the best expression of their learning progression. By construction, Level
184 1 to Level 4 exercises were built according to their increasing cognitive difficulties
185 following bloom's taxonomy [27].

186 All student activities in H5P exercises (self-assessing), and in the learnr tutorials
187 (transitioning smoothly from theory to practice) were recorded in a MongoDB data-
188 base. The {learnitdown} R package (<https://www.sciviews.org/learnitdown/>)
189 provided the code required to manage user login, user identification and activity
190 tracking for this interactive material.

191 Projects containing the data, the analyses and the reports were hosted in GitHub
192 repositories. These repositories were cloned and edited by the students in their vir-
193 tual machines (SciViews Box) with RStudio ([https://www.rstudio.com/products/](https://www.rstudio.com/products/rstudio/)
194 [rstudio/](https://www.rstudio.com/products/rstudio/)), either on their laptops or on the computers in the lab. We encour-
195 aged our students to install the virtual machine for the course on their own com-
196 puter so that they were able to work comfortably at home and could also use
197 it for other activities too. Assignment and creation of the GitHub repositories
198 for each student, or group of students, was orchestrated by GitHub Classroom
199 (<https://classroom.github.com>). Reports were written in R Markdown (<https://rmarkdown.rstudio.com/>),
200 a file format that combines the prose with R code to
201 produce analysis results, plots and tables directly inside the documents. All reposi-
202 tories were ultimately cloned by the teacher in a centralized area on our servers and
203 data about commits (git logs) were collected using git version 2.31.1 and R version
204 4.0.5 [35]. To give an idea of the data recorded in 2020-2021, we had just over 3,500
205 events that were recorded for each student.

206 In distance learning, student support was done via email and Discord (<https://discord.com>).
207 At the end of an academic term, all recorded messages were
208 collected into text files. These files were scraped using custom R code to create
209 a table with key information (basically, who, when, and what) for each message.

TABLE 1. Four levels of increasing difficulties in the exercises.

Level	Description	Type
L1	Interactive exercise in the course, direct feedback	h5p
L2	Tutorial with guided exercises, feedback and hints	learnr
L3	Individual and guided data analysis	individual project
L4	Free data analysis and reporting (by 2 or 4 students)	group project

210 Surveys were done periodically in class through Wooclap questionnaires (<https://www.wooclap.com>). Such questionnaires were used to query perceived workload
 211 of the learnr tutorials. Results were manually exported out of Wooclap by means of
 212 Excel files. These data were then incorporated into a table in our database thanks
 213 to an R script.

215 Data about users, courses, lectures and projects, as well as grading items (on
 216 average, more than 130 grading items were established for each student in 2020-
 217 2021) were pseudonymized: names, emails and all the personal information were
 218 replaced by random identifiers. The different tables were ultimately exported into
 219 CSV files and made public [19]. Data collection, treatment and use respect the
 220 European GDPR (General Data Protection Regulation) since each student had to
 221 agree explicitly with the way data were collected and used (including the research
 222 purpose) before each course started. They were able to visualize their data through
 223 personalized reports at any time.

224 The course material was organized in a way that favored autonomy and self-
 225 assessment (direct feedback in the exercises, hints and retry buttons in case of wrong
 226 answers). Table 1 summarizes the main characteristics of the exercises according to
 227 the difficulty level.

228 R and tidyverse [45] packages (<https://www.tidyverse.org>) were used to pre-
 229 pare the data and the analyses. **The pseudonymized data are available in**
 230 **Zenodo [19, 20].** A GitHub repository with the code used to create the figures and
 231 table in this paper is available at [https://github.com/BioDataScience-Course/](https://github.com/BioDataScience-Course/teaching_data_science_in_biology)
 232 [teaching_data_science_in_biology](https://github.com/BioDataScience-Course/teaching_data_science_in_biology).

233 **2.1. Measured and Perceived Cognitive Workload in learnr Tutorials.** The
 234 average number of trials that were required for each student to find the right an-
 235 swer in learnr tutorial exercises was used as a proxy of measured workload. In
 236 comparison, the perceived cognitive workload was established with a NASA LTX
 237 questionnaire. This questionnaire is composed of six questions on a Likert scale [23].
 238 The questions concern mental load, physical load, time pressure, expected success,
 239 effort required, and frustration experienced during the accomplishment of the task.
 240 The average value for the six questions constitutes a Raw Task Load indeX (RTLX)
 241 [10] that we used to quantify how students felt when using these learnr tutorials.
 242 An analysis of variance test and a Tukey’s post-hoc Honest Significant Difference
 243 (HSD) tests were used for the comparison between courses.

244 **2.2. Students Activity Profiles with Ongoing Assessment.** Data from the
 245 L1-L4 exercises and the student support (email and Discord) were used to char-
 246 acterize the students’ activity profiles. A non-supervised classification technique
 247 called a Self-Organizing Map (SOM) [26] was used to characterize various learning
 248 profiles. The {kohonen} R package was used to compute the model [44].

TABLE 2. Number of students, modules, and exercises for each course. For the learnr tutorials, the first number is the amount of tutorial documents and the second number in brackets is the total number of questions in these tutorials (year 2020-2021).

Course	Students	Modules	H5P	Learnr	Indiv. projects	Group projects
A	59	12	59	24 (211)	10	4
B	45	8	29	11 (108)	12	2
C	26	6	19	7 (37)	7	1

249 **2.3. Transition from Face-to-face to Distance Learning Imposed by the**
 250 **COVID-19 Lockdown.** The transition between face-to-face and distance learning
 251 was studied through the contribution of each student to projects with the commits
 252 (git logs) and by these contributions were then divided by the questions they asked
 253 by email and on Discord. **Though** only descriptive analysis of these data was done,
 254 interesting patterns were observed.

255 **3. Results.** This study was performed on data related to the three successive
 256 courses that comprised 26 modules in total in 2020-2021. Table 2 summarizes the
 257 number of H5P, learnr, individual and group GitHub projects that students had
 258 to complete. Group projects usually spanned over several modules. It should be
 259 noted that for Course C, we also introduced a challenge in machine learning that
 260 replaced one GitHub group project. This challenge is omitted from the present
 261 analysis, being a unique activity that is difficult to compare to the rest. However,
 262 this explains why there was only one group project in Course C.

