The Workflow

National Centre for Research Methods Working Paper

1/17

The Workflow: A Practical Guide to Producing Accurate, Efficient, Transparent and Reproducible Social Survey Data Analysis

Vernon Gayle, Paul Lambert

NCRM National Centre for Research Methods

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

The Workflow: A Practical Guide to Producing Accurate, Efficient, Transparent and Reproducible Social Survey Data Analysis

Professor Vernon Gayle, University of Edinburgh

Professor Paul Lambert, University of Stirling

Abstract

This working paper is concerned with the very practical aspects of organising and undertaking analyses of social surveys. The focus is directed towards the many tasks involved in making good analytical use of social survey data, and it provides practicable advice on the process of how to plan, organise, compute and document analyses of social surveys. Advice on data analysis software is provided and on how to work using syntax files to undertake data analysis. We also discuss the critical role of documentation and aspects of the workflow that are frequently overlooked (e.g. organising directory structures and file naming protocols). The overall goal of this working paper is to provide researchers with information that enables them to efficiently undertake accurate, transparent and reproducible research using social surveys.

Introduction

Social surveys provide rich sources of empirical data but analysing these data often proves to be complicated. In this working paper we highlight the many tasks involved in making good analytical use of social survey data and provide practicable advice. The focus is the process of how to plan, organise, compute and document analyses of social surveys. Following Long (2009) we use the term 'workflow' to describe the series of activities necessary to complete statistically orientated data analysis tasks.

We begin with a simple thought experiment.

> *Have you ever lost a file?*
>
> *Have you ever wondered if you have deleted a file?*
>
> *Have you and a colleague ever unintentionally been working on different versions of a file?*
>
> *Have you ever had difficulty identifying which file is the correct one (e.g. chapter1_2016.dat or chap1_2016.dat)?*
>
> *Have you ever struggled to re-run a statistical analysis?*

If the answer to any of these questions is yes, then your workflow could be improved.

The workflow includes planning, organising, executing and documenting analyses. The process begins with conceptualising analyses and includes all of the steps associated with completing the work. Central to the workflow for survey data analysis is the operation of tasks using statistical software. Software can be operated in different ways but the complexity of most social surveys means that nearly all highly skilled researchers write out software commands using a syntactical or programming format, and we will elaborate upon this issue.

The initial steps in the research process are likely to include applying for ethical approval, applying for access to the data, downloading data, cleaning data, backing up data, and enabling data for analysis. The later steps are likely to include analysing data, presenting results, refining results,

writing up and then publishing findings, and finally archiving files of data and results. The cardinal message is that the workflow should be planned and carefully orchestrated. The workflow should never be *ad hoc* or piecemeal, or developed as a reaction to problems and mistakes. The good news is that the workflow can be improved with only a modest amount of extra effort. The modest outlay of effort will then pay huge dividends. The other piece of good news is that less experienced data analysts have the advantage that they can start from scratch and get into the good habit of developing a systematic workflow for all of their projects. In our view, Long (2009) is the definitive text on organising your social science data analysis workflow. It is clearly written and insightful and we recommend that any data analyst who has not read the text would benefit from doing so regardless of their age or career stage.

The Four Pillars of Wisdom

The central goal of the workflow is to ensure that the provenance of every result can be easily determined. It is useful to think of four priorities that inform successful social survey data analyses. They are accuracy, programming efficiency, transparency and reproducibility (see Long, 2009). We sometimes refer to these as the four pillars of wisdom.

Accuracy relates to minimising information loss and errors in both analyses and outputs.

Programming Efficiency relates to maximising the features offered by software, and when possible automating (or semi-automating) actions. Social science data analysts should be mindful of Drukker's dictum - *never type anything that you can obtain from a saved result* (Long, 2009).

Transparency is central to good social science data analysis practices. When work is appropriately transparent questions of the 'who, what, where, when and why' variety are easily answered. Transparency is especially important when collaborating or working within teams.

Reproducibility is a cornerstone of good social science. In essence, work can be reproduced when the same results are produced every time, despite who undertakes the analysis, or where it is

undertaken. The ability to reproduce results is critical for editing work and essential for activities such as rewriting a thesis or working towards resubmitting a journal article.

It is easy to under-estimate the importance of having a workflow that facilitates reproducible results. It is difficult to over-emphasise the dividends that a planned and organised workflow will return to researchers.

