Reproducible Social Research

Vernon Gayle

For NCRM Online Resources

https://www.youtube.com/watch?v=kT3SjQ2cCzw&t=513s

Good day my name is Professor Vernon Gayle, I'm professor of sociology and social statistics at the University of Edinburgh and I'm a co-director of the ESRC National Centre for Research Methods.

At the current time the National Center for Research Methods are unable to provide any face-to-face teaching. I hope that you and your families are all healthy during this difficult period.

The video that follows will introduce this topic and provide some background information. There is increasing concern across a wide range of academic disciplines the empirical results cannot be reproduced because of a lack of transparency in the research process. Over the last two decades there has been increasing anxiety that it is impossible to verify the results presented in many research papers. There is growing interest in the need for researchers to provide additional materials alongside traditional publications to enable other researchers to understand evaluate and better build upon previous research work. The purpose of these materials is to provide sufficient information for a third party that is unconnected with the original work to reproduce results without any additional information being provided by the original authors. The focus of this video is social science research that employ statistical techniques to analyze observational data, for example social surveys. Many of the issues associated with undertaking transparent and reproducible data analysis pervade other forms of social science research for example qualitative data analysis. Despite the different nature of the data and the analytical techniques that are used. The problem. Conventional publications for example those in paper-based journals do not provide sufficient space for researchers to deal to detail exactly how they undertook the research therefore the final publication might be best regarded as the tip of the iceberg of the research process. In a similar vein nearly 25 years ago Jon Claerbout stated that in engineering a published paper should be considered as the advertisement for the scholarship. In social science research enterprises such as the analysis of social surveys the researcher begins with a raw i.e. unprocessed dataset. Typically this is a dataset that has been downloaded from the National Archive. In practice a great deal of work usually goes into transforming the raw data prior to the analysis commencing. This data enabling or data wrangling process comprises tasks associated with preparing the data the raw data set and transforming it into an analytical data set suitable for statistical analysis. This data enabling work will include operations such as appropriately coding missing values and recoding values into a suitable format that is required for the specific piece of research. Typically in this phase the data analysts must select appropriate measures and decide how to operationalize them. These choices will be guided both by theoretical considerations and by practical requirements. Verbal selection is not a trivial activity however and genuine research datasets may contain a wide range of variables and can commonly contain different measures of key analytical concepts such as income socioeconomic status and education.

Analytical datasets are the products of the decisions that are made and the actions that are taken during the data enabling phase. These include which cases to include, operationalizing and coding measures and so on. These decisions combine and they are ultimately result in analytical datasets that are too complicated to be reverse engineered from the limited information that is routinely provided in conventional published outputs. Indeed it is usually impossible to reverse-engineer analytical data sets from published results such as the output tables of statistical models. Without access to the analytical data set research findings cannot be genuinely reproduced. The findings from a single social science study should sort of be considered as being definitive. In a similar manner to the natural or physical sciences, social science knowledge is cumulative and empirical research is incremental. Social research findings will almost always be strengthened by additional work that verifies its generality or ubiquity. The extent to which a finding can be reproduced in other research domains is therefore an important barometer. The case for greater transparency. Transparency is a central tenant in reproducible research because without it research cannot feasibly be reproduced. Increasingly transparency in statistically orientated social science research is intrinsically attractive for a number of reasons. Greater transparency will increase the capacity to understand how the research was conducted, help other scholars evaluate the analyses undertaken, aid the detection of errors and inconsistencies facilitate the incremental development of work, contribute to limiting negative research practices, provide extra safeguards against nefarious practices and improve confidence in results within and beyond the academic community. Duplication and replication. Following Nicole Janz we argue that it is fruitful to divide reproducibility into two related concepts.

The first concept is duplication. A study can be duplicated if sufficient information is made available which ensures that consistent results can be produced when the same analytical techniques are applied to the same data and an analysis can be duplicated when a third party that is unconnected with the original analysis can produce identical results. The facility to duplicate work is essential for evaluating empirical research. The second concept is replication a replication study extends the original work with additional measures alternative measures new data or different statistical analysis techniques or any combination of these four components. A sensible first stage in a replication study is the duplication of the original results. Replication studies are important because the methodological extension of the work i.e. additional measures alternative measures new data or alternatives to still cool techniques is what we will test the robustness of the original research. A fundamental aspect of working transparently and enabling reproducibility is data availability providing access to the analytical data set represents a major step forward in enabling the duplication of the original published results. Data should be shared in line with the fair principles which ensure that the standards of findability accessibility interoperability and reusability are met.

