Teaching Data Science to Students in Biology using R, RStudio and Learnr: Analysis of Three years Data

1 Abstract

This is the original abstract that should be reworked according to final content of the manuscript.

The courses in biostatistics in biology at the University of Mons, Belgium, were completely refactored in 2018 into data science courses (see http://bds.sciviews.org). The content is expanded beyond statistics to include computing tools, version management, reproducible analyses, critical thinking and open data. Flipped classroom approach is used. Students learn with the online material and they apply the concepts on individual and group projects using a preconfigured virtual machine with R and RStudio. Activities (H5P, learnr or Shiny applications) are recorded in a MongoDB database (300,000+ events for 180+ students and 2,000+ GitHub repositories at https://github.com/BioDataScience-Course). The analysis of these data reveals several trends. (1) There is a relatively long lag period required for the students to get used to the computing environment, the teaching method and the data science in general. (2) Implication is very high, with more than 85% of the students that complete all the activities and got good to excellent assessment. (3) There is a gap between students' own perception of their skills achievements and their assessment results: they tend to underestimate their progress. (4) During COVID-19 pandemic lockdown, the intensity of the activities largely decreased during two weeks before returning to previous level, but for 3/4 of the students only. The remaining fraction never caught up. We hypothesize that the technical requirements or the lack of motivation during the lockdown were detrimental to roughly one student over ten, despite all the efforts the University deployed to reduce the social fracture.

2 Introduction

In a context where there is an exponentially growing mass of data [ref...], a reproducibility crisis in Science (Baker 2016), and a progressive adoption of Open Science practices (Banks et al. 2019), statistics were broaden to a larger discipline called data science. For the Data Science association, "the Data Science means the scientific study of the creation, validation and transformation of data to create meaning" (http://www.datascienceassn.org/code-of-conduct.html). These changes also led to the emergence of data science programs in universities and higher schools (Donoho 2017; Çetinkaya-Rundel and Ellison 2021). One example is the Harvard Data Science initiative (https://datascience.harvard.edu/about) initiated in 2017. With a broader approach, comes also a broaden public. The data science courses are not just limited to computer scientists, mathematicians or statisticians, but also welcome students in humanities, social sciences, and natural sciences (for instance, the data science training at Duke University (Çetinkaya-Rundel and Ellison 2021). Main focus of such courses is for students to develop the ability to deal with "real" datasets in all their complexities and to realize reproducible analyses to interpret these data in the light of knowledge in their field of expertise.

The data management part of the job is a challenge for students with a poor or no background at all in computing. Students that are not used to deal with computer languages enter in a foreign world and have to deal with many exotic concepts, techniques and tools. [même problème concernant les stats] This generates anxiety (see for instance (Onwuegbuzie and Wilson 2003), for students in biology). The course must be organized in a way that such students progress by little steps in order to avoid exposition to much intimidating concepts and tools at once. Hence, a student in computing science already masters one or more computing languages, is acquainted with version control systems, with databases and with the way data

are represented in a computer. A student in mathematics or statistics is familiar with various concept that underpin the techniques to analyse the data. On the other hand, students in biology, medicine, psychology, social sciences, economics, … have very different *a priori* knowledges.

[TODO: to write this section] Teaching of git & GitHub (Hsing and Gennarelli 2019; Fiksel et al. 2019), R Markdown and projects management integrated in data science courses [Baumer et al. (2014); Xie2019]. Git offers the double advantage to track changes in student's projects, and to teach them one valuable tool they will use in their future career.

Suitable computer hardware and software environments are required in the practical sessions of these courses. Different approaches range from inline software (RStudio Cloud, Chromebook data science) to local installation on the Student's computers. The later raises problems of license for proprietary software, but also installation and configuration issues. An intermediary solution uses preconfigured virtual machines, or containers (e.g., Docker) (Çetinkaya-Rundel and Rundel 2018) [refs]. This allows to play with the concepts and work on the various projects anywhere (in the computer lab, at home, using a laptop, …). To fix theoretical concepts through applied exercises is a key aspect of learning data science (Larwin and Larwin 2011) [ref take.Let Them Eat Cake First! ?] Correct choice of software is critical and exposing students early with the tools they are most susceptible to use later in their work is desirable. This was highlighted by (Auker and Barthelmess 2020) for instance, for the analysis of ecological data.