A diagrammatic overview of the stages of the workflow in social survey data analysis is provided in Figure 1. Central to the workflow is the concept of having an 'audit trail'. The 'audit trail' is a sequential account of the activities undertaken in the research process. It is critically important because within the statistical analysis of social surveys minor decisions have major consequences. For example, it is often difficult to remember which cases have been included in an analysis, which version of a variable was used, or how a variable was recoded. Any of these seemingly minor decisions can potentially have major consequences within an analysis. Keeping track of even the most seemingly minor actions in the workflow is therefore important as it facilitates transparency. Transparency makes contributions to accuracy, effectiveness and reproducibility, and ultimately to the overall success of the research project.

A good workflow entails consistent practices, and the repetition of clear practices should lead to greater efficiency. The workflow must be easy to use and the protocols must not be too difficult or cumbersome because this will make them unattractive to adhere to. The workflow should be compatible with how the researcher (or the research team) prefer to work, whilst still being systematic and consistent.

We often joke that a platinum standard workflow would mean that if a researcher on a project was incapacitated in a freak accident then the Principal Investigator would be able to appoint a replacement who could understand all of the work that had been undertaken and continue to move it forward. This is a lofty ambition but it is worth aspiring to. The gold standard would be that every action, from the most minor variable recode through to research outputs being produced and finally materials being archived, was clearly documented and was subsequently traceable. In reality most

good survey analysts aim for the gold standard, but probably achieve something that resembles a lower standard, which we will call a bronze standard. The bronze standard of workflow could be improved upon, but it is suitably functional. Even a bronze standard will be much more beneficial than an *ad hoc* or unplanned workflow. A bronze standard workflow will not be achieved accidentally however, and will be the result of researchers devoting some thought to the workflow and being sufficiently disciplined to follow a systematic set of procedures throughout the lifecycle of the project.

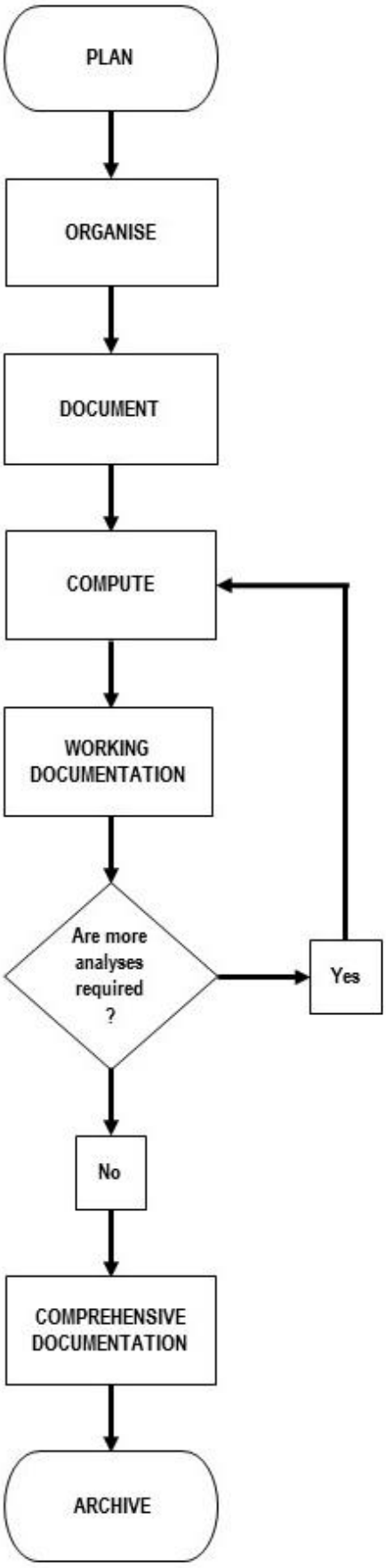
Data Analysis Software

It is unrealistic to undertake anything more than extremely basic analyses of survey data without using data analysis software. There are several statistical data analysis packages available to social science researchers. The three that are currently most prevalent are SPSS, Stata[1] and R. The software package SPSS (the original title was Statistical Package for the Social Sciences) is long established (see Nie, 1983). SPSS has traditionally been popular in areas such as sociology, social policy and politics and to a lesser extent psychology. The software package Stata was initially developed in the 1980s (see Pinzon, 2015). More recently R has emerged, and it is popular with statisticians and methodologists (see Ihaka, 1998).

Conversations on which statistical data analysis package social science survey data analysts should use sometimes degenerate into a volley of retorts where an advocate of one software package suggests that another software package is unable to perform certain operations. The conversation on the merits and limitations of software resemble the Christmas pantomimes that we remember from childhood, with one side chanting “oh no it can’t”, and the other side chanting “oh yes it can”.

