Change in self-reported confidence in data cleaning and analysis skills among students enrolled in a graduate level course on quantitative data management using R

Sheena E. Martenies1

1Department of Health and Kinesiology, University of Illinois Urbana-Champaign

**1. Introduction**

Background here.

The objective of this study was to assess changes in the level of confidence

students enrolled in HK 513 (Fall '24) feel about their ability to conduct

specific data management tasks using R. The analysis compared the self-reported

confidence felt by students at the start of the semester (Week 1) and the end of

the semester (Week 14) to see if there was a change in confidence as a result

of engaging in the course.

**2. Methods**

2.1 Participants and Study Setting

This study included students at the University of Illinois Urbana-Champaign enrolled in a graduate level course (HK 513) introducing key concepts related to quantitative data management using R software. HK 513 is a seated course designed to introduce graduate students to the basic methods involved in cleaning, analyzing, and visualizing quantitative data. Most students in the course were pursuing a Master of Public Health degree, though graduate students from across were welcome to enroll in the course. Data collected occurred during the Fall 2024 semester. Students were eligible for this study if they were enrolled in the course during this time.

Because data were collected for course demonstration purposes, this work is not considered research. These results will not be disseminated outside this course. Students received course credit for completing the questionnaires are part of their weekly participation assignment. All data provided by students was deidentified prior to analysis.

2.2 Survey Instrument

During Week 1 and Week 14 of the course, students were asked to complete a questionnaire outlining their level of experience using various software programs for quantitative data management and the level of confidence completing data management tasks. These questions were designed to align with the course objectives as outlined in the syllabus. Students were asked to indicate their level of experience using R/S Studio, Python, SAS, Stata, SPSS, Excel, Tableau, and Power BI from 1 to 5, where 1 represented “no experience” and 5 represented “a great deal of experience.” They were also asked to indicate how confident they felt completing the following tasks using R: A) using software to organize quantitative data; B) using software to clean quantitative data; C) using software to analyze quantitative data; D) using software to visualize quantitative data; E) write a methods section that conveys their data analysis steps; F) write a report that summarize their data analysis results; G) share their work with others so they can reproduce it effectively; and H) find answers to data management questions using resources they find and vet on their own. Responses to these questions could range from 1 to 4, where 1 represented “not confident at all” and 4 represented “very confident.”

For the Week 1 questionnaire, students were asked to identify goals for the course. A part of the Week 14 questionnaire, students were asked if they met their course goals; students could select “yes,” “somewhat,” or “no” for this question.

Our primary outcomes of interest were the changes observed in the level of confidence to complete data management tasks in R over the course of the semester. We were also interested in the level of experience students had with R before the course and in whether or not students felt they met their semester-long goals. We dichotomized our pre-course experience variable as no experience (1) vs. any experience (2-5).

2.3 Data Analysis

Results from hard-copy surveys were entered into an Excel data base by the investigator. A random sample of 10% of the surveys were entered in duplicate to check for data entry errors. To deidentify the data, a random ID between 1 and 200 was assigned to participants.

We first examined the distributions of our experience and confidence variables using histograms and frequency tables. We also use a frequency table to examine the number of students who reported meeting their goals for the semester. To test our main hypothesis that self-reported confidence in performing data management tasks would increase over the course of the semester, were compared pre- and post-course confidence scores using the paired Wilcoxon Signed Rank test. We chose to use a non-parametric test because of the non-Normal distribution of self-reported confidence scores. We considered a p-value of <0.05 to be evidence that the difference between pre- and post-course confidence scores was statistically significant.

All data cleaning, analysis, and visualization tasks were conducted using R version 4.4.1 (R Core Team 2024). Data cleaning tasks were performed using functions from the tidyverse set of packages (Wickham et al. 2019). Data visualizations were created using ggplot2 and the viridis packages (Simon Garnier et al. 2024; Wickham 2016).

**3. Results**

A total of 12 students completed the baseline (pre-class) questionnaire during Week 1 of the semester. Nine students completed the follow-up (post-class) questionnaire during Week 14, resulting in a completion rate of 75%.

The baseline level of experience with R and R Studio as reported by students was low. Of the 12 students enrolled in the course, 5 (42%) reported having no experience with R at all. Only 2 of the 12 students (17%) reported having “much experience” or a “great deal of experience” using R and R Studio for quantitative data management.

