ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202211.00463

# Toward Training and Assessing Reproducible Data Analysis in Data Science Education Postprint

**Authors:** Yu, Bei, Hu, Xiao

**Date:** 2022-11-29T00:00:00+00:00

## Abstract

Reproducibility is a cornerstone of scientific research. Data science is not an exception. In recent years scientists were concerned about a large number of irreproducible studies. Such reproducibility crisis in science could severely undermine public trust in science and science-based public policy. Recent efforts to promote reproducible research mainly focused on matured scientists and much less on student training. In this study, we conducted action research on students in data science to evaluate to what extent students are ready for communicating reproducible data analysis. The results show that although two-thirds of the students claimed they were able to reproduce results in peer reports, only one-third of reports provided all necessary information for replication. The actual replication results also include conflicting claims; some lacked comparisons of original and replication results, indicating that some students did not share a consistent understanding of what reproducibility means and how to report replication results. The findings suggest that more training is needed to help data science students communicating reproducible data analysis.

## Full Text

## Preamble

**RESEARCH PAPER**
**Toward Training and Assessing Reproducible Data Analysis in Data Science Education**

Bei Yu[1] & Xiao Hu[2]†
[1]Syracuse University, Ringgold Standard Institution
320 Hinds Hall, Syracuse, New York 13244-1100, USA
[2]University of Hong Kong, Ringgold Standard Institution
Room 209, Runme Shaw Building, Pokfulam, Hong Kong, China

**Keywords:** Data science education; Reproducibility; Reproducible data analysis; Communication; Action research

**Citation:** B. Yu & X. Hu. Toward training and assessing reproducible data analysis in data science education. *Data Intelligence* 1(2019), 333-344. doi: 10.1162/dint_a_{00053}

**Received:** May 27, 2019; **Revised:** August 18, 2019; **Accepted:** August 22, 2019

## ABSTRACT

Reproducibility is a cornerstone of scientific research, and data science is no exception. In recent years, scientists have expressed concern about a large number of irreproducible studies. Such a reproducibility crisis in science could severely undermine public trust in science and science-based public policy. While recent efforts to promote reproducible research have focused primarily on established scientists, student training has received far less attention. In this study, we conducted action research with data science students to evaluate their readiness to communicate reproducible data analysis. The results show that although two-thirds of students claimed they could reproduce results from peer reports, only one-third of reports provided all necessary information for replication. The actual replication results also included conflicting claims, and some lacked comparisons between original and replication results, indicating that some students did not share a consistent understanding of what reproducibility means or how to report replication results. These findings suggest that data science students need more training to help them communicate reproducible data analysis effectively.

## 1. INTRODUCTION

Reproducibility—the ability to replicate an experiment and obtain the same result—is a cornerstone of scientific research [1]. Science relies on reproducibility to weed out unreliable claims and to self-correct when scientific misconduct occurs. The process of providing access to experimental materials with sufficient precision necessary for replication has long been considered "a deeply established part of the scientific process" [2].

However, in recent years, both scientists and the public have become increasingly concerned about reproducible research. A *Nature* survey of more than 1,500 scientists asked whether there was a reproducibility crisis in science; 52% responded "yes, a significant crisis," and 38% said "yes, a slight crisis" [3]. The survey also found that more than 70% of researchers had tried and failed to reproduce another scientist' s experiments, and more than half had failed to reproduce their own experiments.

While such concerns may have been more prominently reported in natural science disciplines such as biomedicine (e.g., Kaiser [4]), the increasing use of big data and data analytics in virtually all disciplines has made this crisis closely relevant to data science as an emergent discipline in its own right. Indeed, many previous replication studies have focused on reproducing data analysis processes and results in various disciplines, including empirical economics [5], epidemiology [6], and psychology [7].

Discussions dedicated to reproducibility in data science have also emerged in recent years from various perspectives [8]. Unlike many other domains where reproducibility has been discussed primarily as a science policy problem—particularly in the contexts of tenure and promotion review and the publication review process [2, 9, 10]—discussions in data science have paid special attention to reproducibility as an infrastructure problem [11, 12], which involves how to share code and data as well as how to control the computing environment to support replication studies. Other discussions frame reproducibility as a communication problem, arguing that although the infrastructure problem is largely solved, the reproducibility problem persists due to failed communication [13].

To date, discussions on improving reproducibility have mainly targeted established researchers. In contrast, data science students are rarely mentioned in these discussions, let alone receiving dedicated training on the concept of reproducibility and how to practice it in their own data analyses. Despite some pioneering efforts such as Howe [12], reproducibility training and assessment in data science education has been largely neglected, especially among undergraduate and master' s students in professional schools such as iSchools, probably because these students are typically considered non-research-oriented.