[1] Stata is written as ‘Stata’ and not ‘STATA’. We must confess that in the past we have made this error (see http://www.restore.ac.uk/Longitudinal/ accessed 27.06.16).

Figure 1 A High Level Overview of the Workflow in Social Survey Data Analysis



Survey data analysts are free to use whatever software they prefer. At the current time we recommend using Stata for all analyses of social survey data. We are social scientists and are very actively engaged in the analysis of large-scale and complex social surveys. We have also been engaged in teaching statistics and data analysis to undergraduate and postgraduate students for the last two decades. In addition we have delivered a sizeable number of training and capacity building activities, such as training workshops for early career researchers and researchers working in non-academic settings. Taken together these experiences have informed our choice of Stata for the analysis of social surveys. Stata stands out as a sensible choice because it is a popular commercial package with a wide community of social science users who exploit it in a manner that accommodates effective practices in social research which include supporting easy documentation, a wide range of analytical capabilities, and ongoing developmental activities (Lambert et al., 2015). We have found that overall it is the single most effective and efficient tool for undertaking and successfully completing survey data analysis. The tasks associated with data enabling, exploratory data analyses, building statistical models and organising presentation-ready and publication-ready outputs (by which we mean high-quality graphs and tables of modelling results), can all be undertaken using Stata in a single uniformed environment.

A leading US social scientist and survey data analyst Donald J. Treiman makes the following comments on Stata,

> *'for many years, SPSS was the package of choice among sociologists probably because it was written by and for sociologists…Although it is still widely used by social scientists in Europe and Asia it has lost its market share in leading U.S. research universities… [Stata] has gotten better and better over time, so that by now it can happily be used as a general-purpose package. Stata is powerful and fast which makes it viable to carry out analysis on a PC… Overall, Stata is a very good choice for our kind of work' (Treiman, 2009 p.66).*

The UK Data Service[2] provides most large-scale social surveys in SPSS and Stata format. SPSS is far more restricted in the range of statistical models that it can estimate. SPSS currently has few options for estimating models that are suited to longitudinal data. Stata is able to offer more comprehensive facilities to analyse survey datasets with complex designs and selection strategies. This is a clear benefit for social scientists working with contemporary datasets such as the UK Household Longitudinal Study (Understanding Society)[3] and the UK Millennium Cohort Study[4].

The freeware R provides a viable alternative with a substantial volume of analytical options and considerable programming flexibility (Long, 2011). Over the years, we have often found ourselves being told by people who are not routinely engaged in survey data enabling or analysis activities, and don't themselves regularly use R, that R is the superior statistical package. We are not sure why researchers with less experience of analysing surveys are confident in these assertions, but we can only imagine that there is something beguiling about R that convinces them it is more suitable than packages like Stata. The advanced programming and statistical skills which are necessary to effectively exploit R through textual programming seem unlikely to lead to its universal adaptation amongst the wide ranging user-communities within the social sciences (Lambert et al., 2015). A further limitation is that R is currently not well suited to the analysis of large-scale social surveys. For example when using R it is difficult to effectively combine the numeric codes for variables along with both their value and variable labels. This means that users are not able to effectively exploit the meta-information on measures that is helpful for routine survey data analysis tasks. As well as providing data in SPSS and Stata format the UK Data Service provides data in a more package agnostic tab-delimited format. Some R users advocate importing data in this format. In our experience this format can prove challenging to work with especially when matching and merging files and undertaking data analysis enabling tasks. R users can sometimes encounter memory capacity problems when processing large-scale data files. A further limitation of R is that there is a

[2] See https://www.ukdataservice.ac.uk/ accessed 10.11.16.
[3] See https://www.understandingsociety.ac.uk/ accessed 10.11.16.
[4] See http://www.closer.ac.uk/study/millennium-cohort-study/ accessed 10.11.16.

lack of clear and concise help files which contain applied examples that relate to the analysis of social science datasets.

There are many resources aimed at providing Stata training. We recommend Kohler and Kreuter (2012) because this is a well organised and informative text for anyone who needs to learn Stata. A number of colleagues reported positive experiences using Pevalin and Robson (2009). We also recommend Treiman (2009) which is an excellent text covering a range of more advanced topics and the challenges and issues that researchers are likely to encounter. We enthusiastically promote the Institute for Digital Research and Education (IDRE), University of California Los Angeles, web resources[5] and draw attention to their Stata pages[6]. We also direct readers wishing to improve their Stata skills to IDRE's very helpful annotated outputs pages which contain example programs and output with footnotes explaining the meaning of information, measures and statistical tests reported in outputs[7]. These pages are designed to help researchers read the output more effectively and to enable them to have a more comprehensive understanding of results delivered by software packages. These resources have been extremely useful over the years and have had very positive feedback from our students.