These data science course represent thus several challenges to pedagogy because various, numerous and unfamiliar concepts must be acquired by a population of potentially very diverse students. Learning objectives span a large range of cognitive abilities (Krathwohl 2002). [explain here in 2-3 sentences main approaches used in these courses + refs]. [main points are: active pedagogy, development of student's autonomy in flipped classrooms, continuous evaluation and project pedagogy.] The flipped classroom approach allows students to be active in their learning, which has the benefit of improving student outcomes (Freeman et al. 2014).

[Partie pédagogie à détailler un peu , probablement sur 2 ou 3 paragraphes]

Recently, data science is also used to analyze the effect of different pedagogical practices on the outcome of these courses (Estrellado et al. 2020). With numerical tools, a vast amount of data can be collected on students activities, and the analysis of these data allows to compare the impact of different pedagogical approaches, or to quantify and document the impact of changes in the courses.

At the University of Mons, in Belgium, we have started to rework our biostatistics courses in the biology curriculum in 2018. A series of Data Science courses were introduced, both for our undergraduate and graduate students. These courses are inspired from precursor initiatives cited here above. The goal of these courses is to form biological data scientists capable to extract meaningful information from raw biological data, and to do so in a reproducible way and with correct application of statistical tools and an adequate critical mindset. A preconfigured VirtualBox virtual machine with R, RStudio, Rmarkdown, git, and a series of R packages preinstalled is used (url sciviews box?), as a very flexible way to deploy the same software environment both on the university computers and on student's own laptops.

As our course were completely reworked, we also decided to use flipped classroom and progressive adoption of suitable pedagogical practices with a cyclical approach that consists in stating goals, building pedagogical material with a large emphasis on numerical tools and collection of student's activities, and analysis of these data. Conclusions of these analyses initiate another cycle the following academic year with refined goals and pedagogical material or techniques. Here, we present the main results spanning on three successive academic years from 2018 to 2021, including two particular periods where distance learning was forced due to COVID pandemic lockdown.

[TODO: ajouter une informaiton sur notre objectif : former des étudiants pouvant analyser des données.]

[TODO: present here the 3-4 research questions that will be elaborated in the manuscript.]

3 Material and methods

Niveau	description	type d'exercices
N1	Exercice court intégré directement dans le cours en ligne avec un feedback immédiat	h5p, shiny
N2	Exercice guidé dans un tutoriel avec un feedback immédiat	learnr
N3	Exercice cadré sous la forme d'un projet individuel	ind github
N4	Exercice libre sous la forme de projet de groupe	group github

ajouter une colonne avec la taxonomie de bloom

The NASA-LTX indicator is composed of six questions on a Likert scale to quantify the perceived workload to complete a tutorial (Hart and Staveland 1988). The questions concern mental load, physical load, time pressure, expected success, effort required, and frustration experienced during the accomplishment of the task. The average value for the six questions constitutes a Raw Task Load indeX (RTLX) (Byers, Bittner, and Hill 1989).

The SUS ….

rechercher une ref plus récente type meta analyse SUS et NASA LTX

4 Results

Liste des idées :

- exam versus project 2018 et 2019 => elimination de l'examen
- profiles analyse SOM => sous-groupes + analyse des groupes y compris comparaison avec les grades
- temporel présentiel -> distanciel
- learnrs (+ perception) -> apprentissage sur le très long terme (> 1 ou deux quadris)

La transition du cours classique vers une cours en classe inversée a menée à l'intégration de nouveaux outils permettant de diversifier les types d'exercice proposés aux étudiants (tab M&M). Le tableau XX indique la répartition des exercices pour chaque cours. La collecte des données pour chaque exercice permet de construire une note objective pour chaque étudiant. La note des étudiant est construite sur l'évaluation des 5 niveaux d'exercices complémentaires allant des exercices les plus simples (N1) aux exercices les plus complexes (N5). Une moyenne pondérée pour chaque niveau d'exercice est employée pour obtenir une note finale par cours.