263 **Retrospective data 2019-2020** [20] were also used when pertinent. It should
 264 be kept in mind that the pedagogical material was written and improved progres-
 265 sively over the three academic years. The 15 years of experience gathered all along
 266 these former courses was critical in the redesign of the new ones. The H5P exercises
 267 and the auto-checking of learnr answers were not available before 2020-2021.

268 **3.1. Measured and Perceived Cognitive Workload in learnr Tutorials.** In
 269 our courses, learnr tutorials helped students to transition from the theory (online
 270 book chapters) to practice (projects). These tutorials are online interactive doc-
 271 uments that recall main concepts, and take the students by the hand to perform
 272 their first data analysis step by step. At each step, they have at least one exercise
 273 or one quiz. The exercise consists of writing R code, or filling missing parts in R
 274 code to progress through the analysis.

275 Our goal with these tutorials was to optimally prepare the students for the prac-
 276 tice of data science. The usefulness of these tutorials was qualitatively determined
 277 by observing the behavior of the students when they started their practical work.
 278 The number of retries necessary to complete an exercise on average, the number of
 279 exercises correctly answered, or the time needed to complete one tutorial are quan-
 280 titative measurements and could be analyzed in order to optimize these tutorials.

281 A few tutorials were elaborated during the academic year 2018-2019, and positive
 282 feedback on their utility (both from direct observation of the students and thanks to
 283 their remarks) led us to systematize them into what we now call Level 2 activities
 284 (see Table 2) in the form of learnr documents in 2019-2020. The tutorials were
 285 further refined in 2020-2021: we added contextual hints with the {gradethis} R

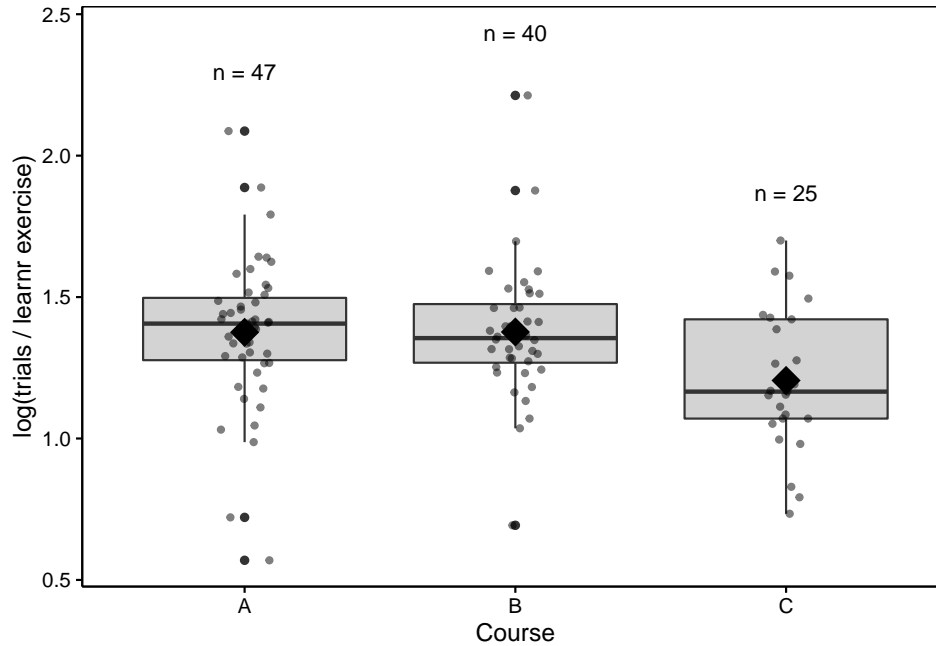


FIGURE 1. Logarithm of the average number of retries that were required for each student to find the right answer in learnr tutorials exercises (2020-2021 academic year). This measure is used as a proxy to quantify the cognitive workload (with caution as explained in the text). **The big black diamond is the average for the whole class. The small black dots are the outliers of the boxplot. The number above each box is the number of students.**

286 package (<https://pkgs.rstudio.com/gradethis/>). In their latest version, when
 287 students submit their answer to the exercises, the R code is parsed, analyzed and
 288 the result is compared with the solution. In case of differences, heuristics are used
 289 to provide contextual hints. Students can then refine their solution and resubmit
 290 it. This appears very efficient in self-learning and self-assessing their competences
 291 before switching to the practice with confidence.

292 A fully objective quantitative measurement of the cognitive workload is near
 293 impossible to obtain in these asynchronous activities done at home. It was thus
 294 estimated by using a proxy: the logarithm of the average number of trials that were
 295 required for each student to find the right answer in the learnr tutorial exercises
 296 (Fig. 1). Caution is required here as a high number of trials could also be the result
 297 of students that are just guessing. However, answers being pieces of R code, pure
 298 guessing most probably leads to nothing useful. A certain level of understanding of
 299 both the R syntax and the question are required to obtain correct answers. Only
 300 data from students that correctly answered most of the exercises ($> 90\%$) were used
 301 here, as it also rules out the students that appeared to have insufficient knowledge
 302 to master the concepts in the tutorials and were probably just guessing. For Course
 303 A, 12 out of 59 students did not pass this filter, 5 out of 45 for Course B and 1 out of
 304 26 for Course C. **The averages and standard deviations of logarithm of the**

