Miriam A. Golden
Mamoor Ali Khan

Standard Operating Procedures for Research Assistants

Version 1.2; 6 July 2023

Contents

Overview

a. **Purpose:** This SOP outlines the guidelines and procedures for Research Assistants (RAs) who work remotely on academic projects under the supervision of PI Miriam A. Golden and designated co-PIs. The main goal is to provide RAs information about how to organize project materials and workflow.

b. **Reproducibility:** Your work contributes to scientific knowledge. For scientific knowledge to cumulate, it must be reproducible by others.

c. **Transparency:** You are working as part of a team. Everything you do affects all other team members. For your team to be able to understand your work, all steps must be transparent. As you work, you should assume that someone else will eventually need to make sense of your work and will not be able to ask you any questions directly. Thus, you should document your work extensively and make sure that it operates self-sufficiently.

d. **Work Load:** A fully reproducible workflow means higher up-front time investment for reduced errors and reduced later work load. RAs should expect to spend as much as 30 percent of their time documenting and annotating their work. You should prioritize transparency and reproducibility over implementing quick and dirty solutions.

e. **Scope:** This SOP applies to all RAs involved in research work on academic projects with PI Golden.

Communication and Reporting

a. **Weekly Check-Ins:** RAs will have weekly check-ins with the PIs to discuss project progress, clarify objectives, and address any questions or concerns. RAs are expected to be available at a standard time every week unless notified by a PI. If an RA will be unavailable, notify the PIs in advance.

b. **Communication Channels:** The primary mode of written communication will be through a dedicated Slack channel. Weekly check-ins will take place via a video conferencing tool (Zoom) at the same location every week (PI Golden's Zoom room).

c. **Response Time:** RAs should respond to communication from a PI and from other team members promptly and maintain open lines of communication. In general, responses are expected within 48 hours, or sooner if a matter is highly urgent. RAs are asked to report problems that might affect other team members immediately via Slack. This might include problems such as an inability to pull from or push to the Github repo, an inability to format a document, or an inability to process data.

d. **Absences:** RAs who will be unavailable for a period of time due to personal or medical reasons should notify the PI, giving as much advance notice as possible.

Project Management

a. **Project Plan:** PIs will provide RAs with a clear project plan outlining objectives, deliverables, and timelines. You are encouraged to provide feedback on aspects of the plan that you consider would benefit from modifying.

b. **Task Assignment:** PIs will assign specific tasks to RAs based on what the PIs understand of the RA's expertise. RAs are encouraged to learn new skills on the job but should inform PIs if they do not yet have the skills to complete a specific task.

c. **Task Reassignment:** If you are assigned a task and discover that your skills are not fully adequate to it, or that it is taking you far longer than expected, you should make the PIs aware of this. The PIs may reassign the task to someone else or guide you in skill acquisition.

d. **Task Prioritization:** RAs should prioritize tasks based on the project plan and seek clarification from PIs if there are conflicting priorities.

e. **Timelines and Deadlines:** RAs are responsible for meeting the agreed-upon deadlines for each task and informing PIs in advance if they anticipate delays. RAs are encouraged to provide feedback if specific timelines or deadlines seem unrealistic.

f. **Task Documentation:** RAs should maintain detailed documentation of their work, including methods, results, and any challenges encountered. The standard place to do this is in a project logbook. Each RA usually maintains a single logbook.

g. **Project Documentation:** Projects typically reside in a dedicated Github repository, to which RAs will be granted access. RAs should pull from the repo before starting work and should push frequently. In some cases, PIs will monitor commits before they are finalized.

h. **Public Posting:** Once a project is complete, it will be put in the public domain. As a result, you should keep in mind that eventually there will need to be a data codebook and that all code underlying everything reported in a publication will have to be reproducible. Even if you are not working directly on writing a codebook, all data will have to be documented by someone when that person writes the codebook. Try to anticipate this and to make sure your work easily permits it. Assume that the person writing the codebook cannot contact you directly for information.

i. **Interoperability:** All the work you do on a project must be designed to run on other computers (i.e. the computers used by the PI and other team members). This requires you build in from the start code that is platform-independent.