Our in-service training includes 26 complementary modules over 3 years and 3 consecutive courses from bab2 to MA1. The number of exercises per type is shown in Table XXX. The goal and the difficulty levels of the exercises are presented in table xxx. The completion of each exercise is recorded in a database which allows an objective grade to be constructed for each student.

```
# Tab of number of users and number of exercices by type ----------------------
# users
users %>.%
  filter(., institution == "UMONS" & term == "Q1" & state == "regular") %>.%
  group_by(., course) %>.%
  summarise(., user = n()) %>.%
  ungroup(.) %>.%
  filter(., course != "D")-> us_tab

# learnr
learnr %>.%
  filter(., !app %in% c("A06Lb_recombinaison", "A99La_avis", "B00La_rappel", "B99La_avis","C99La_avis")
  mutate(., course = substr(app,1,1),
    app_label = paste0(app, label)) %>.%
```

```
  filter(., course %in% c("A", "B", "C")) %>.%
  group_by(., course) %>.%
  summarise(., app = length(unique(app)), questions = length(unique(app_label))) -> learnr_tab

# projects
projects %>.%
  filter(., type %in% c("ind. github", "group github") & course != "D") %>.%
  group_by(., course, type) %>.%
  count(.) %>.%
  pivot_wider(., names_from = "type", values_from = "n") %>.%
  ungroup(.) %>.%
  select(., course, `ind. github`, `group github`)-> projects_tab

# tab number of exercices by type ---
assessments %>.%
  filter(., type == "h5p") %>.%
  mutate(.,
    app_type = paste0(app, "_" ,'type'),
    course = substr(app, start = 1, stop = 1),
    ) %>.%
  group_by(., course) %>.%
  summarise(., h5p = length(unique(app_type))) -> h5P_tab

us_tab %>.%
  mutate(., module = c(12, 8, 6)) %>.%
  left_join(., h5P_tab) %>.%
  left_join(., mutate(learnr_tab, learnr = paste0(app, " (", questions, ")"), .keep = "unused")) %>.%
  left_join(., projects_tab) %>.%
  knitr::kable(., caption = "Number of users, modules, exercises with h5p, learnr tutorial,individual p
```

Table 2: Number of users, modules, exercises with h5p, learnr tutorial,individual project and group project by course. The number of quesion by learnr tutorial are in the parentheses.

course	user	module	h5p	learnr	ind. github	group github
A	42	12	59	24 (211)	10	4
B	40	8	29	11 (108)	12	2
C	25	6	19	7 (37)	7	1

4.1 Exams versus project

```
assessments18 %>.%
  #select(., -coral_growth, result = biometry) %>.%
  mutate(., result = (biometry+coral_growth)/2, acad_year = "2018-2019") -> assess_result18

q1_18_regular <- left_join(rename(assess_result18, icourse = course), courses18) %>.%
  filter(., user %in% users18$user[users18$institution == "UMONS" & users18$term == "Q1" & users18$stat

assessments19 %>.%
  group_by(., course, evaluation, github_project, project, user) %>.%
  summarise(., result = round(sum(score*weight),4)) %>.%
  filter(., evaluation == "Q1") %>.%
```