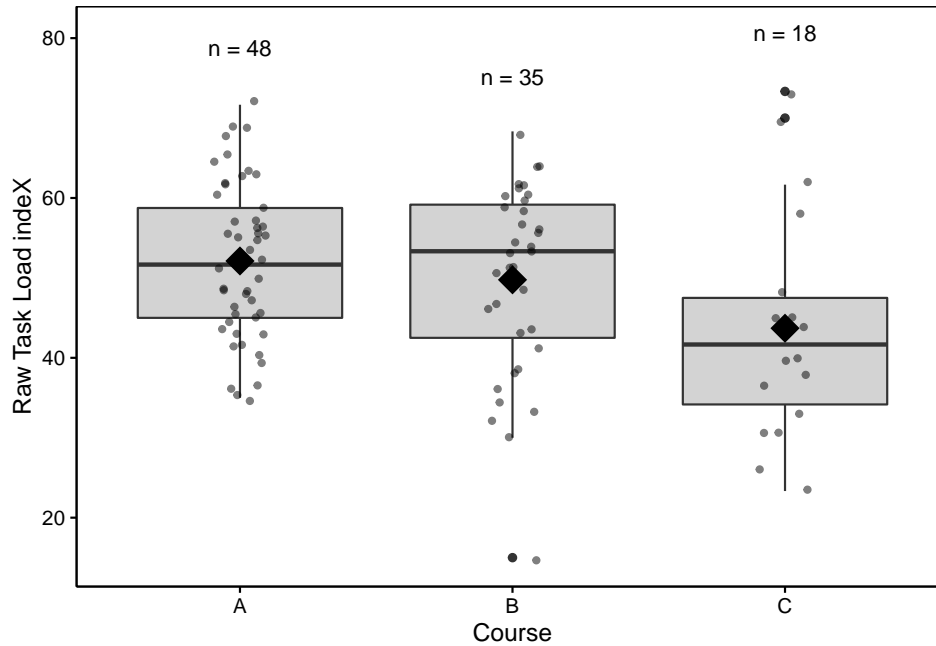


FIGURE 2. Perceived workload for the learnr tutorials in the three courses (year 2020-2021). The big black diamond is the mean RTLX value. **The small black dots are the outliers of the boxplot. The number above each box is the number of anonymous respondents.**

average number of retries are 1.375 ± 0.260 for the Course A, 1.377 ± 0.239 for the Course B and 1.205 ± 0.252 for the Course C. At least one course differs significantly from the other two courses (ANOVA, $F(2, 109) = 4.49$, $p\text{-value} = 0.013$). The homogeneity of variances (Bartlett Test, $K^2 = 0.30$, $df = 2$, $p\text{-value} = 0.86$) and the Normal distribution of the residuals using a quantile-quantile plot were verified. The students on Course C need significantly fewer trials to find the right answer than students on Courses A at α level of 5% (Tukey HSD, $t = -2.74$, $p\text{-value} = 0.019$) and B (Tukey HSD, $t = -2.68$, $p\text{-value} = 0.023$).

The perceived cognitive load required to perform these exercises was also determined on the same students and for the same exercises. The Raw Task Load indexX (RTLX) measured the emotional state of the students after having completed a tutorial. This has, as far as we know, not yet been done. We used a NASA LTX questionnaire to assess it across all three courses. Participation in the survey was high: 48/59 (81%), 35/45 (78%) and 18/26 (69%) for Courses A, B, and C respectively.

The averages and standard deviations of RTLX are 52.12 ± 9.70 for the Course A, 49.76 ± 1.90 for the Course B and 43.70 ± 14.19 for the Course C. The difficulty of the course, and thus, of the exercises in the tutorials increased from one course to the other. However, we did not observe any increase, neither in the number of retries, nor in the RTLX (Fig. 2). On the contrary, these

appeared significantly lower for Course C than for Course A at the α level of 5% (ANOVA, $F(2, 98) = 3.59$, p-value = 0.031; Tukey HSD, $t = -2.68$, p-value = 0.023). The homogeneity of variances (Bartlett Test, $K2 = 4.17$, $df = 2$, p-value = 0.12) and the distribution of the residuals using a quantile-quantile plot were verified and indicated that they did not depart significantly from a Normal distribution. The cognitive load perceived by the students diminished at the same pace as their ability to find the right answer in fewer trials. This may be a consequence of a more fluent R coding ability and a better mastery of the software environment.

3.2. Student' Activity Profiles with Ongoing Assessment. The flipped classroom approach and the proactive attitude we expected from our learners (they had to formulate questions correctly whenever they faced a problem) led to different and contrasted learning strategies. Not all students asked questions. Some of them tried to find solutions on their own. Others preferred to ask their questions privately, while others had no problems exposing their difficulties in a public forum (the Discord channel dedicated to the course). The way and the timing learners progressed in the exercises also varied largely. The schedule was not tight and only suggested a rate of progression. No students were penalized if the exercises were done later, as long as they were completed before the deadline. As expected, a part of our students preferred to stick to the proposed schedule, while others procrastinated and delayed the completion of their exercises. Some strategies are probably more efficient than others. We analyzed the records of the students' activities to distinguish the different learning profiles and we compared them with the grade they obtained at the end of the course.

In 2020-2021, to support the ongoing assessment without a final exam, the activity of each student in Level 1 (H5P) and 2 (learnr) exercises was exhaustively recorded in a database. For the GitHub projects (Levels 3 and 4 exercises), the GitHub repositories and the git log data were analyzed. During the lockdown periods, exchanges with students and answers to their questions were exclusively done by email, text or voice messages on Discord, or on private or public channels. Students were allowed to freely choose their favorite tool to interact with the teachers and between each other. All these exchanges were recorded too.

The degree of completion of all the exercises was used to establish the final grade for the course, with a much higher weight on individual and, especially, on group projects. The weight was adjusted from course to course according to the importance of the different projects. To give an idea, for Course A during the second term, Level 1 H5P exercises accounted for 5%, 10% for Level 2 learnr tutorials, 35% for Level 3 individual projects and 50% for Level 4 group works. On average, each student received more than 130 assessment items that accounted for their final grade. Two thirds of these assessments were established manually, using evaluation grids based on their work in the various projects. The remaining third was made of scores automatically calculated from the online exercises.

For the three courses, we recorded more than 450,000 events, which makes on average almost 3,500 events for each student. These data contain information that we used to characterize the behavior and learning patterns that the students exhibited. They are summarized into sixteen metrics.

For H5P exercises:

- trials/H5P ex.: the average number of trials for each H5P exercise until the right answer is found (students can retry as many times as they wish and they have immediate feedback on whether or not their answer is correct),
- correct H5P ex.: the fraction of H5P exercises that were answered correctly.

For learnr tutorial exercises:

- trials/learnr ex.: the average number of trials for each learnr exercise until the right answer is found (here also, students can retry as many times as they want), excluding quizzes,
- hints/learnr ex.: in learnr exercises, students can display hints to help them solve the problems (but they lose 10% of the exercise score for each hint they reveal). This is the average number of hints per exercise that were displayed for each student,
- correct learnr ex.: the fraction of learnr exercises that were completed with a correct answer,
- time/learnr ex.: the average time required to finish one learnr exercise involving R code writing, thus excluding quizzes.