However, data science is science, and data analysis is research. A typical data science curriculum often includes significant coursework in data analytics, which exemplifies research skills in data science. Data analytics requires students to dive into data, find patterns, and provide evidence to prove that these patterns are reliable and useful. Therefore, training in Reproducible Data Analysis (RDA) should be an indispensable component of data science education. Students should understand the concept of reproducibility and use it to guide their data analysis design and result reporting.

As a first step toward training RDA, this study focuses on the communication aspect and investigates students' current understanding of reproducibility, whether they can communicate RDA effectively, and if not, what skills are missing and what types of training could be helpful.

We designed a data analysis task with controlled computing infrastructure and asked each student in a data mining class to carry out this task, write a report of the data analysis process and results, and then replicate another student' s analysis based solely on that student' s report. We then designed a content analysis schema to code the completeness of descriptions and replication outcomes. Specifically, this study aims to answer two research questions: (1) How reproducible are students' data analysis reports, as perceived by peers? (2) Do students share a consistent understanding of the concept of reproducibility?

## 2. RELATED WORK

For the purpose of this study, we identified two perspectives in the literature on reproducible data analysis (RDA): the first considers RDA an infrastructure problem, while the second considers RDA a communication problem. Below we review literature from both perspectives.

### 2.1 RDA as an Infrastructure Problem

Prior studies have identified access to data and code as the major barrier to reproducibility in data science [8]. Therefore, open access efforts have been made to develop repositories, protocols, and platforms that allow researchers to deposit and share data and code, such as the Inter-university Consortium for Political and Social Research (ICPSR) and GitHub.

As computer code execution depends on computing environment factors such as operating systems and software versions, Howe [12] argued that cloud computing is an ideal solution for reproducibility because virtualization provides controlled computing environments for replicating data analysis. Stodden and Miguez [11] also reviewed best practices for reproducible research regarding software infrastructure and environments and concluded that current technologies are sufficient to provide infrastructure support for RDA.

### 2.2 RDA as a Communication Problem

As Peng [13] argued, although the infrastructure problem is largely solved for disseminating reproducible research, the language and communication problem still exists and is actually a bigger and deeper problem. He drew an analogy that sharing code and data to communicate data analysis is like sharing an audio recording to communicate music, which is not as effective as the original score for understanding how the musician wrote the piece. This is because the recording provides too much information–although a trained ear can reconstruct the score, it would be a difficult and time-consuming task.

This viewpoint is supported by studies such as Dewald et al. [5], which found that inadvertent errors were common in published papers and computer programs, meaning that even sharing data and code could not guarantee reproducibility. The further need for communicating data analysis is also demonstrated in various efforts to develop new documentation tools that can generate dynamic documents consisting of text, code, and data for the convenience of reproducing experiments [14].

The idea itself is not new. Knuth [15] proposed WEB, a programming language that would allow programmers to consider a computer program a document to explain to human beings what we want a computer to do, rather than a script to instruct a computer what to do. Fast forward to today, some data analysis communities have developed new documentation tools to make replication easier. For example, R, one of the most popular data analysis tools [16], provides the "knit" function to allow users to write paragraphs of descriptions and explanations along with R source code into one R Markdown (RMD) file [17]. After weaving/knitting, a Word or PDF file is generated that includes the source code, the accompanying explanations, and the output of each block of source code. The iPython Notebook tool provides similar functions.

However, the availability of effective communication tools does not necessarily guarantee effective communication. We can identify two different situations depending on the purpose of information sharing. In the first situation, information providers are demanded to share, but they do not necessarily have the need to share. This is the situation mainly discussed in the current reproducibility literature, in which researchers are asked to share for the sake of reliable research.

In the second situation, information providers genuinely need to share information to seek help. For example, if a learner asks a data analysis question on Stack Overflow and does not provide sufficient information for others to replicate the problem, or if the information is unclear or excessive, they might not receive an answer. Therefore, the ability to communicate reproducible data analysis is needed not only for research in data science but also for learning in data science.

In fact, the problem of poor communication of reproducible data analysis is best illustrated by relevant discussions on community question-answering sites, among which Stack Overflow is one of the most popular. Stack Overflow maintains a help page [18] that defines three criteria for reproducible questions: (1) minimal—use as little code as possible to reproduce the same problem; (2) complete—provide all information needed to reproduce the same problem; and (3) verifiable—test the code you are going to provide to make sure it reproduces the problem.