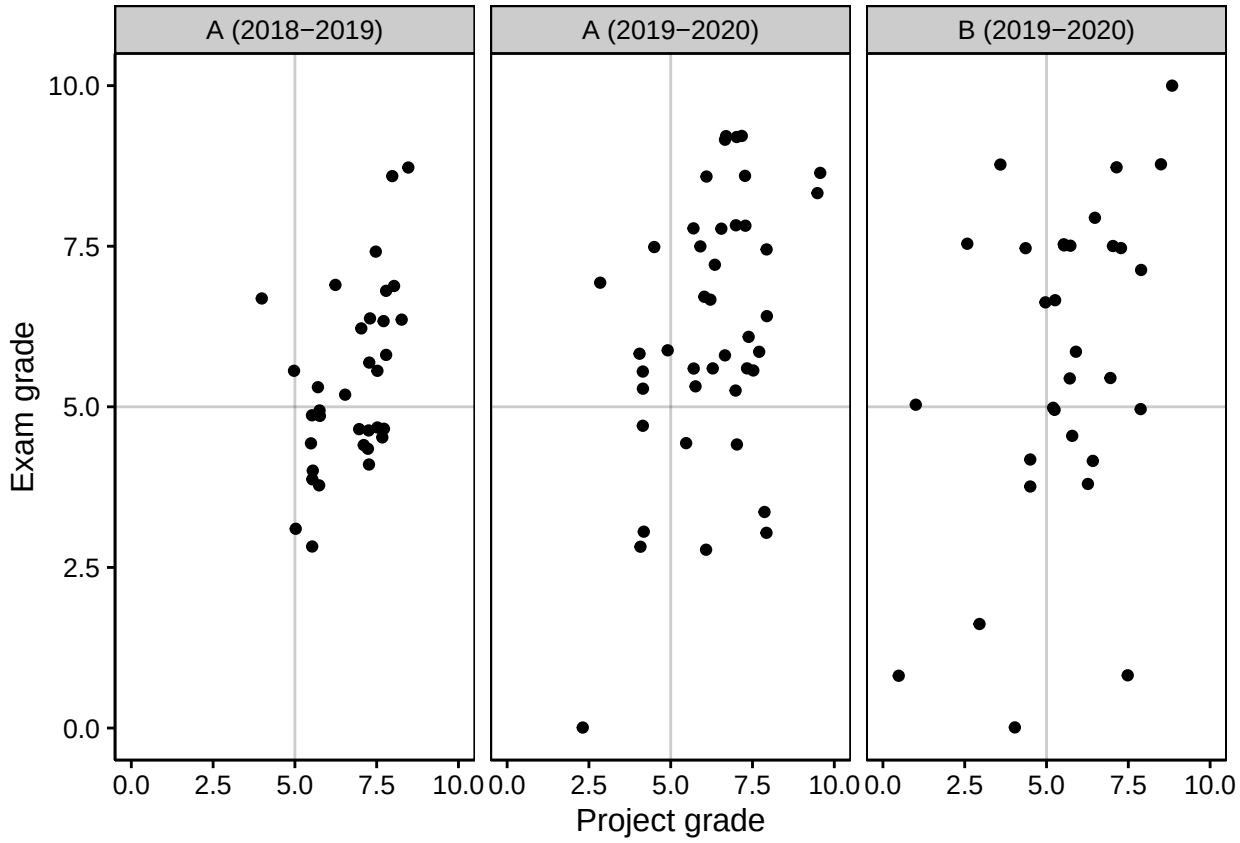
```
  left_join(exam19, .) %>.%
  replace_na(., list(result = 0)) %>.%
  mutate(., acad_year = "2019-2020")-> assess_result19

q1_19_regular <- left_join(rename(assess_result19, icourse = course), courses19) %>.%
  filter(., user %in% users19$user[users19$institution == "UMONS" & users19$term == "Q1" & users19$stat
```

```
q1 <- bind_rows(
  select(q1_18_regular, user, acad_year, course, result, exam),
  select(q1_19_regular, user, acad_year, course, result, exam)
  )
#table(q1$link) /nrow(q1)
```

Until the first four months of the 2019-2020 academic year, students' grades were based on the completion of a project and a more conventional examination during the examination period. The projects are assessed using a grading grid since the 2019-2020 academic year.

```
q1 %>.%
  mutate(., course_year = paste0(course, " (",acad_year,")")) %>.%
  chart(., exam ~ result | course_year) +
  geom_vline(xintercept = 5, alpha = 0.2) +
  geom_hline(yintercept = 5, alpha = 0.2) +
  geom_jitter(alpha = 1, width = 0.05, height = 0.05, show.legend = FALSE) +
  ylim(c(0,10)) +
  xlim(c(0,10)) +
  labs(y = "Exam grade", x = "Project grade")
```



The comparison of the marks obtained between the project and the exam mark shows a strong disparity between these two types of evaluation.