For individual and group projects:

- commits/ind. projects: the average number of commits done by a student in one individual project,
- contributions/ind. projects: the number of lines changed -added or subtracted- in the R Markdown reports by one student in one individual project (this includes embedded R code for the processing, analysis and plotting of data),
- commits/group projects: same as above, but for group projects,
- contributions/group projects: same as above, but for group projects,
- percentage of contributions to group projects: the fraction of work the student did, relative to all the work done in group projects.

For support:

- questions/module: the number of questions students asked, divided by the number of modules in the course,
- percent of public questions: the fraction of questions that the student posted in a public channel (the Discord channel dedicated to the course that all the other students of the class can read),
- contributions/question: a metric that catches the relative “productivity” of the student related to the number of questions they ask.

Finally, global measurements:

- work done: the fraction of all exercises that the student completed,
- work done in time: the fraction of exercises done in the right time, that is, within the proposed schedule.

In our courses, we have a few students in mobility that come from various origins. The *a priori* knowledge is important in education. So, to avoid biases due to the past curriculum of the students, we restricted this analysis to the subpopulation that comes from the first year of the Bachelor in Biology at UMONS only. A Kohonen’s self-organizing map was used to create student profiles according to their activities (Fig. 3). A three-by-three hexagonal map was chosen, and students were thus classified into nine different groups.

In Fig. 3, the small peripheral plots in gray scale show how selected metrics distribute in the nine cells, from lowest value in white to highest value in black.

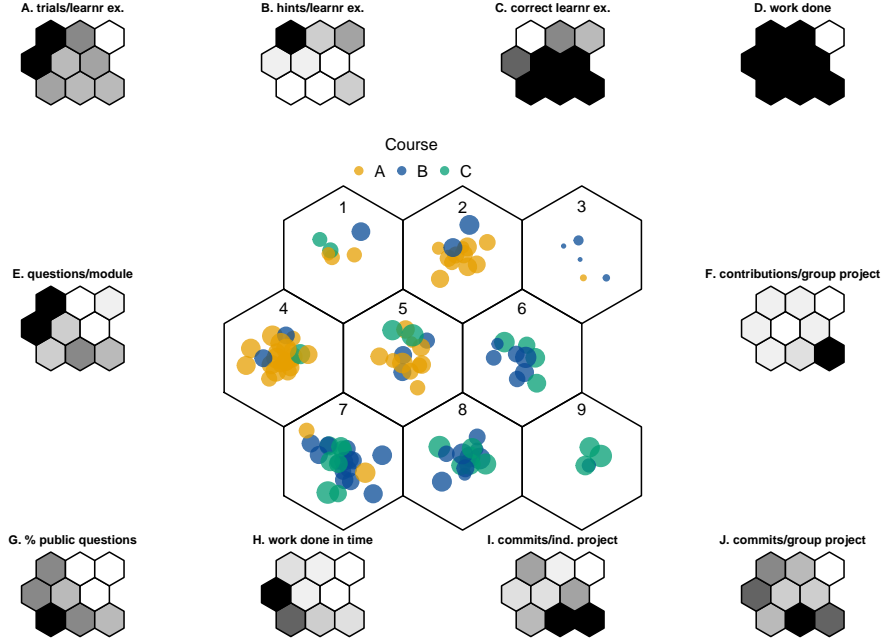


FIGURE 3. Self-organizing map of the student activities across the three courses (year 2020-2021). See the text for the explanations.

They help to decipher the way students behave according to their profile. Metrics that are not represented in the figure exhibit similar patterns to others (for instance H5P metrics have a similar pattern to learnr metrics). **The codebook vectors of the self-organizing map is presented in the Appendix. These codebook represent the importance of each variable on each cell is presented in the Appendix.** Dots in the central plot are the various students, with the color representing the course and the diameter of the dots indicating the grade the students obtained at the end of the course. The following paragraphs detail the information in this figure. The numbers between parentheses are the cell numbers in the central plot, and the uppercase letters in parentheses refer to the peripheral subplots.

Each cell (1-9) represents a learning strategy used by the students. **The metric used to determine the effectiveness of the learning strategy is the score. A very high grade is associated with an optimal learning strategy and a low grade with a poor learning strategy.** Strategies that are considered as suboptimal are not only related to the grade but they can be related to the time students need to learn (e.g., large number of trials per exercise) or to a lack of interaction in the group (e.g., very few questions asked).

Although most students completed all, or almost all of the exercises (D), Cell (3) collects the few students that did only a tiny part of these exercises. These students obtained very low grades. They belong to Courses A and B. On the other hand, heavy workers are at the bottom (I & J), and good performers in learnr tutorials (C) are in Cells (5-9).

- 441 • Cells (2) and (6) collect students that seldom asked questions (E), and that
 442 rarely appeared on the public channel (G). Minor differences separate them.
 443 For instance, learners in Cell (2) sometimes used hints (B), while those in
 444 Cell (6) never did so, also because they found the correct answer to the ex-
 445 ercises more often by themselves (C). Asking questions is at the core of our
 446 pedagogical approach. So, these students did not play the game. However,
 447 they were possibly successful anyway. Some of them probably exchanged with
 448 other students through different channels that we did not monitor. Cell (2)
 449 -more difficulty with learnr tutorials- mainly contains students belonging to
 450 Course A, while cell (6) contains students of Courses B and C. There is a
 451 clear evolution in their behavior from one course to the other in terms of
 452 ease in carrying out the exercises, even though they remained silent at the
 453 teacher-learner interaction level.
- 454 • Among the students that had a hard time figuring out the answers to the
 455 auto-evaluation exercises, Cell (1) reassembles people that most heavily relied
 456 on hints (B), and were among those who had to retry the exercises more
 457 often before they worked out the correct answer (A), a characteristic they
 458 share with Cell (4). These students also asked a lot of questions (E), both
 459 on the public and private channels (G, mid gray indicating a balance between
 460 public and private messages). The main difference between these two groups
 461 is that students in Cell (4) tried harder to find the answer without looking at
 462 the hints, while in Cell (1) they gave up more rapidly. Also these students
 463 respected the proposed schedule much more closely than all others (H). We
 464 have students coming from all courses there, but a majority from Course A.
- 465 • Cells 1-4 plus 6 contain students that exhibit suboptimal behaviors in one or
 466 the other way. The remaining Cells (5, 7-9) correspond to learner profiles that
 467 perform better from this point of view. Cell (5) is primarily represented by
 468 students from course A, but secondly, also from Courses B and C. These are
 469 average actors in all metrics, except they are fluent with Level 1 (H5P, not
 470 shown) and Level 2 (C) exercises.
- 471 • Moving from Cell (5) to (7), (8) and (9), we encounter increasingly good
 472 performers. The number of students from Course A becomes progressively
 473 lower, while Course B and, especially C, dominate in these groups. In Cell (7),
 474 they intensively used the public channel (G) and also respected the schedule
 475 quite well (H) as main differences from those from Cell (5). Students in
 476 Cells (8) and (9) were not so often in time, but this is because they were
 477 heavier workers in the projects, both in the individual (I) and in the group
 478 (J) activities. This obviously needed more time. In cell (9) we also find the
 479 students that contributed the most to the reports in terms of lines added or
 480 deleted (F).

