9th International Conference on Computational Social Science IC^{2}S^2
July 17-20, 2023, Copenhagen, Denmark

In-class Data Analysis Replications: Feasibility, Student Learning Outcomes, and Challenges

Keywords: reproducibility, data analysis, education, open science, citizen science

Extended Abstract

Introduction. Across a number of fields, researchers have been raising caution about the low reproducibility rates of scientific publications [1]. The focus on novel, confirmatory, and statistically significant results leads to substantial bias in the scientific literature. Replication challenges exist throughout the entire pipeline of a scientific study [3], from the intent of a study (including research question, experimental design, and analysis plan) and what was performed in the conduct of the study (data collected), all the way to the very last mile—data analysis and asserting conclusions. A recent body of work [2, 4] has proposed a possible step towards a solution: educating undergraduate and graduate students to perform data analysis replications. Universities are well positioned to introduce replications as class assignments in methods training in order to establish a culture of reproducibility. However, despite the postulated benefits for both students and science, it is unclear whether and how data analysis replications can be incorporated into large university classes and what impact the data analysis replication tasks could have on the students. What are the educational benefits of data analysis replication? What do students expect, and how can an in-class replication activity turn out?

Methods. To better understand the challenges and benefits of this approach, we incorporated data analysis replications in the project component of a large data analysis course taught at a major university. Based on a set of surveys conducted over the course of one semester, we aim to understand students' expectations about the difficulty of the exercise before performing the replication and their impressions of how hard the task actually was once completed. We aim to identify how students' expectations about the difficulty of replication tasks change when they perform actual replications.

In particular, as part of the data analysis replication exercise, pre-surveys recorded the expectations about the time required, the difficulty of replicating findings from data science papers, and the perceived reproducibility of papers in the field. Students answered a questionnaire to measure expectations on how easy and time-consuming it will be to reproduce. Students recorded their expectations for the likelihood that each predefined result will replicate, along with their expectations about the time it will take for each stage of the replication. Then, students performed the replications. Each student was assigned one paper. We specified two figures or tables to replicate, a basic one and an advanced one. Students recorded their results and recorded how long things actually took in post-surveys. This is compared with their expectations from before they started.

Contributions and results. We present the following contributions:
(1) A possible design of a large-scale in-class replication exercise. We establish the feasibility of a large class (354 students) performing data analysis replication exercise, in a real-world educational scenario. We describe one possible way of integrating a data analysis replication activity into an existing class.

(2) Pre-registered findings. Through the conducted in-class study, we then report pre-registered findings about the discrepancies between expectations and reality of data analysis replication that let us understand how the activity impacts students. We find that there is a significant difference between the time students take to perform the data analysis replication and the time they expect to take ($p = 0.030$), with students, on average, expecting to take 9.01 hours, and taking 10.53 hours, as illustrated in Fig. 1a and Fig. 1b. Furthermore, we discover discrepancies between the predicted and the true distribution of time spent on the three core activities (data wrangling, data analysis, and interpretation). Students spent more time than expected on data analysis and interpretation, which include performing data analysis and modeling, as well as evaluating results and comparing them with the results in the paper, interpreting findings, and redoing analysis, if necessary. Although taking more time than expected, data analysis replication activity was not perceived as significantly more challenging than expected. Students took time iteratively redoing the data analyses and interpreting the results, which was perceived as time-consuming, although not technically challenging.
(3) Exploratory findings. Through the conducted in-class study, we also report exploratory analyses of open-text responses that let us describe how this activity is perceived by the students. Qualitatively analyzing students' responses, we investigate the difficulties students experienced during the replication activity. In particular, we identify four frequent topics: poor description, expertise requirements, limited resources, and time requirements (the computation time and resources required to process large datasets represented a limitation for students working with personal laptops). Poor description, the most common topic, is reported in 60% of the responses. Students mentioned issues such as too little detail on the data analysis process and the pre-processing steps, missing description of the parameters used in the modeling (e.g., size of the random forest model), and limited information about the data origin and format. This issue was summarized by one student as: "[...] *it's almost a guessing game as to what method or inclusion I might be doing differently. This lack of hints was fairly difficult to navigate*".

Discussion. In this study, we demonstrate the potential of in-class data analysis replication exercises, which can be beneficial for both students and science. Our insights about the discrepancies between students' expectations and true outcomes will be informative for future efforts aiming to incorporate data analysis replications into existing educational practices. We conclude by discussing "lessons learned" and we provide reflections on cost-benefit of incorporating data analysis replication exercises into in-class scenarios.

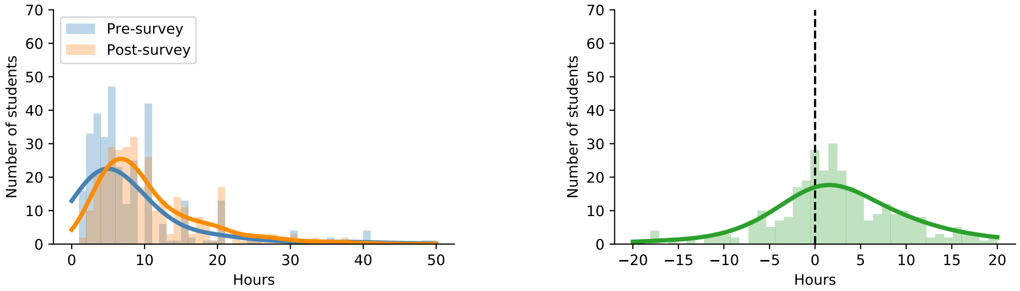
References

[1] M. Baker. Reproducibility crisis. *Nature*, 2016.

[2] J. M. Hofman, D. G. Goldstein, S. Sen, F. Poursabzi-Sangdeh, J. Allen, L. L. Dong, B. Fried, H. Gaur, A. Hoq, E. Mbazor, N. Moreira, C. Muso, E. Rapp, and R. Terrero. Expanding the scope of reproducibility research through data analysis replications. *Organizational Behavior and Human Decision Processes*, 2021.

[3] P. Patil, R. D. Peng, and J. T. Leek. A visual tool for defining reproducibility and replicability. *Nature Human Behaviour*, 2019.

[4] D. S. Quintana. Replication studies for undergraduate theses to improve science and education. *Nature Human Behaviour*, 2021.



(a)

(b)

Figure 1: **Expectations vs. reality of a data analysis replication exercise**. (a) Across students (y-axis), the histogram of the expected number of hours (x-axis) data analysis replication is expected to take in the pre-survey (in blue), and the actual number of hours that replication took (in orange). (b) Across students (y-axis), the histogram of the difference (x-axis) between the number of hours data analysis replication took, and the number of hours data analysis replication the student expected to take.