Undertaking Transparent and Reproducible Data Analysis

There is increasing concern across a wide range of academic disciplines that empirical results cannot be reproduced because of a lack of transparency in the research process (Baker, 2016). Over the last decade there has been increasing anxiety that it is impossible to verify the results presented in many research papers (Christensen, Freese, & Miguel, 2019).

There is a growing interest in the need for researchers to provide additional materials alongside traditional publications to enable other researchers to understand, evaluate and 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 entry is social science research that employs statistical techniques to analyse observational data (e.g. social surveys). Many of the issues associated with undertaking transparent and reproducible data analysis pervade other forms of social science research (e.g. 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 detail exactly how they undertook the research. Therefore, the final publication might best be regarded as the tip of the iceberg of the research process. In a similar vein nearly twenty five years ago Claerbout (1994) stated that in engineering a published paper should be considered as an advertisement of 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 a national archive. In practice a great deal of work usually goes into transforming the ‘raw’ data prior to analyses commencing. This ‘data enabling’ or ‘data wrangling’ process comprises tasks associated with preparing the ‘raw’ dataset and transforming it into an ‘analytical’ dataset suitable for statistical analysis (Gayle & Lambert, 2017; Long, 2009).

This data enabling work will include operations such as appropriately coding missing values and re-coding variables into a suitable format that is required for the specific piece of research. Typically, in this phase the data analyst must select appropriate measures and decide how to operationalize them. These choices will be guided both by theoretical considerations and practical requirements. Variable 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, socio-economic status, and education (see Connelly, Gayle, & Lambert, 2016).

Analytical datasets are the products of the decisions that are made, and the actions that are taken in choosing which cases to include, and operationalizing and coding measures in the data enabling phase. These decisions combine and they ultimately result in analytical datasets that are too complex to be 'reversed engineered' from the limited information that is routinely provided in conventional published outputs. Indeed, it is usually impossible to reverse engineer analytical datasets from published results such as the output tables of statistical models. Without access to the analytical dataset research findings cannot be genuinely reproduced.

The findings from a single social science study should seldom 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 tenet 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

1. Increase the capacity to understand how the research was conducted
2. Help other scholars evaluate the analyses undertaken
3. Aid the detection of errors and inconsistencies
4. Facilitate the incremental development of work
5. Contribute to limiting negative research practices
6. Provide extra safeguards against nefarious practices
7. Improve confidence in results within and beyond the academic community

Duplication and Replication

Following Janz (2016) 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. 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 (King, 2003).

The second concept is ‘replication’. A replication study extends the original work with

1. additional measures
2. alternative measures
3. new data
4. different statistical analytical 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 alternative statistical techniques) is what will test the robustness of the original results.

Data Sharing and Citing Data

A fundamental aspect of working transparently and enabling reproducibly is data availability. Providing access to the analytical dataset 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 (Wilkinson et al., 2016). Established data archives such as the UK Data Archive and the Inter-University Consortium for Political and Social Research Data Archive are extremely well placed to promote data sharing, and to assist in ensuring data are professionally curated and made easily accessible to other researchers.

In many instances social science data are only available via an end user license (e.g. from a national data archive) and cannot be shared publically. Examples of datasets with end user licenses and restrictions on data sharing include the UK Household Longitudinal Study, the German Socio-Economic Panel and the Household, Income and Labour Dynamics in Australia survey. Many other social surveys and national Censuses have similar restrictions. Even tighter controls are placed on some administrative data resources, especially those containing personal and sensitive information, or measures that might identify individuals or invade their privacy.

In studies where data cannot be shared then it is imperative that researchers clearly identify which ‘raw’ dataset has been used. Some protocols (e.g. from national data archives) are emerging. These protocols indicate how researchers should ensure that they appropriately identify the dataset, 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 a static format) identifying the specific version of 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 Long (2009) we use the term ‘workflow’ to describe the process of planning, organising, executing and documenting social science analyses. The process begins with conceptualising 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 to the data, downloading the ‘raw data’, producing the ‘analytical dataset’. The later steps are likely to include analysing data, presenting results, refining results, writing up and then publishing findings, and finally 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 datasets into analytical datasets or to undertake statistical analyses without using a computer. It is commonplace for both ‘data enabling’ and data analyses 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 used by social scientists.