481 To summarize, at the top of the SOM map, Cells (1-4, plus 6) contain students
 482 with suboptimal behaviors, Cell (5) are average students, and Cells (7-9) at the
 483 bottom exhibit profiles corresponding to the best performers. The pattern is also
 484 visible between Courses A (mainly distributed at the top or center of the map) to
 485 B and C (more represented at the bottom). This probably suggests that students
 486 needed time to get used to the course, its pedagogical approach, and/or the soft-
 487 ware environment they had to use. Since only a small fraction of the participating
 488 students failed, excluding the failing ones in Cell (4), the intercourse pattern can

hardly be explained by a filter that eliminated the low performers from one course to the other.

3.3. Transition from Face-to-face to Distance Learning Imposed by the COVID-19 Lockdown. Due to the COVID-19 lockdown periods, distance learning had to be adopted abruptly. We analyzed the activity collected during the 2019-2020 and 2020-2021 academic years to **describe** the impact of these transitions on the progression of the students. In Fig. 4, the academic term is divided into seven work periods of approximately two weeks each (note that this was the suggested rate of the courses: one module every other week). The classes of the second term started in period Y1P09, since period Y1P08 was reserved for the first term exams session. The courses of the first term of 2020-2021 began in Y2P01. First lockdown started in period Y1P11 for one month and a half. Second lockdown started in Y2P03 and lasted to the end of the second term (Y2P15). During the first lockdown, we quickly opened the dedicated Discord channels that were available without any latency.

Contributions per student (Fig. 4a) was relatively constant during the second year, starting essentially in Y2P03, when the second lockdown was established. The highest activity is observable at the end (Y2P15), although there was no module taught during that period. This is because of the late students that finalized their reports at the last minute. Y2P01, Y2P02 and Y2P09 exhibit the lowest activity, and these were the start of the first and second terms. Y2P01 and Y2P01 were also taught in face-to-face and they correspond to the start of all three courses.

The learners-teachers interactions, especially during the lockdown, are quantified here by the number of questions. Figure 4b) shows the amount of contributions divided by the number of questions as a measure of work that was done on average by the students for each interaction. A higher value means more autonomy. A lower value indicates more problems or difficulties that require learners-teacher interactions to be solved. As this measure spreads over several orders of magnitude from one student to the other, a logarithmic scale is used. However, the median value -the bar inside the boxes- varies much less. the global number of questions during each period is less variable, as are the absolute contributions, leading to a rather stable ratio. The highest median ratios are observed at the last period of each term (Y2P07 and Y2P15) although no module was taught during that time. More contributions are observed relative to the questions at the end: students essentially finalize their reports.

The first year shows a different pattern. First, the lockdown period was restricted to the very end of the second term. Only the last module in both Courses A and B remained. In Y1P12, when distance learning was first imposed, we observed a marked decrease in the contributions per student (Fig. 4a). It is heavily compensated, and even overcompensated, in periods Y1P13 and Y1P14, which were by far the busiest periods of all. Period Y1P15 was not represented because it is after the deadline to finish all work that year.

The intensity of support during the first year shows a similar pattern as for the second year: extremely widespread from student to student. The median value is similar too, if not among the highest during periods Y1P11, Y1P13 and Y1P14. The productivity was thus not affected during that first lockdown, after a short lag time observable in Y1P12.