The 2018-2019 year is the first year of transition to data science courses. Only one student failed the project, while almost one third of the students failed their exams. The level of requirements is being raised for the 2019-2020 academic year. The new expectations and the introduction of evaluation grids to assess projects show a greater disparity in students' grades.

Despite an examination that includes theoretical and practical questions, this type of assessment does not assess a student's ability to correctly process and analyse biological data.

Following these results and the monitoring of more precise exercises, the examination is definitively abandoned for the 2020-2021 academic year to be replaced by a continuous assessment.

4.2 Students' profiles

[TODO : Add SOM analyses, This analysis is stand by as we validate the metrics]

4.3 temporality of work

[TODO : add analyse on the commit or h5P/learnr exercices to follow the student]

4.4 Learnr tutorials

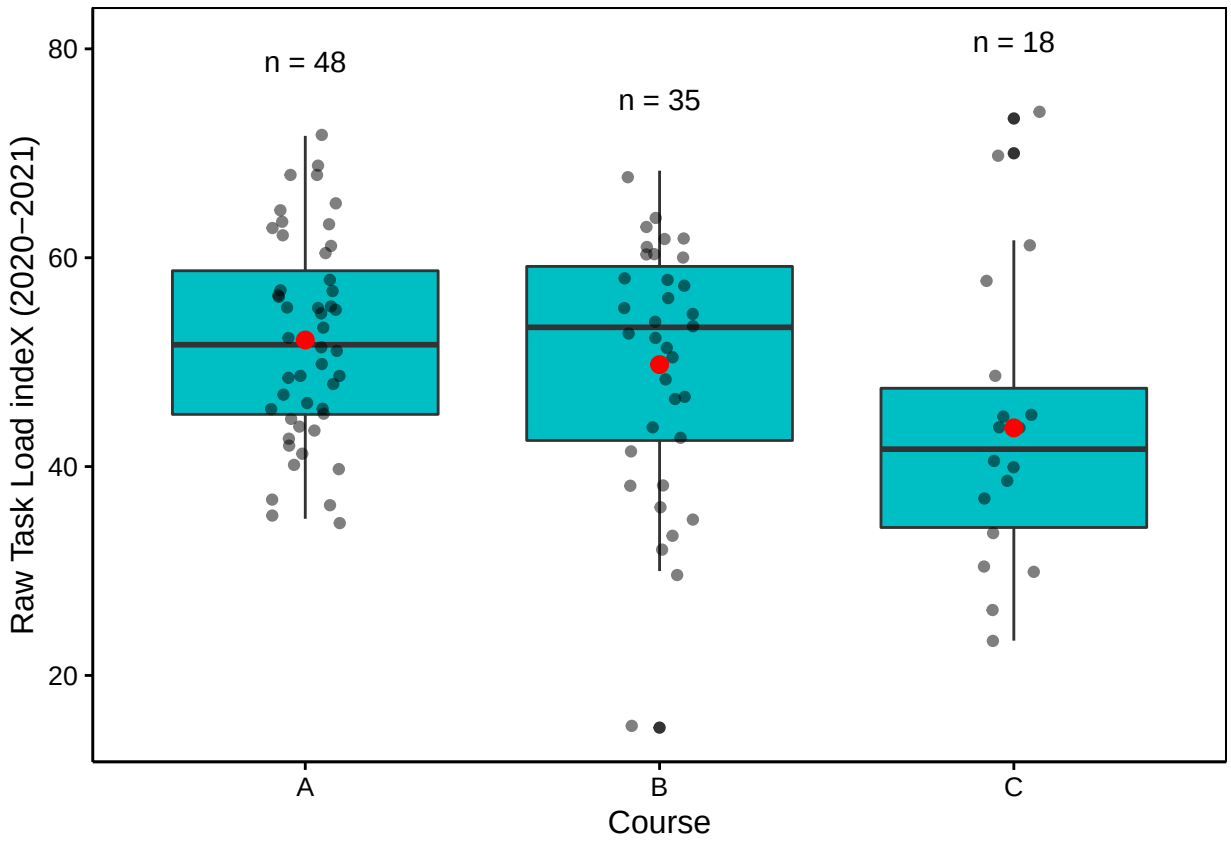
The learnr tutorials are an essential part of the learning method to link theory and practice. The cognitive load required to perform tutorials is studied via a NASA LTX questionnaire.

```
c("A99Wa_perception", "B99Wa_perception", "C99Wb_perception:perception") %>%
  purrr::map_dfr(learnr_feeling, df = wo, label = "Q4") %>.%
  mutate(., course = substr(app, start = 1, stop = 1)) -> learnr_workload

learnr_workload %>.%
  pivot_longer(.,cols = c(mental, physical, time_pressure, performance, effort, frustration),
  names_to = "category", values_to = "grade")  %>.%
  left_join(., dplyr::distinct(courses, course,name), by = "course")-> workload

workload %>.%
  group_by(., user, app, course) %>.%
  #filter(., user != "ECAYEO033") %>.%
  summarise(., rtlx = 10*mean(grade)) -> workload_rtlx

set.seed(222)
chart(workload_rtlx, rtlx ~ course) +
  geom_boxplot(fill = "#00BFC4") +
  geom_jitter(alpha = 0.5, width = 0.1) +
  labs(y = "RTLX", x = "Course") +
  stat_summary(fun.y="mean", color = "red")+
  stat_summary(fun.data = n_fun, geom = "text", hjust = 0.5) +
  labs( y = "Raw Task Load indeX (2020-2021)")
```