## 3. METHOD

Viewing reproducibility from the communication perspective, we designed the following experiment to evaluate a convenience sample of iSchool students' understanding of reproducibility and their ability to communicate RDA. The experiment was carried out in a data mining course and thus qualified as action research [19].

### 3.1 Data Sample

A graduate-level text mining course at a reputable iSchool in the US was chosen as a convenience sample. All 22 registered students participated in this study, including 4 doctoral students and 18 Master's students. The majority of students were from the information management program (16, 73%), with 6 (27%) others from accounting, finance, library science, linguistics, and political science. As data science is highly interdisciplinary, students in data science courses typically come from a wide range of fields. It is noteworthy that this study is exploratory in nature, and due to the moderate sample size, the results are not intended for generalization.

### 3.2 Experiment Design

Students were first given an individual assignment to test the hypothesis that stemming can help sentiment classification (hereafter noted as the "Stemming hypothesis" ). This task required only a basic understanding of data mining algorithms, which helped control possible biases introduced by students' familiarity with the algorithms. Similarly, this experiment did not involve programming, as students were required to use a graphic-based data analysis tool (Weka GUI). To control the computing infrastructure, students were required to use the same data analysis tool (Weka), the same algorithm (SMO), and the same movie review dataset provided by the instructor. Both the assignment and the replication task (see below) were conducted in lab sessions of the course where all students used computers in an iSchool lab. All software packages had been pre-installed on these computers, ensuring a consistent computational environment.

This assignment task involved two steps: vectorization and classification. Students were given the freedom to change parameters in each step, such as vocabulary size, word weighting schemes, and the kernel function in SMO. Students were asked to report whether the Stemming hypothesis was consistently confirmed or disconfirmed across different parameter settings. Students were also reminded to provide necessary information for others to replicate their analyses. It is noteworthy that, as our goal was to test students' current understanding of reproducibility, no additional training was provided on what information should or should not be reported.

After students submitted their reports, the reports were printed with author names redacted. The reports were then randomly ordered and assigned ID numbers from 1 to 22. In the next class (one week later), students were randomly given another student's report and asked to independently replicate the analyses described in that report.

Upon completing the replication, students were asked to write a short report on their replication results and post it to a discussion forum on Blackboard. Specifically, students were prompted to answer the following questions in their replication reports:

**Q1:** Were you able to replicate the results?
**Q2a:** If yes, did the replicated results support the Stemming hypothesis?
**Q2b:** If no, what information was missing for replicating the results? Were you able to recover the missing information through trials?

By analyzing answers to these questions, we expected to evaluate students'ability to communicate reproducible data analyses and the extent to which students shared a consistent understanding of what reproducibility meant in this particular data analysis task. If a shared understanding existed, we should see consistent answers to the above questions; otherwise, self-conflicting answers would occur. For example, a student who did not fully understand reproducibility might claim that a report was reproducible while simultaneously reporting that some information was missing or drawing an opposite conclusion about the Stemming hypothesis.

## 4.1 Coding Scheme on Reproducibility

The replication reports were downloaded from Blackboard. Student answers were independently annotated by both authors. The complete annotation schema with category definitions and examples is presented in Table 1 , where students are referred to as "evaluator." Answers to Q1 were inductively categorized into three types: Yes, No, and Partial. For Q2a, five types of answers occurred: (1) both original and replication reports supported the hypothesis (co-support); (2) both refuted the hypothesis (co-refute); (3) replication supported the hypothesis but the original report did not (own-support); (4) replication refuted the hypothesis but the original report did not (own-refute); and (5) the replication report did not provide relevant information (unknown). This classification approach helps us assess not only reproducibility but also students' ability to communicate replication results. Q2b received three types of answers: NM (No Missing information), M-R (Missing information but Recovered), and M-NR (Missing information, Not Recovered).

Intercoder reliability between the two coders was measured by Cohen' s kappa coefficient, which achieved 1.00 for Q1, 0.74 for Q2a, and 0.80 for Q2b, indicating substantial to almost perfect agreement [20]. Disagreements were resolved through discussion.

## 4.2 Reproducibility Result

The annotation result is reported in Table 2 . Note that one student' s answer to Q2a and Q2b was coded as "incomprehensible."

## 4.3 Peer-Reported Level of Reproducibility

As a direct answer to RQ1—how reproducible students' data analysis reports are as perceived by peers—Table 2 shows that 15 out of 22 students reported they could reproduce results from original reports. Therefore, the peer-reported reproducibility level is 68% (15/22).