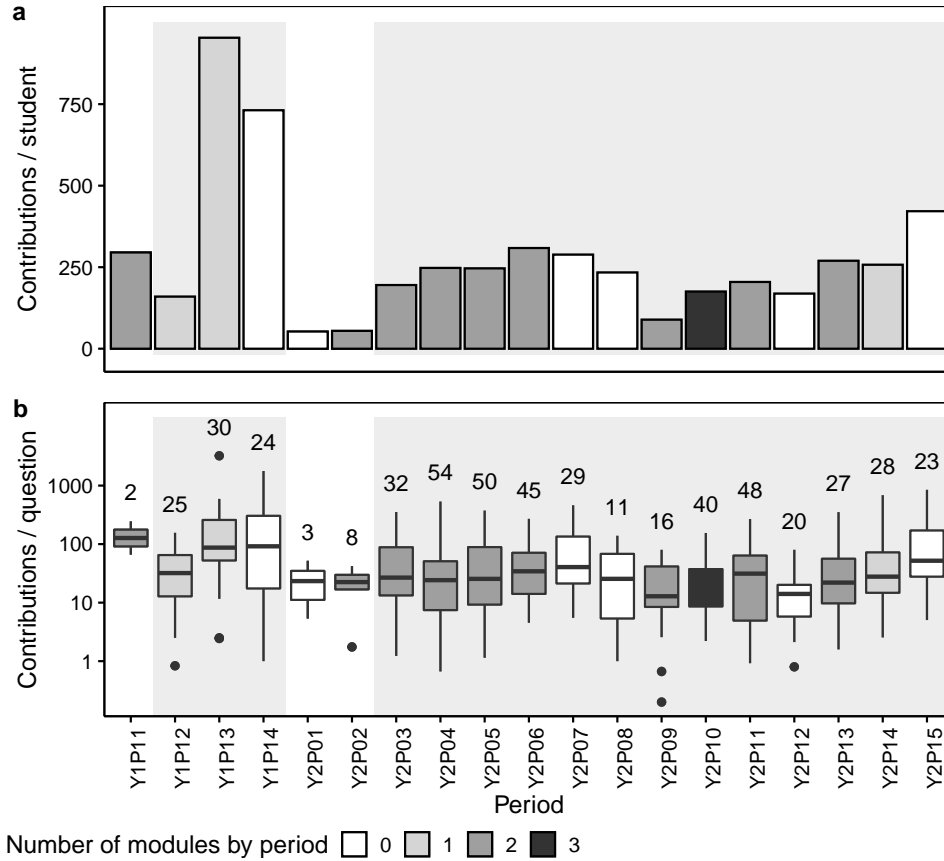


FIGURE 4. a. Average student contributions to the projects per two-week course. b. Contributions per question asked (log scale) for each student as a measurement of the intensity of teacher-learner interactions relative to the progression. Light gray background indicates periods where distance teaching was mandatory due to the COVID-19 lockdown. The number of students that interacted during each period is indicated on top of the boxplots (Y1 is 2019-2020, Y2 is 2020-2021).

536 **4. Discussion.** Teaching data science to a population of students that are not very
 537 used to advanced computer techniques and tools, and that have only basic knowl-
 538 edge in mathematics and statistics is a hard task [38]. In order to let them learn
 539 progressively, the courses were stretched out over a very long period of time: five
 540 successive terms spanning three consecutive years (undergraduate and graduate).
 541 That way, the different concepts they had to learn were broken down into subunits
 542 (26 modules) that lasted for two weeks each. We also used flipped classrooms and
 543 blended teaching and learning (following Spadafora & Zopito’s definition of “any
 544 educational model where online delivery ranges from 50% to 80%” [39]), with an
 545 emphasis on proactive exchanges with the teachers: students had to ask questions to
 546 progress. Overall, these appear to be winning choices because most of our students

were successful, excluding a few who failed. Compeau also obtained good results using the flipped classroom with the course targeting students in biology [14].

Despite our course framework, students are more used to a traditional face-to-face approach made of lectures followed by exercises where important concepts are repeated at the beginning of the practical sessions. They tend to have a passive attitude during lectures and they expect teachers and assistants to feed them with the key concepts. That attitude does not purposely work here. Proactive behavior and development of autonomy are required [18]. They thus have to engage themselves in a very different way of learning. The transition between the theory they read in a book and the projects where they have to apply these concepts is too sharp without a progression that facilitates students' engagement. The four stages of the progression were: (L1) auto-evaluation exercises directly in the online book, (L2) recall of the main concepts and a guided step by step analysis of a first dataset with the learnr tutorials, (L3) at least one guided individual project with another dataset, before (L4) where they are presented yet a different dataset to analyze with limited instructions this time.

4.1. Measured and Perceived Cognitive Workload in learnr Tutorials. The learnr tutorial was immediately spotted as a key activity in the learning process during the 2018-2019 academic year. As such, we focused our attention on these learnr documents. In 2020-2021, the use of a heuristic engine {gradethis} to provide contextual feedback on the errors students made in their answers was much appreciated.

The number of trials needed to find the right answer is considered here as a workable proxy for the cognitive workload. Contextual feedback allowed the student to correct these answers on their own to find the expected answer. A high number of trials may definitely indicate that students had a hard time to find the right answer. They could misunderstand the concepts, but they could also be trying many solutions at random in order to find the right solution. This indicator alone is not sufficient, so we combined it with the perceived cognitive workload. In the future, the measured RTLX index will serve as a reference to gauge possible optimization of the tutorials, with lower perceived workload without sacrificing the content. The significant decrease in the RTLX value from Course A to C indicates that there is still a margin of progression. We would like to observe such a decrease sooner, perhaps already in the second course. Monitoring the perceived and measured cognitive workload is indeed important “to maintain reflective and systematic approaches in both the development and evaluation of [our] blended approach” [39].

4.2. Student Activity Profiles with Ongoing Assessment. Activity tracking in the exercises, primarily set up for the ongoing assessment, also offers the opportunity to study the way learning happens (or not). Indeed, learning analytics provides opportunities to monitor learning events as well as to adjust teaching to improve student outcomes [30, 34], even if they are primarily used to early predict success or failure. In our courses, the failure rate was already rather low and essentially limited to a few defeating students that did not work at all. We were more interested in classifying our participating students according to their behaviors. This paved the way toward a more inclusive pedagogy by spotting different kinds of suboptimal patterns (for instance, never asking questions, looking at hints too quickly without really trying to figure out the answer, being shy about discussing problems on public channels, etc.). Once these patterns are evidenced, we can

then consider countermeasures. As Martin and Ndoe highlight from other studies, “benefits that the online learning platform provides with respect to assessment include better monitoring opportunities for student learning and immediate feedback [...], and individual practice opportunities” [30]. As an example, for students that rarely post their questions publicly, we will test an alternate discussion channel where teachers never post but have read access. In case of an error, the teacher will contact the student privately to explain to him what is wrong. That student would then have the responsibility to reexplain the point correctly to their classmates. This way, this error is never publicly pointed out by the teacher. With tools like the self-organizing map, we should be able to predict suboptimal student profiles early. We could engage in a discussion with the students concerned to determine the cause and find a solution as early as possible. Learning analytics used in this way could promote a differential pedagogical approach, a key for more inclusive teaching [37].

Group projects are one of the keys to our method. Sometimes, groups do not work well, and one student has to do most of the work. This is a clear weakness of this approach, especially if one of the failing students in Fig. 3 Cell (3) is involved. If we could identify the profile of the different students relatively early during the course, we would be able to create better groupings with a blend of different complimentary profiles to enrich the experience of all learners. Maybe should we work exclusively with groups of four to mitigate the impact of one failing student? The balance of heterogeneous and complimentary competences are essential in such a group in order to create mutual emulation and efficacy [32]. Working on group composition will thus be one of our future challenges.