The difficulty of the tutorial exercises increases from course to course. However, we observe a significantly lower RTLX value for the course C than for the course A (Tukey HSD, p-value = 0.023). The cognitive load perceived by the students is therefore lower.

```
workload_rtlx %>.%
  mutate(., course = as.factor(course)) -> workload_rtlx

#kruskal.test(data = workload_rtlx, rtlx ~ course)
#summary(kw_comp. <- nparcomp::nparcomp(data = workload_rtlx, rtlx ~ course))

anova. <- lm(data = workload_rtlx, rtlx ~ course)
anova(anova.)
```

```
## Analysis of Variance Table
##
## Response: rtlx
##           Df  Sum Sq Mean Sq F value  Pr(>F)
## course     2   926.9  463.47  3.5883 0.03134 *
## Residuals 98 12658.0  129.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#bartlett.test(data = workload_rtlx, rtlx ~ course)
#plot(anova., which = 2)

summary(anovaComp. <- confint(multcomp::glht(anova.,
  linfct = multcomp::mcp(course = "Tukey"))))
```

```
##
```

```
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = rtlx ~ course, data = workload_rtlx)
##
## Linear Hypotheses:
##            Estimate Std. Error t value Pr(>|t|)
## A - B == 0    2.356      2.526   0.933   0.6183
## C - B == 0   -6.058      3.296  -1.838   0.1607
## C - A == 0   -8.414      3.141  -2.679   0.0229 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

5 Discussion

Les études portant sur le changements d'attitudes au sein de semestre ne montre pas différence significative. La comparaison entre les 3 cours met en avant qu'il faut plusieurs cours en continue afin d'observer une chagement de la charge cognitive des étudiants.

apprentissage en continu sur 3 années successives (cohérence entre le programme et l'approche pédagogique), les résultats sont meilleurs vers la 3ieme années.

6 Conclusions

- Exam classique évalue mal la capacité d'evaluer des données biologiques par eux même
- Les biologistes non expert de l'informatique est une challenge vu le nombre important de notions a apprendre utilisation d'un ordi, gestion de projet, statistique. Il faut décomposer ces notions en 5 quadrimestre continue
- Le logiciel reste vu comme pointu et diffficile d'utilisation (SUS).
- l'evaluation continue et l'analyse de projet via des grilles critérié semble une approche intéressante pour juger de la capacité des étudiant à bosser.
- la catégorisation des étudiants démontre une grande diversité de profils. Premier élément vers une pédagogie différencié vers

References

Auker, Linda A., and Erika L. Barthelmess. 2020. "Teaching R in the undergraduate ecology classroom: approaches, lessons learned, and recommendations." *Ecosphere* 11 (4): e03060. https://doi.org/https://doi.org/10.1002/ecs2.3060.

Baker, Monya. 2016. "1,500 Scientists Lift the Lid on Reproducibility." *Nature* 533 (7604): 452–54. https://doi.org/10.1038/533452a.

Banks, George C., James G. Field, Frederick L. Oswald, Ernest H. O'Boyle, Ronald S. Landis, Deborah E. Rupp, and Steven G. Rogelberg. 2019. "Answers to 18 Questions About Open Science Practices." *Journal of Business and Psychology* 34 (3): 257–70. https://doi.org/10.1007/s10869-018-9547-8.