Software can be operated in different ways but the structure of many raw social science datasets, and the intricacy of the variety of tasks associated with transforming the raw data into an analytical dataset, 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 documenting software commands in a programming or syntactical format rather than using graphical user interfaces (e.g. drop down menus) is far more effective. The software commands for SPSS are usually written within syntax files, within .do files in Stata, and within scripts in the R environment.

The software commands required for transforming the ‘raw data’ into an ‘analytical dataset’ 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 ‘raw data’ into ‘analytical datasets’. Openly sharing data analysis research code allow 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 foundation of reproducible social science research.

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 was not connected with the original work. In particular, ineffective organisation 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 also gain useful insights from the paradigm of literate programming (see Knuth, 1992). Knuth (1984) 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 explaining to human beings what the researcher wanted the computer to do. In essence, the computer code 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 (e.g. the SPSS syntax file, the Stata .do file or the R scripts) is adequately supported by comments which explain the particular element of the workflow. Examples in the data enabling phase might include the motivation behind the decision to select certain cases for analysis, or the strategy for recoding missing values. Examples in the data analysis phase might include the motivation for choosing a statistical technique, the strategy for estimating a suite of statistical models, or how the most appropriate model was chosen.

Current resources such as Jupyter Notebooks, R Markdown or Stata Dynamic Documents 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 in a potentially more appealing fashion than more common plain text code files.

Making the Workflow Public

To enable transparency and facilitate reproducibility the research code should be shared alongside the output (e.g. 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 or outlet. Currently most social science journals do not require researchers to share their data analysis code but something of a quiet revolution might be underway. The Transparency and Openness Promotion (TOP) guidelines are a set of standards which aim to improve the transparent reporting of research findings in academic journals (Nosek et al., 2016). The TOP guidelines consist of eight standards including data citation standards, data transparency, analytic methods (code) transparency, research materials transparency, and design and analysis transparency. Many journals have signed up to the TOP guidelines, and therefore have signaled their commitment to adopt these good practices in transparent and reproducible research.

If a journal is unable to host research code or additional materials alongside a published manuscript, there are other options available to researchers. A simple pragmatic solution is for researchers to share research code on their own, or a research team’s, website. The downside of this approach is that these resources will seldom meet the FAIR principles. A practical alternative is the use of GitHub. This is a software development platform which allows users to create repositories to upload code. Integral to GitHub is the concept of version control, which allows teams to efficiently collaborate when developing, and editing code. The functionality of GitHub lends itself well to developing public repositories of social science workflows. The Open Science Framework (OSF) is a specialist platform where research code can be shared alongside further project related materials such as conference presentations and preprints

(Foster & Deardorff, 2017). It is currently in infancy, but the OSF platform shows promising signs that it could emerge as a dominant eco-system for transparent and reproducible social science.

An Example of Current Good Practice

Connelly and Gayle (2019) published a paper analysing existing large-scale social science datasets, 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 Jupyter Notebook.

The accompanying notebook included full details of all of the stages of the analysis process, from the initial step of data acquisition (i.e. downloading the data from the UK Data Archive), and then through the stages of data enabling, exploratory data analysis, statistical modelling, sensitivity analyses and reporting results. The intricacies of the analytical process, for example decisions and actions for selecting cases, the protocol and techniques for handling missing data and the construction and coding of measures are all fully disclosed. The notebook provides sufficient information for a third party, who is unconnected with the original work, to reproduce results without any additional information being provided by the original authors. The Jupyter Notebook is available as online supplementary material on the journal's webpage. In line with the FAIR principles (findability, accessibility, interoperability and reusability) the Jupyter Notebook was also made available on GitHub and the Open Science Framework.

Conclusion

The capacity to understand exactly how research was conducted will be revolutionised by researchers making their complete workflows publically available. Having access to the analytical dataset, or the information required to reconstruct the analytical dataset 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 or alternative statistical data analysis techniques. Replication studies offer great potential to improve the capacity to evaluate social science findings, and 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 reported 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-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 researchers' file drawers, and this in turn leads to biases in published research (see Franco, Malhotra, & Simonovits, 2014). 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 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 include theoretical decisions relating to the research question, pragmatic decisions on choosing data, and practical exigencies associated with developing the analytical dataset. Researchers will also make theoretical and practical judgements in order to select measures. They will 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 technical issues such as model choice. Ultimately, decisions will be made about which aspects of the analyses are emphasised in reporting. This spectrum of subjective decisions is sometimes described as 'researcher degrees of freedom' (Silberzahn et al., 2018). The positive aspect of the degrees of freedom afforded to researchers it that it enables the suitable formulation of empirical work. The negative aspect of this freedom is that it opens opportunities for pernicious research practices such as p-hacking and HARKing (see Kerr, 1998; Rubin, 2017).