In studies where data cannot be shared then it is imperative that researchers clearly identify which raw data set has been used. Some protocols e.g. from the national data archives are emerging. These protocols indicate how researchers should ensure that they are appropriately identifying the data set so that other researchers are able to access exactly the same data resource. An important element of the data citation is that it must include detailed information preferably in the static format identifying the specific version of the data which has been used in order to ensure that they are identical to the data used in the original study. The workflow and code sharing. Following J. Scott Long we use the term workflow to describe the process of planning organizing executing and documenting social science

analyses. The process begins with conceptualizing analyses and includes all of the steps associated with completing the work. The initial steps in the research process are likely to include applying for

ethical approval applying for access the data downloading the raw data and producing the analytical data set. The latter steps are likely to include analyzing the data, presenting results refining results, writing up and then publishing findings. And the final step will be archiving files associated with the project. The central spine of the workflow is the audit trail the audit trail can be thought of as a useful path of breadcrumbs back through the research process. It is implausible for social scientists to expect to transform raw data sets into analytical data sets or to undertake statistical analyses without using a computer. It's commonplace for both data enabling and data analysis to be undertaken using a data analysis software package or a statistical programming language at the current time SPSS Stata and R are the most commonly used statistical data analysis programs in social science.

Software can be operated in different ways but the structure of many raw social science data sets and the intricacy of the variety of tasks associated with transforming the raw data into an analytical data set means that writing out software commands in a programming or syntactical format is a highly effective approach. Similarly the complexity of many analyses means that documented software commands in a programming or syntactical syntactical format rather than using graphical user interfaces gooeys i.e. drop-down menus is far more effective.

The software commands for SPSS are usually written with in syntax files. Within Stata they are written in .do files and in R they'll be written in scripts. The software commands required for transforming the raw data into an analytical data set and the software commands that drive statistical data analyses are referred to as research code. Openly sharing data enabling research code allows third parties who are unconnected with the original research to transform the raw data into an analytical data set. Openly sharing data analysis research code allows researchers who are unconnected with the original to duplicate published results. Therefore making the workflow that produces a published study openly available is fundamental to research transparency and is the foundation of reproducible social science. Documenting the workflow. Sharing research code is essential for understanding all of the steps undertaken to produce the research output. The effectiveness of shared code for reproducing results is completely contingent on how easily it can be understood by a third party that is not connected with the original work. In particular ineffective organization and insufficient documentation are central issues that limit how easily and how well others can comprehend research code and ultimately reproduce work. Social scientists can gain useful insights from the paradigm of literate programming Don Knuth suggested that the traditional attitude to the construction of computer programs should change. Instead of imagining that the main task is to instruct a computer the emphasis should be on explain to human beings what the researcher wanted the computer to do. In essence the codebook is reported alongside an explanation of its logic in a human-readable format e.g. plain English.

At the most fundamental level literate programming involves ensuring that the research code for example the SPSS syntax file the Stata .do file or the R script is adequately supported by comments which explain the particular element of the workflow. For example here is a literate comment about constructing a variable which is along side the code used to undertake that operation. Current resources such as Jupiter notebooks may prove useful in the production of more literate social science workflows. This is because they allow researchers to weave a narrative alongside both statistical data analysis code and results. This is appealing and improves upon plain text code files. Making the workflow public. To

enable transparency and facilitate reproducibility the research code should be shared alongside the output. For example the journal article as an online supplementary material. In practice the format and

location of these materials will depend on the policies and practices of the academic journal. Currently most social science journals do not require researchers to share their data analysis code but something of a quiet revolution is underway. The transparency and openness promotion top guidelines are a set of standards which aim to improve the transplant reporting of research findings in academic journals.