4.3. Transition from Face-to-face to Distance Learning Imposed by the COVID-19 Lockdown. Forced distance learning, due to the COVID-19 lockdown did not appear to be a barrier in the production of our students in their projects. A pattern was observed during the first lockdown with a marked decrease in their contributions, followed by a large, compensatory activity. All this happened in a time frame of a couple of weeks. That was the time needed to adapt to the new situation. **Several causes can be put forward to explain this adaptation time. Our study does not attempt to determine the main cause or causes. However, the first problem identified** was the access to a powerful-enough computer for roughly 15 to 20% of our students. In a normal situation, these students had access to computers on the university premises, both in-class and outside of class time. **When the lockdown was imposed, these students suffered a lack of hardware.** However, to reduce the social numeric divide, the university quickly reacted and computers were lent to them. During the second lockdown, a larger part of our students had acquired their own computer, and solutions were immediately available for the others.

While the contributions/questions remained globally at a similar level in face-to-face and distance learning, their impact on the teacher’s timetables was very different. In distance learning, students worked at very different times. Their questions were thus less concentrated during the course periods. Also, an alternation between asynchronous work at home and synchronous work in the computer lab was more beneficial to interactions between students. The social and human components of teaching and learning are key factors that tend to vanish in exclusive distance learning. Contacts through videoconferences only partly compensate for a lack of interactions because in-class presence remains different to video chats. Blended

learning combines the best of the two practices if pedagogical setups are accurate and well balanced [7].

5. Conclusion and Perspectives. Teaching data science comes with challenges. The discipline is quite young, and we are still seeking the best pedagogical approach. After three years of teaching data science to undergraduate and graduate students on a biology curriculum with revised pedagogical practices, we have had our first cohort that passed all three courses. There are still two optional courses available in the second year of the Master if they want to push their data science skills further on. However, the three mandatory courses are designed to be self-supported. Globally, most students acquired the expected competences during these courses. We have the feeling that they are more mature and that they effectively acquired the intended outcomes in data science more than with our previous courses in biostatistics, which was given in a more traditional way. The impact of the revised approach to teach biological data science on the way learners manage data and data analysis will be observable during the following years. We will monitor how these students apply their skills in their Master's thesis, and later, in their career or during their PhD thesis. Meanwhile, we will continue to improve our courses by further exploiting the data we accumulate on the activity of our students. Experience gathered during forced distance learning during the COVID-19 lockdown will also be used to improve our courses framework. The radical changes that were required in that context showed that students can accommodate, to a large extent, but also that the diversification of the activities is beneficial to guarantee their engagement [39, 46]. Regarding diversification, in 2020-2021 we successfully tested a kaggle-like challenge (<https://www.kaggle.com/competitions>) in one of the machine learning modules. Such playful activities could also contribute to the diversification of pedagogical practices, interest and motivation of the students [1]. We would also be happy to share experiences with other teachers in data science. Altogether, we are on the way to reshaping the post-COVID teaching landscape, and it will probably be quite different than what we were used to!

REFERENCES

- [1] C. Alonso-Fernandez, A. Calvo-Morata, M. Freire, I. Martínez-Ortiz and B. Fernandez-Manjon, Applications of data science to game learning analytics data: A systematic literature review, *Computers & Education*, **141** (2019), 103612.
- [2] H. Alvarenga da Silva and A. Sampaio, Teaching introductory statistical classes in medical schools using RStudio and R statistical language: Evaluating technology acceptance and change in attitude toward statistics, *Journal of Statistics Education*, **28** (2020), 212–219.
- [3] L. A. Auker and E. L. Barthelmess, Teaching R in the undergraduate ecology classroom: approaches, lessons learned, and recommendations, *Ecosphere*, **11** (2020), e03060.
- [4] M. Baker, 1,500 scientists lift the lid on reproducibility, *Nature*, **533** (2016), 452–454.
- [5] G. C. Banks, J. G. Field, F. L. Oswald, E. H. O'Boyle, R. S. Landis, D. E. Rupp, S. G. Rogelberg. Answers to 18 Questions About Open Science Practices, *Journal of Business and Psychology*, **34** (2019), 257–270.
- [6] B. Baumer, M. Cetinkaya-Rundel, A. Bray, L. Loi and N. J. Horton, R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics, *Technology Innovations in Statistics Education*, **8** (2014).
- [7] R. M. Bernard, E. Borokhovski, R. F. Schmid, R. M. Tamin and P. C. Abrami, A meta-analysis of blended learning and technology use in higher education: from the general to the applied, *J Comput High Educ*, **26** (2014), 87–122.
- [8] C. Boettiger, An Introduction to Docker for Reproducible Research, *SIGOPS Oper. Syst. Rev.*, **49** (2015), 71–79.