Directory and File Structure and Naming

a. **Directory Structure:** Each project will be organized in a hierarchical file structure, with the topmost level typically named for the overall project. Every team member will have an identical copy of the project directories and files.

b. **Directory and File Names:** Directory and file names should not contain any blank spaces, and should only use capital letters if required for readability. The underscore character should be used to divide words for readability as necessary.

c. **Dates:** Dates should be formatted according to the International Organization for Standardization (ISO) guidelines, i.e. yyyymmdd.

d. **Directory and File Ordering:** To the extent possible, directory and file names should be organized sequentially, i.e. using a two-digit leading number system so that they hang in order of operations (e.g. 01_admin, 02_lit_bib, 03_funding, 04_design, 05_pap, 06_irb, 07_analysis, 08_presentations, 09_papers, 10_scripts, etc.).

e. **Master Version:** There should only be a single canonical version of any file; e.g. we should never encounter files such as "data_v1" and "data_v2" or "paper_v1" and "paper_v2_myinitials." If you need to temporarily generate different versions of the same file for a specific purpose, you may wish to branch in Git or to work locally and only push when you have fully resolved an issue.

f. **File Names:** Files should be named for their contents and not for their authors. We should find files such as "results_section.tex" and "female_voters1950.png." We should not encounter files such as "results_joe.tex" or "female_voters1950_Lucas_version.png."

g. **READMEs:** Every directory and subdirectory should contain a plain text "readme" file that describes what is contained in the directory, who assembled the material, when it was assembled, where it was sourced, when the readme was written or updated, and any other essential information.

h. **Tables and Figures:** Tables and figures should be output into appropriately labelled directories and should be pulled into documents from those locations. Tables and figures should be manipulated using code, not manually. File names for tables and figures should be comprehensible, meaning a longer, more precise name is preferred to a shorter, more ambiguous name.

i. **Confidential Information:** If you have confidential project information that should not be shared with all team members, please tag relevant files and directories using .gitignore so you retain the information but it does not pass into the repo.

Data Management

a. **Data Security:** RAs should adhere to all institutional and ethical guidelines regarding data security and privacy. No data or project documents should be placed on devices that are not owned personally by the RA and to which the RA does not have exclusive access without explicit permission of the PIs. For legal and ethical reasons, data that contains identifying information about research subjects must be handled with particular care. RAs should exercise vigilance in protecting their hardware from theft.

b. **Data Organization:** RAs should organize data in a structured manner, labeling files and folders appropriately, as described under **Directory and File Structure and Naming**.

c. **Raw Data:** Raw data should be exclusively stored in separate directories, appropriately labelled. Generally, raw data is organized by source and/or type (e.g. electoral_data, census_data, VDEM_data, etc.) although occasionally, it will be organized along other lines (e.g. raw data for U.S. state elections is organized by state; raw data for Pakistani provinces by province). Each raw data directory should include a readme that documents sourcing details and other relevant information that will be included in a codebook. Once assembled, raw data should be locked; i.e. it must remain exactly as input, scraped, or downloaded. It should not be mixed with cleaned or transformed data.

d. **Download and Access Documentation and Dates:** Data that is downloaded from the internet should include the full url location and the download date in an accompanying readme.

e. **Clean Data:** Cleaned data should be stored separately from raw data. Cleaned data should (usually) be stored in directories whose structure parallel those of the raw data.

f. **Variable Names:** Variables should be named with comprehensible English-language labels (e.g. "party") and not with incomprehensible labels (e.g. "var1", "var2", etc.). Binary variables should be named according to the meaning assigned to 1; e.g. instead of "gender" (which introduces doubt about the meaning of 0 and 1), name the variable "male" if men are coded 1 and women 0.

g. **Variable Transformations:** In projects that incorporate variable transformations, some useful conventions are:

    a. In case of language translation, a variable might end with *_eng* or *_urdu*.
    b. In case of inverting the order, a variable might end with *_rev*.

h. **Value Labels:** Variables that take different values should be assigned value labels within the dataset; e.g. "sex" = (0,1,2,3), labelled 0 = "does not disclose," 1 = "male," 2 = "female," 3 = "transgender."