Baumer, Ben, Mine Cetinkaya-Rundel, Andrew Bray, Linda Loi, and Nicholas J. Horton. 2014. "R Markdown: Integrating A Reproducible Analysis Tool into Introductory Statistics." *Technology Innovations*

in Statistics Education 8 (1). https://doi.org/10.5070/t581020118.

Byers, J C, A Bittner, and S Hill. 1989. "Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary? In A." In.

Çetinkaya-Rundel, Mine, and Victoria Ellison. 2021. "A Fresh Look at Introductory Data Science." *Journal of Statistics and Data Science Education* 29 (sup1): S16–26. https://doi.org/10.1080/10691898.2020.1804497.

Çetinkaya-Rundel, Mine, and Colin Rundel. 2018. "Infrastructure and Tools for Teaching Computing Throughout the Statistical Curriculum." *American Statistician* 72 (1): 58–65. https://doi.org/10.1080/00031305.2017.1397549.

Donoho, David. 2017. "50 Years of Data Science." *Journal of Computational and Graphical Statistics* 26 (4): 745–66. https://doi.org/10.1080/10618600.2017.1384734.

Estrellado, Ryan A., Emily A. Bovee, Jesse Mostipak, Joshua M. Rosenberg, and Isabella C. Velásquez. 2020. *Data science in education using R*. London, England: Routledge. https://datascienceineducation.com/.

Fiksel, Jacob, Leah R. Jager, Johannna S. Johanna S Hardin, and Margaret A. Taub. 2019. "Using GitHub Classroom To Teach Statistics." *Journal of Statistics Education* 27 (2): 110–19. https://doi.org/10.1080/10691898.2019.1617089.

Freeman, Scott, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. "Active learning increases student performance in science, engineering, and mathematics." *Proceedings of the National Academy of Sciences of the United States of America* 111 (23): 8410–15. https://doi.org/10.1073/pnas.1319030111.

Hart, Sandra G., and Lowell E. Staveland. 1988. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." *Advances in Psychology* 52 (C): 139–83. https://doi.org/10.1016/S0166-4115(08)62386-9.

Hsing, Courtney, and Vanessa Gennarelli. 2019. "Using GitHub in the Classroom Predicts Student Learning Outcomes and Classroom Experiences: Findings from a Survey of Students and Teachers." In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 672–78. SIGCSE '19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287324.3287460.

Krathwohl, David R. 2002. "A Revision of Bloom's Taxonomy: An Overview." *Theory Into Practice* 41 (4): 212–18. https://doi.org/10.1207/s15430421tip4104_2.

Larwin, Karen, and David Larwin. 2011. "A Meta-Analysis Examining the Impact of Computer-Assisted Instruction on Postsecondary Statistics Education." *Journal of Research on Technology in Education* 43 (3): 253–78. https://doi.org/10.1080/15391523.2011.10782572.

Onwuegbuzie, Anthony J., and Vicki A. Wilson. 2003. "Statistics Anxiety: Nature, etiology, antecedents, effects, and treatments–a comprehensive review of the literature." *Teaching in Higher Education* 8 (2): 195–209. https://doi.org/10.1080/1356251032000052447.