An example of current good practice. Connelly and Gayle 2019 published a paper analyzing existing large-scale social science data sets and they provided an open and transparent workflow. This analysis of social inequalities used existing data from two of the UK's long-running birth cohort studies. The 1958 national child development study and the 1970 British cohort study. The entire workflow that produced the paper was published within a Jupiter notebook. The accompanying notebook included full details of all of the stages of the analysis process.

From the initial stage of data acquisition i.e. downloading the data from the UK data archive and then through the stages of data wrangling, exploratory data analysis, statistical modeling, sensitivity analysis and writing up and reporting results. The intricacies of the analytical process for example decisions and actions for selecting cases the protocol and technique for handling missing data and the construction and coding of measures are fully disclosed. The notebook provides sufficient information for a third party who is unconnected with the original work to reproduce the results without any additional information being provided by the original authors. In line with the fair principles findability accessibility interoperability and reusability the jupiter notebook was also made available on GitHub and the Open Science framework. The capacity to understand exactly how research was conducted will be revolutionized by researchers making their complete workflows publicly available.Having access to the analytical data set all the information required to reconstruct the analytical data set allows scholars to duplicate research. The ability to duplicate research results not only helps others in the field to evaluate analyses but also dramatically aids the detection of errors and inconsistencies. The capacity to duplicate results is foundational for replication studies. Replication studies extend the original work. For example with additional measures alternative measures new data alternative statistical data analysis techniques replication studies offer great potential to improve the capacity to evaluate social science research findings. They also allow us to appropriately locate them within the corpus of existing research evidence. Replication studies are also critical in establishing the extent to which findings constitute empirical regularities.

Increased transparency has the potential to limit negative research practices. A notable example is publication bias the term used to describe the phenomenon of a distortion in reporting knowledge. One invidious form of publication bias is the greater likelihood of statistically significant results being published in academic journals as publishers may be reticent to publish non statistically significant results. An interconnected issue is researchers selective reporting of non significant empirical findings which is often referred to as the file drawer problem. This terminology conveys the notion that undesirable results often go no further than the researchers file drawers and this in turn leads to bias in published research. Open access to the complete workflow makes a contribution to limiting publication bias this is because it provides opportunities for the wider research community to have access to the results that hitherto would have been inaccessible because they were unpublished. There are a range of subjective decisions that researchers must make to motivate any study. For example these actions might

include theoretical decisions relating to the research question, pragmatic decisions on choosing data and practical efficiencies associated with developing the analytical data set. Researchers will also make theoretical and practical judgments in order to select measures. They'll also be required to make

decisions on which data analytical methods to employ. Statistical methods are not mechanical and further decisions will be required on the technical issues such as model choice. Ultimately then decisions will be made about which aspects of the analyses are emphasized in reporting. This spectrum of subjective decisions is sometimes described as researcher degrees of freedom. The positive aspect of the degree of freedom afforded to researchers is that it enables the suitable formulation of empirical work. The negative aspect of this freedom is that it opens up opportunities for pernicious research practices such as p-hacking and harking hypothesizing after the result is known. An obvious practice which provides extra safeguards against pernicious research practices is pre-registering a pre-analysis plan. This requires researchers to submit a document describing the analysis they plan to carry out which forms a public record.

The use of pre analysis plans has grain traction in some areas for example randomized control trials RCTs it is more difficult however to provide credible pre analysis plans for the analysis of observational data such as social surveys this is during part to the form of the raw data sets and the data enabling work that is almost always routinely required prior to data analysis. Some progress has however been made in areas such as economics. In the absence of pre-registering pre analysis plans open access to the complete workflow especially when it is appropriately documented is a positive development in providing safeguards as well as providing protection against pernicious research practices improved transparency provides extra safeguards against the various practices such as data fraud. In conclusion increasing transparency and facilitating the duplication of results and the incremental development of empirical research through replication is likely to improve confidence in social science results both within and beyond the academic community.

At the current time the UK is in lockdown due to the coronavirus emergency the National Center for Research Methods would have been providing face-to-face training but at the current time this is not possible. I would like to thank my colleagues Dr. Roxanne Connelly and Dr. Christopher Playford who I had intended to deliver a workshop with on the topic of undertaking transparent and reproducible data analysis but the current time this isn't possible. The video that you've just watched will be followed up with more information that you'll be available from the NCRM website.