- 694 [9] R. Burton, S. Borruat, B. Charlier, N. Coltice, N. Deschryver, F. Docq and al., Vers une ty-
695 pologie des dispositifs hybrides de formation en enseignement supérieur, *Distances et savoirs*,
696 **9** (2011), 69–96.
- 697 [10] J.C. Byers, A. Bittner and S. Hill, Traditional and raw task load index (TLX) correlations:
698 Are paired comparisons necessary?, *Advances in Industrial Ergonomics and Safety*, 1989,
699 481–485.
- 700 [11] M. Cetinkaya-Rundel and C. Rundel, Infrastructure and Tools for Teaching Computing
701 Throughout the Statistical Curriculum, *American Statistician*, **72** (2018), 58–65.
- 702 [12] M. Cetinkaya-Rundel and V. Ellison, A Fresh Look at Introductory Data Science, *Journal of*
703 *Statistics Education*, **0** (2021), 1–27.
- 704 [13] W.S. Cleveland, Data science: An action plan for expanding the technical areas of the field
705 of statistics, *International Statistical Review*, **1** (2001), 21–26.
- 706 [14] P. Compeau, Establishing a computational biology flipped classroom, *PLoS Computational*
707 *Biology*, **15** (2019), 1–8.
- 708 [15] D. Donoho, 50 Years of Data Science, *Journal of Computational and Graphical Statistics*, **26**
709 (2017), 745–766.
- 710 [16] R.A. Estrellado, E. A. Emily, J. Mostipak, J. M. Rosenberg and I. C. Velasquez, *Data science*
711 *in education using R*, 1st edition, Routledge, London, England, 2020.
- 712 [17] J. Fiksel, L. R. Jager, J. S. Hardin and M. A. Taub Using GitHub Classroom to teach statistics,
713 *Journal of Statistics Education*, **27** (2019), 110–119.
- 714 [18] D. R. Krathwohl, Active learning increases student performance in science, engineering, and
715 mathematics, *Proceedings of the National Academy of Sciences*, **111** (2014), 8410–8415.
- 716 [19] P. Grosjean and G. Engels, Biological data science courses at UMONS, Belgium: student’s
717 activity for 2020-2021 (1.0.0) [Data set]., *Zenodo*, (2022).
- 718 [20] P. Grosjean and G. Engels, Biological data science courses at UMONS, Belgium: student’s
719 activity for 2019-2020 (1.0.0) [Data set]., *Zenodo*, (2022).
- 720 [21] P. Grosjean and G. Engels, Biological data science courses at UMONS, Belgium: student’s
721 activity for 2018-2019 (1.0.0) [Data set]., *Zenodo*, (2022).
- 722 [22] L.M. Guzman, M.W. Pennell, E. Nikelski and D.S. Srivastava, Successful integration of data
723 science in undergraduate biostatistics courses using cognitive load theory, *CBE-Life Sciences*
724 *Education*, **18** (2019), ar49 1–10.
- 725 [23] S. G. Hart and L. E. Staveland, Development of NASA-TLX (Task Load Index): Results of
726 Empirical and Theoretical Research, *Advances in Psychology*, **52** (1998), 139–183.
- 727 [24] C. Hsing and V. Gennarelli, Using GitHub in the Classroom Predicts Student Learning Out-
728 comes and Classroom Experiences: Findings from a Survey of Students and Teachers, In Pro-
729 ceedings of *the 50th ACM Technical Symposium on Computer Science Education (SIGCSE*
730 *’19)*, 2019, 672–678.
- 731 [25] P.A. Kirschner and C. Hendrick, *How learning happens: seminal works in educational psy-*
732 *chology and what they mean in practice*, 1st edition, Routledge, London, New York, 2020.
- 733 [26] T. Kohonen, *Self-Organizing Maps*, 1st edition, Springer-Verlag, Berlin Heidelberg, 1995.
- 734 [27] D. R. Krathwohl, A Revision of Bloom’s Taxonomy: An overview, *Theory Into Practice*, **41**
735 (2002), 212–218.
- 736 [28] K. Larwin and D. Larwin, A meta-analysis examining the impact of computer-assisted in-
737 struction on postsecondary statistics education: 40 years of research, *Journal of Research on*
738 *Technology in Education*, **43** (2011), 253–278.
- 739 [29] L. Lloyd-Smith, Exploring the Advantages of Blended Instruction at Community Colleges and
740 Technical Schools, *Journal of Online Learning and Teaching*, **6** (2010).
- 741 [30] F. Martin and A. Ndoeye, Using Learning Analytics to Assess Student Learning in Online
742 Courses, *Journal of University Teaching & Learning Practice*, **13** (2016), 69–96.
- 743 [31] V. Marx, The big challenges of big data, *Nature*, **498** (2013), 255–260.
- 744 [32] R. Mucchielli, *Le travail en équipe*, 1st edition, ESF Editions, Paris, France, 1996.
- 745 [33] J. A. Onwuegbuzie and V. A. Wilson, Statistics Anxiety: Nature, etiology, antecedents, effects,
746 and treatments—A comprehensive review of the literature, *Teaching in Higher Education*, **8**
747 (2003), 195–209.
- 748 [34] C. Romero and S. Vetura, Educational data mining and learning analytics: An updated
749 survey, *WIREs Data Mining Knowl Discov*, **10** (2020), e1355.
- 750 [35] R Core Team (2021), R: A Language and Environment for Statistical Computing, R Founda-
751 tion for Statistical Computing, Vienna, Austria. <https://www.R-project.org>
- 752 [36] RStudio team (2015), RStudio Cloud, Boston, MA: RStudio, Inc.

- [37] G. Siemens, Learning Analytics: The Emergence of a Discipline, *American Behavioral Scientist*, **57** (2013), 1380–1400.
- [38] B. Sousa and D. Gomes, Teaching With R—A Curse or a Blessing?, In Proceedings of *the Tenth International Conference on Teaching Statistics (ICOTS10 '18)*, 2018, 672–678.
- [39] N. Spadafora and Z. Marini Self-Regulation and “Time Off”: Evaluations and Reflections on the Development of a Blended Course, *The Canadian Journal for the Scholarship of Teaching and Learning*, **9** (2018).
- [40] A. S. Theobald, A. Hancock and S. Mannheimer, Designing Data Science Workshops for Data-Intensive Environmental Science Research, *Journal of Statistics and Data Science Education*, **29** (2021), S83–S94.
- [41] University of Mons Study programme 2021-2022 : Data Sciences I visualisation and inference, Accessed March 21, 2022, <https://applications.umons.ac.be/web/en/pde/2020-2021/ue/US-B2-SCBIOL-006-M.htm>
- [42] University of Mons Study programme 2021-2022 : Data Sciences II analysis and modelling, Accessed March 21, 2022, <https://applications.umons.ac.be/web/en/pde/2020-2021/ue/US-B3-SCBIOL-006-M.htm>
- [43] University of Mons SStudy programme 2021-2022 : Data Sciences III exploration and prediction, Accessed March 21, 2022, <https://applications.umons.ac.be/web/en/pde/2020-2021/ue/US-M1-BIOECO-004-M.htm>
- [44] R. Wehrens and J. Kruisselbrink, Flexible Self-Organizing Maps in kohonen 3.0, *Journal of Statistical Software, Articles*, **87** (2018), 1–18.
- [45] H. Wickham et al, Welcome to the Tidyverse, *Journal of Open Source Software*, **43** (2019), 1686.
- [46] J. Young, Hybrid teaching seeks to end divide between traditional and online instruction, *The Chronicle of Higher Education*, **48** (2002), A33–34.

Received xxxx 20xx; revised xxxx 20xx.

E-mail address: Guyliann.Engels@umons.ac.be

E-mail address: Philippe.Grosjean@umons.ac.be

E-mail address: Frederique.Artus@umons.ac.be