i. **Missing Values:** Missing values are generally coded "NA" or some other standard identifier, such as a period or -999. Different kinds of missingness should be precisely indicated; e.g. missing because unavailable from source material is different from missing because the unit does not exist. These differences must be documented to the extent possible.

j. **Data Manipulation:** The standard order of operations for data manipulation is to bring in raw data, clean and store the data, and create new variables. These operations should be performed in distinct and clearly labelled files, operating over distinct and clearly labelled directories. Directories and subdirectories will commonly take the form: 01_data <– 01_raw_data; 02_clean_data; 03_datasets, where 03_datasets are the final processed datasets to be analyzed. The various files performing these operations will be named according to the following type of conventions: 02_cleaning_code <– 00_master.R, 01_build_rawdatasource1.R, 02_build_rawdatasource2.R, 03_build_rawdatasource3.R, 04_create_voteshares.R, 05_create_partyid.R, 06_add_covariates.R, and so forth. Often many variable transformations are required and these should be done in well-labelled files that operate on the dataset

before it is analyzed in order to reduce repetition. The 00_master.R file should perform a full clean run of all processes.

k. **Codebook:** It is usually most efficient to draft a codebook as soon as a dataset is built rather than waiting until we are preparing to post the dataset, when many details will have been forgotten. This also allows team members to quickly ascertain what variables have been created and how. RAs should ask the PI for sample codebooks.

l. **Clean Runs:** A clean run of the dataset assembly process should begin by removing all objects in memory and by removing all previously generated versions of temporary files and output datasets. The code to do this should be at the top of the file. This may be relaxed in very large simulation or assembly processes, where removing all objects would then require too lengthy a process.

m. **File Headers:** Every file that includes code should begin with a standard header that includes: purpose, author, initial date programmed, and any other essential information that permits other team members to operate on the file (e.g. source files, output files, where the code falls in the pipeline, etc.).

n. **Data Backup:** RAs should push all materials to Github frequently so that they are backed up on the server. In general, this means that work should be pushed every time the RA gets up from the computer, or at a minimum at the end of the day.

o. **Git Conflicts:** It is common to experience conflicts when pushing to a Git repo. Simple conflicts can usually be easily resolved by going into the offending file, reviewing the conflicts (demarcated by HEAD), and removing whichever version seems obsolete or wrong. Before doing this, you may want to communicate with other users to investigate the origin of a conflict. This may be especially helpful if you know that someone else is simultaneously editing the same document as you. If you experience more complex conflicts, please contact a PI.

p. **Data and Project Sharing:** All project materials should be considered confidential and RAs should not discuss specifics of any project with anyone other than other team members, unless instructed by the PIs. In discussions with outsiders, RAs may refer to a project in terms of its general topic (e.g. "I work for PI Golden on a project that assembles Pakistani electoral data" or "I work for PI Golden on a project that surveys politicians about corruption") without divulging specifics to others.

Literature and Bibliography

a. **Literature:** RAs will generally not be asked to conduct literature reviews. If an RA locates a relevant article that is not known to the PIs, please communicate this to the PIs with a copy of the article included.

b. **Bibliography:** References are stored in a bibtex file, which is located in the "lit_bib" directory under the project.

c. **Reference Keys:** In written work, citation keys take the format "authorlastnameYY." In cases where a single author has multiple publications in the same year, the format changes to "authorlastnameYYa," "authorlastnameYYb," etc. For clarity, articles should be stored under names that are identical to the reference key that will be used.

Collaboration and Teamwork

a. **Collaboration Tools:** RAs should utilize collaboration tools such as shared document platforms or version control systems as instructed by the PIs. As a rul of thumb, this means working on Github. If you are unfamiliar with Github, you might start by reading the Github manual under Resources at https://www.miriamgolden.com/.

b. **Team Meetings:** RAs will normally be required to attend weekly virtual team meetings to discuss progress, share insights, and collaborate with other team members.

c. **Peer Support:** RAs should actively support and collaborate with their fellow RAs, sharing knowledge and expertise to enhance the overall project outcomes.