An obvious practice which provides extra safeguards against pernicious research practices is pre-registering a pre-analysis plan (Nosek & Lakens, 2014). This requires researchers to submit a document describing the analyses they plan to carry out, which forms a public record. The use of pre-analysis plans has gained traction in some areas for example Randomised 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 due in part to the form of 'raw' datasets and the enabling work that is almost always routinely required prior to data analysis. Some progress has however been achieved in areas such as economics however (see Burlig, 2018). 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 nefarious 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.

Further Readings

Chambers, C. (2019). The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice. Princeton University Press.

Christensen, G., Freese, J., & Miguel, E. (2019). Transparent and Reproducible Social Science Research: How to Do Open Science. University of California Press.

Figueiredo Filho, D., Lins, R., Domingos, A., Janz, N., & Silva, L. (2019). Seven Reasons Why: A User's Guide to Transparency and Reproducibility. Brazilian Political Science Review, 13(2), 1-37.

Gandrud, C. (2016). Reproducible research with R and R studio. Boca Raton, Florida: CRC Press.

Ioannidis, J. P. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124.

King, G. (1995). Replication, Replication. Political Science & Politics, 28(3), 444-452.

Long, J. S. (2009). The workflow of data analysis using Stata (p. 379). College Station, Texas: Stata Press.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., … Van der Laan, M. (2014). Promoting transparency in social science research. Science, 343(6166), 30-31.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. Science, 348(6242), 1422-1425.

References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452-454. doi: 10.1038/533452a

Burlig, F. (2018). Improving transparency in observational social science research: A pre-analysis plan approach. *Economics Letters, 168*, 56-60.

Christensen, G., Freese, J., & Miguel, E. (2019). *Transparent and reproducible social science research: How to do open science*. Oakland, California: University of California Press.

Claerbout, J. (1994). *Seventeen years of super computing and other problems in seismology.* Paper presented at the National Research Council meeting on High Performance Computing in Seismology.

Connelly, R., & Gayle, V. (2019). An investigation of social class inequalities in general cognitive ability in two British birth cohorts. *The British journal of sociology, 70*(1), 90-108.

Connelly, R., Gayle, V., & Lambert, P. S. (2016). Modelling key variables in social science research: Introduction to the special section. *Methodological Innovations, 9*. doi:10.1177/2059799116637782

Foster, E. D., & Deardorff, A. (2017). Open science framework (OSF). *Journal of the Medical Library Association: JMLA, 105*(2), 203.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505.

Gayle, V. J., & Lambert, P. S. (2017). *The workflow: A practical guide to producing accurate, efficient, transparent and reproducible social survey data analysis*. Retrieved from Southampton:

Janz, N. (2016). Bringing the gold standard into the classroom: replication in university teaching. *International Studies Perspectives*, *17*(4), 392-407.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196-217.

King, G. (2003). The future of replication. *International Studies Perspectives*, *4*(1), 100-105.

Knuth, D. E. (1984). Literate programming. *The Computer Journal, 27*(2), 97-111.

Knuth, D. E. (1992). Literate programming. *CSLI Lecture Notes, Stanford, CA: Center for the Study of Language and Information (CSLI), 1992*.

Long, J. S. (2009). *The workflow of data analysis using Stata*. College Station, TX: Stata Press

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S., Breckler, S., . . . Christensen, G. (2016). Transparency and openness promotion (TOP) guidelines. doi:https://osf.io/vj54c/

Nosek, B. A., & Lakens, D. (2014). Registered Reports: A method to increase the credibility of published reports. *45*(3), 137-141.

Rubin, M. (2017). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology, 21*(4), 308-320.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., . . . Bonnier, E. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337-356.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Nature, 3*.