Baseline levels of confidence as reported by students on the Week 1 questionnaire were generally low, but we observed statistically significant increases in confidence reported by students in the Week 14 questionnaire for some data management tasks (Table 1). The median levels of confidence for organizing quantitative data using R (Task A), cleaning quantitative data using R (Task B), analyzing quantitative data using R (Task C), and visualizing quantitative data in R (Task D) each increased from 2 (somewhat confident) to 3 (confident; all p values < 0.05). We also saw significant increases in reported confidence in ability to write methods and results sections (Task E and Task F). Median levels of confidence in the ability to share work in a reproducible way (Task G) and findings answers to questions (Task H) also increased, but these increases were not statistically significant.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Table 1. Self-reported level of confidence students feel about their ability to complete quantitative data management tasks in R. | | | | | | | |
|  |  | N (%) | | | |  |  |
| Task | Period | 1: Not confident at all | 2: Somewhat confident | 3: Confident | 4: Very confident | Median (IQR) | p1 |
| A) Use software to organize quantitative data | Pre-class | 3 (25) | 5 (41.7) | 3 (25) | 1 (8.3) | 2 (1.25) | 0.02 |
| Post-class | 0 (0) | 0 (0) | 5 (55.6) | 4 (44.4) | 3 (1) |
| B) Use software to clean quantitative data | Pre-class | 3 (25) | 5 (41.7) | 3 (25) | 1 (8.3) | 2 (1.25) | 0.02 |
| Post-class | 0 (0) | 1 (11.1) | 4 (44.4) | 4 (44.4) | 3 (1) |
| C) Use software to analyze quantitative data | Pre-class | 4 (33.3) | 3 (25) | 4 (33.3) | 1 (8.3) | 2 (2) | 0.03 |
| Post-class | 0 (0) | 1 (11.1) | 4 (44.4) | 4 (44.4) | 3 (1) |
| D) Use software to visualize quantitative data | Pre-class | 4 (33.3) | 4 (33.3) | 3 (25) | 1 (8.3) | 2 (2) | 0.03 |
| Post-class | 0 (0) | 2 (22.2) | 4 (44.4) | 3 (33.3) | 3 (1) |
| E) Write a methods section | Pre-class | 4 (33.3) | 4 (33.3) | 3 (25) | 1 (8.3) | 2 (2) | 0.03 |
| Post-class | 0 (0) | 0 (0) | 7 (77.8) | 2 (22.2) | 3 (0) |
| F) Write a report summarizing results | Pre-class | 3 (25) | 4 (33.3) | 3 (25) | 2 (16.7) | 2 (1.25) | 0.04 |
| Post-class | 0 (0) | 0 (0) | 6 (66.7) | 3 (33.3) | 3 (1) |
| G) Share work with others in a reproducible way | Pre-class | 4 (33.3) | 4 (33.3) | 3 (25) | 1 (8.3) | 2 (2) | 0.11 |
| Post-class | 0 (0) | 3 (33.3) | 3 (33.3) | 3 (33.3) | 3 (2) |
| H) Find answers to data management questions | Pre-class | 4 (33.3) | 3 (25) | 4 (33.3) | 1 (8.3) | 2 (2) | 0.06 |
| Post-class | 0 (0) | 1 (11.1) | 4 (44.4) | 4 (44.4) | 3 (1) |
| 1 p value is for the paired Wilcoxon Ranked Sign test (alpha = 0.05, two-tailed test) | | | | | | | |

Of the 9 students who completed the Week 14 questionnaire, 8 (88%) reported meeting their course goals for the semester and 1 (11%) reported somewhat meeting their goals for the semester (Figure 1). None of the students reported not meeting their goals.

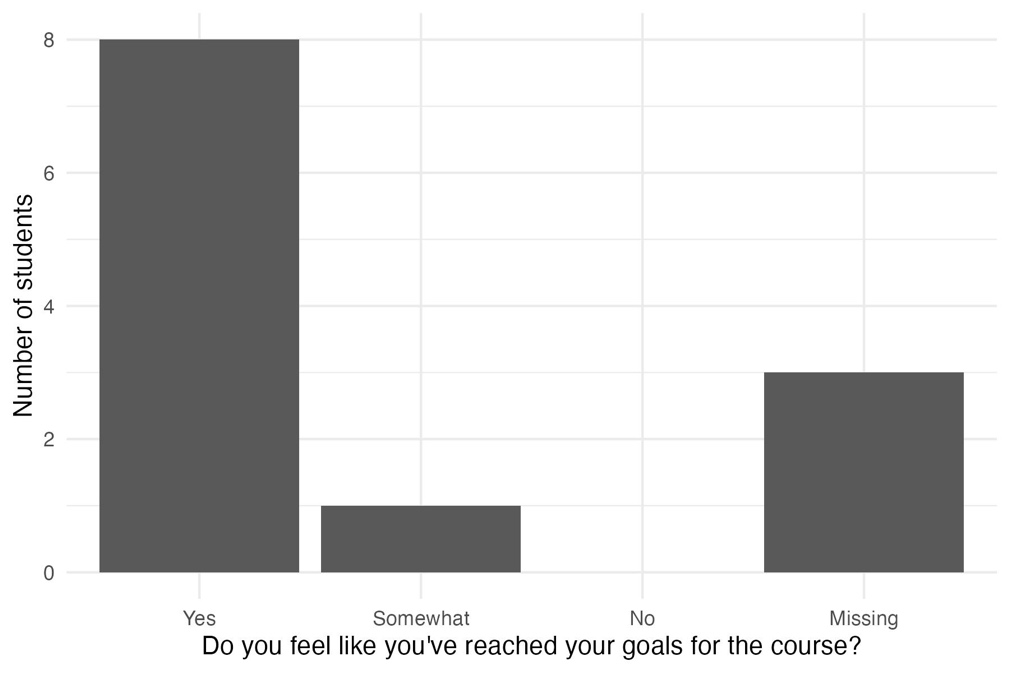


Figure 1. Student responses to the Week 14 survey question about whether they have reached their goals for the course.

**4. Discussion**

This is the section where you would contextualize your results in the broader lecture.

Here is where you would highlight the strengths and mention the limitations of your work.

**5. Conclusions**

Here is where you would emphasize the main take-away point of your work.

**Acknowledgements**

Thank you to the students of HK 513 (Fall 2024) for their participation this semester and their contributions to this work.

The code and deidentified data used in this analysis are available on the authors GitHub page: <https://github.com/smartenies/hk513_fall24_repro_deliverable>.

**References**

R Core Team. 2024. R: A language and environment for statistical computing.

Simon Garnier, Noam Ross, Robert Rudis, Antônio P. Camargo, Marco Sciaini, Cédric Scherer. 2024. viridis(Lite) - Colorblind-Friendly Color Maps for R.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. 2019. Welcome to the Tidyverse. Journal of Open Source Software 4:1686; doi:10.21105/joss.01686.