Research Ethics

b. **Research Ethics:** RAs should adhere to ethical guidelines in all aspects of their work, including data collection, analysis, and reporting. Any ethical concerns should be communicated to the PI.

c. **Human Subjects Training:** If you will be interacting with human subjects for data collection, you will need to have received a certificate from CITI showing that you have been appropriately trained in the protection of human subjects. This requires you take the CITI course on Social-Bahavioral-Educational (SBE) Foundations (https://about.citiprogram.org/course/social-behavioral-educational-sbe-foundations/). The course costs $129 and is a reimbursable research expense. Please check with the PI about whether you need a CITI certificate and if you should enroll, pay, and seek reimbursement or if another access mechanism will be provided.

d. **Regulatory Compliance:** In addition to approval by an Institutional Review Board at the PI's home institution, the PI is committed to complying with regulations in the country where the research is situated. As part of these processes, your name and other identifying details will be provided to the university where the PI is based and/or the U.S. university of record for the IRB, the funding agency, the contracting institution, and possibly government and/or university authorities in your own country.

e. **Plagiarism:** If you contribute to written work, you should make sure to cite your sources appropriately. Do not copy material off the web or from ChatGPT without first consulting the PI about whether that is appropriate for the specific task.

f. **Fraud:** In order to protect the project from potential accusations of data fraud, it is essential to retain meticulous, highly detailed documentation about data collection procedures.

Professional Development

a. **Training Opportunities:** PIs may provide RAs with access to training materials, webinars, or workshops to enhance their skills and knowledge in relevant areas. RAs who encounter training opportunities that they wish to access should feel free to ask the PI if those could be included as a formal part of the research job.

b. **Skill Development:** RAs should actively seek opportunities to develop their research skills and stay updated with the latest developments in their field.

c. **Professional Conduct:** RAs should maintain professionalism in all interactions and conduct themselves with integrity and respect towards fellow team members, with the public, and with stakeholders. RAs should report any instances of harassment by the public, stakeholders, or other team members to the PI. All reports will be handled confidentially.

Contracts

a. **Logging Hours:** RAs are hired using scarce research funds and should be mindful of that. In order to ensure that you are paid fairly, you will log your hours on Top Tracker (https://www.toptal.com/tracker).

b. **Partial Hours:** If your tasks require frequent short activities, try to bunch them together. You should log work in minimum increments of 15 minutes even if the actual task requires less time.

c. **Payments:** Your contract will be issued by a university or research institute, which is responsible for issuing payment. You should inform the PIs if you are not paid in a timely fashion or if there are any payment irregularities.

d. **Means of Payment:** Generally, you will be paid via bank transfer or an international wire. You must therefore have a bank account in your own name and a tax id.

e. **Tax Withholding:** Depending on the specifics, taxes may be withheld from your payment. All income earned is taxable in your home country.

f. **Contract Length:** Your contract will specify a certain number of hours and specific dates. As you approach the maximum number of hours and/or termination date, you should remind the PIs of this.

g. **Institutional Recourse:** The PIs are committed to fair compensation and to creating an ethical and safe work environment that is fully compliant with all the laws and regulations in the countries where the PIs and their institutions are located as well as in the countries where research is undertaken and RAs are located. If you are concerned that any research task you are asked to undertake may be illegal in your country, please inform the PIs. You may contact the human resources department of the contracting institution if you feel you are not being fairly compensated or if you are asked to undertake activities that are unethical, illegal, or unsafe. If you do so, you will be accorded whistleblower protection.

Performance Evaluation

a. **Performance Assessment:** PIs will evaluate the performance of RAs based on their progress, quality of work, adherence to deadlines, and overall contribution to the project.

b. **Feedback:** PIs will provide regular feedback to RAs, highlighting areas of improvement and acknowledging their strengths. To the extent possible, the PIs will allow RAs to work on tasks that use their strengths and that they enjoy.

c. **Performance Improvement:** RAs should proactively address any performance-related concerns raised by the PIs and take necessary steps to improve their performance.

Conclusions

This SOP serves as a guideline for RAs working remotely on academic projects under the supervision of PIs. Adhering to these procedures will ensure effective communication, efficient project management, and high-quality research outcomes.