## 4.4 Shared Understanding of Reproducibility

RQ2 asks whether students shared a consistent understanding of reproducibility. We answer this question by examining results from two aspects: (1) consistency between reported reproducibility and reported missing information, and (2) consistency between conclusions in original and replication reports.

**Consistency between reported reproducibility and missing information:** Excluding one incomprehensible answer, 32% (7/22) of replications reported no missing information, whereas 64% (14/22) reported that some critical information was missing, including 6 recovered (27%) and 8 not recovered (36%). Compared to the 68% peer-reported reproducibility in RQ1, only 32% of original reports provided all necessary information for replication. Among the 15 reports that claimed reproducibility, only 7 did not miss information, 5 claimed missing information but recovered it, and 2 did not recover the information. The last two cases are self-conflicting, as evaluators could not recover missing information but still claimed to have reproduced the results in the original reports. Other self-conflicting answers explicitly claimed reproducibility while simultaneously reporting some results that did not match those in the original reports.

Below are examples of self-conflicting answers:

> "Able to reproduce the result: Yes; Are the results same: No"

> "I am able to reproduce the 'stemming for sentiment analysis' experiment. Some of the results I get are the same as the result on report but some are not. For unigram term frequency, I could not see the number of term frequency in the report so I don' t know what I need to put when I reproduce the result."

> "I am able to reproduce the 'stemming for sentiment analysis' experiment. ···However, I get different confusion matrix for all the four models."

Overall, a total of 5 answers (24%) contained self-conflicting statements regarding reproducibility, indicating problems in shared understanding of reproducibility.

**Consistency between conclusions in original and replicated results:** We then examined whether replication reports drew the same conclusions as original reports. Table 3 shows the distribution of conclusion accordance, which reveals 8 "co-support," 4 "co-refute," and 2 "unknown" among evaluators who claimed results were replicated. For the 6 irreproducible cases, only 1 reported their own result and the other 5 were "unknown." Therefore, only 12 students (55%) were able to replicate and report the replication result, including 8 "co-support" and 4 "co-refute." One student was not able to replicate but was able to communicate the replication result clearly. The other 9 students (41%) did not provide clear conclusions on their replication results, suggesting they may lack knowledge about how to communicate replication results.

## 5. DISCUSSION

In this section, we summarize the results regarding common issues in communicating data analysis in a reproducible manner and offer actionable suggestions.

First, although 68% of original data analysis reports written by students were claimed to be reproducible by peers, only 32% of replications reported no missing information. One important skill for communicating reproducible data analysis is assessing what information is necessary and what is not, in order to provide just sufficient information for replication. This finding warrants the necessity of enhancing RDA training in data science programs and courses. A possible training approach would be to have students try to replicate others' work as an opportunity to experience the information needs involved in data analysis replication.

Second, 24% of replication reports contained self-conflicting claims, indicating a lack of consistent understanding of what reproducibility means. Furthermore, 41% of replication reports did not draw appropriate conclusions based on comparing original and replication results, indicating a lack of skills to communicate replication results. The concept of reproducibility may seem intuitive, but the reality is more complex. Therefore, data science curricula need to explicitly cover the definitions and principles of RDA as one of the core topics and learning outcomes.

Third, based on the results, all irreproducible results reported missing information in the original reports that was not recoverable. In contrast, most (13 out of 15) reports that claimed reproducibility reported either no missing information in the original reports or that such information was recoverable. This once again proves the importance of documenting all critical steps in data analysis procedures, including parameter settings and pre-processing. In educational settings, realizing this importance should be included as a learning outcome. In other words, data science educators need to train students to provide detailed information necessary for replication, rather than neglecting replicability or simply assuming other people would know what to do.

Unlike specific knowledge or analytical skills, communication skills require relatively long-term training. It is thus desirable for data science programs to establish consistent teaching guidelines on these skills across multiple courses and capstone projects. Seminars and workshops on reproducibility are also viable ways to train students, especially those in related programs with few data science courses.

## 6. CONCLUSION

We conducted an experiment in the classroom to evaluate data science students' ability to communicate reproducible data analysis. The results show that two-thirds of original reports were perceived as reproducible by peer students; however, about a quarter of students loosely defined reproducibility and accepted irreproducible results as reproducible. Two-thirds of the original reports lacked necessary information for replication, and one-third of replication reports did not compare original results with replication results, indicating that students need more training on understanding the rigorous definition of reproducibility and practical skills in communicating reproducible data analysis. Based on these results, we discuss skills essential for communicating reproducible data analysis and propose suggestions for facilitating student learning in this important aspect of research and practice in data science.

## AUTHOR CONTRIBUTIONS

B. Yu (byu@syr.edu) proposed the research problems, performed the research, collected and analyzed the data, and wrote and revised the manuscript. X. Hu (xiaoxhu@hku.hk) proposed the research problems, analyzed the data, and wrote and revised the manuscript.

## REFERENCES

[1] National Institutes of Health. Providing new incentives to foster open science and increase reproducibility. Available at: https://obssr.od.nih.gov/providing-new-incentives-foster-open-science-increase-reproducibility/.

[2] G. King. Publication, publication. *PS: Political Science and Politics* 39(1)(2006), 119–125. doi: 10.1017/S1049096506060252.

[3] Nature. 1500 scientists lift the lid on reproducibility. Available at: http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970.

[4] J. Kaiser. Rigorous replication effort succeeds for just two of five cancer papers. Available at: http://www.sciencemag.org/news/2017/01/rigorous-replication-effort-succeeds-just-two-five-cancer-papers.

[5] W.G. Dewald, G.J. Thursby, & R.G. Anderson. Replication in empirical economics: The Journal of Money, Credit and Banking project. *The American Economic Review* 76(4)(1986), 587-603. doi: 10.2753/PET1061-

[6] R. D. Peng, F. Dominici, & S.L. Zeger. Reproducible epidemiologic research. *American Journal of Epidemiology* 163(9)(2006), 783–89. doi:10.1093/aje/kwj093.

[7] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* 349(6251)(2015), 943. doi:10.1126/science.aac4716.

[8] R.D. Peng. Reproducible research in computational science. *Science* 334(6060)(2011), 1226–1227. doi: 10.1126/science.1213847.

[9] G. King. Replication, replication. *PS: Political Science and Politics* 28(3)(1955), 444–452. doi:10.2307/420301.

[10] D.L. Donoho. An invitation to reproducible computational research. *Biostatistics* 11(3)(2010), 385–88. doi: 10.1093/biostatistics/kxq028.

[11] V. Stodden, & S. Miguez. Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. *Journal of Open Research Software* 2(1)(2004), p.e21. doi: 10.5334/jors.ay.

[12] B. Howe. Communicating data science results. MOOC course on Coursera. Available at: https://www.coursera.org/learn/data-results/home/welcome.

[13] R. Peng. Disseminating reproducible research is fundamentally a language and communication problem. Available at: http://simplystatistics.org/2016/05/13/reproducible-research-language/.

[14] R. Gentleman, & D.T. Lang. Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics* 16(1)(2012), 1–23. doi: 10.2307/27594227.

[15] D. E. Knuth. Literate programming. *The Computer Journal* 27(2)(1984), 97–111. doi: 10.1093/comjnl/27.2.97.

[16] K. Dnuggets. Python eats away at R: Top software for analytics, data science, machine learning in 2018: Trends and analysis. Available at: https://www.kdnuggets.com/2018/05/poll-tools-analytics-data-science-machine-learning-results.html.

[17] C. Gandrud. *Reproducible research with R and R studio*. Boca Raton, FL: CRC Press, 2013. isbn:

[18] Stackoverflow. How to create a minimal, complete, and verifiable example – help center –stack overflow. Available at: http://stackoverflow.com/help/mcve.

[19] E. T. Stringer. *Action research in education*. Upper Saddle River, NJ: Pearson Prentice Hall, 2008. isbn:

[20] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1960), 37–46. doi: 10.1177/001316446002000104.

## AUTHOR BIOGRAPHY

**Bei Yu** is an Associate Professor at the School of Information Studies at Syracuse University. She is also the Faculty Lead of the Certificate of Advanced Study in Data Science. Her research area is applied natural language processing, especially sentiment and opinion analysis. Before joining Syracuse, she was a postdoctoral researcher at the Kellogg School of Management at Northwestern University. Dr. Yu earned her PhD degree from the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign. She holds her Bachelor’ s and Master’ s degrees in Computer Science.

**Xiao Hu** is an Associate Professor at the Faculty of Education, the University of Hong Kong, China. She is also the Program Director of Bachelor of Arts and

Science in Social Data Science. Her main research interests are applications of data mining and machine learning in the domains of Information, Education, and Culture, including learning analytics, music information retrieval, affective computing, and data science. Dr. Hu holds a PhD degree in Library and Information Science, a Master' s degree in Computer Science, and a Bachelor' s degree in Electronics and Information Systems.

*Note: Figure translations are in progress. See original paper for figures.*

*Source: ChinaXiv —Machine translation. Verify with original.*