Data Enabling

Large-scale social surveys are almost never delivered to the data analyst in a form that renders them immediately ready for comprehensive analyses. This is typically because there are several different data files supplied. There are often a number of different measures (e.g. different socioeconomic classifications) in the dataset that could be used, and often metadata are also supplied with the dataset. A weakness of some analyses is that they do not take full advantage of the richer data resources available in the survey because they are restricted to exploiting only the most readily available information.

[5] See https://idre.ucla.edu/ accessed 06.04.16.
[6] See http://www.ats.ucla.edu/stat/stata/ accessed 06.04.16.
[7] See http://www.ats.ucla.edu/stat/AnnotatedOutput/ accessed 06.04.16.

Social surveys almost always require some preparation before analyses can be undertaken. We use the term 'data enabling' to describe this phase of the workflow. As a minimum this might only be recoding some variables, or coding some missing values. In reality, even well curated social science datasets require some work to be undertaken to enable data analyses.

Longitudinal survey datasets generally require more data enabling than cross-sectional datasets before analyses can be undertaken. In the case of the large-scale household panel surveys, data files will be released on an annual basis (e.g. each wave or year) and there will usually be a number of different files associated with individuals and households. Even for very simple analyses the data analyst will have to link up a series of files containing information on the respondent for each wave of the survey. In addition the data analysts might want to match household level information at each wave to data on the individual respondent. The data analyst might also require information on the individual's spouse or others sharing the household. The collation of information from several files and appropriately identifying individuals and households, matching them up and then merging relevant information is a frequent and fundamental aspect of organising longitudinal data in the data enabling phase of the workflow.

Documentation is fundamental to achieving a workflow that is accurate, efficient and transparent, and which facilitates reproduction. In his authoritative book on the workflow Long (2009) posits Long's law which states that it is easier to document today than it is tomorrow. Long further states that there are two corollaries. First, nobody likes to write documentation. Second, nobody ever regrets writing documentation. We suggest that in the history of social survey data analysis no one has ever said "this work is too well documented".

We encourage social survey data analysts to make copious written notes and comments because these will be the backbone of documentation. There is a long history in the natural sciences of researchers continuously making notes that contribute to high quality documentation. Nobel Prize winner Linus Pauling used bound notebooks to keep track of the details of his research, and the forty

six notebooks spanning a period from 1922 until 1994 are available online[8]. Professor Pauling's notebooks include calculations, experimental data, scientific conclusions, ideas for further research and numerous autobiographical reflections (e.g. Notebook 24 p.151 contains an entry detailing his golden wedding anniversary).

Over the course of our careers we have learned an important lesson. The process of data enabling usually takes five times as long as initially anticipated. Tasks that researchers initially think will take an hour usually take five hours, and similarly tasks that are estimated to take a day commonly take five days. This is an important lesson to internalise, and our advice is to try to ensure that you build in adequate time for enabling your data and try to remember that data enabling tasks are usually very time consuming.

Our graduate students refer to the idea of 'five times' as the 'Gayle-Lambert constant'. We prefer to think of it as a guideline, because there will undoubtedly be social scientists with better programing skills than ours, who can work faster and program more efficiently. Unless you are certain that your programming skills are at an advanced level then the 'fives times' guideline is worth being mindful of. We recommend that survey data analysts try to remember that data enabling tasks are usually very time consuming, and in all cases we strongly advise you to ensure that you build in adequate time for enabling your data prior to analysis.

Protocols for Directories

There are a number of prosaic activities that assist in maintaining a good workflow. We recommend that survey data analysts adopt a consistent file directory structure. The directory structure should be transparent (or at least diaphanous), and therefore it should be clear where files should be stored. The advice contained in the old adage 'a place for everything, and everything in its place', should be observed.

[8] See scarc.library.oregonstate.edu accessed 06.04.16.

Because social survey datasets usually comprise many different data files, and their analysis generates more files and outputs, a key feature of workflow planning involves decisions about the location of files. Existing statistical data analysis software requires exact file locations to be specified. Traditionally it was assumed that file locations were on local machines and it can sometimes be difficult to combine standard software with emerging facilities for file storage such as cloud-based systems. Our current advice is that researchers should think carefully about developing a strategy that both suits their workflow and fits within the constraints of their university's information technology systems and cloud-based facilities.

Figure 2 A Rudimentary Directory Structure

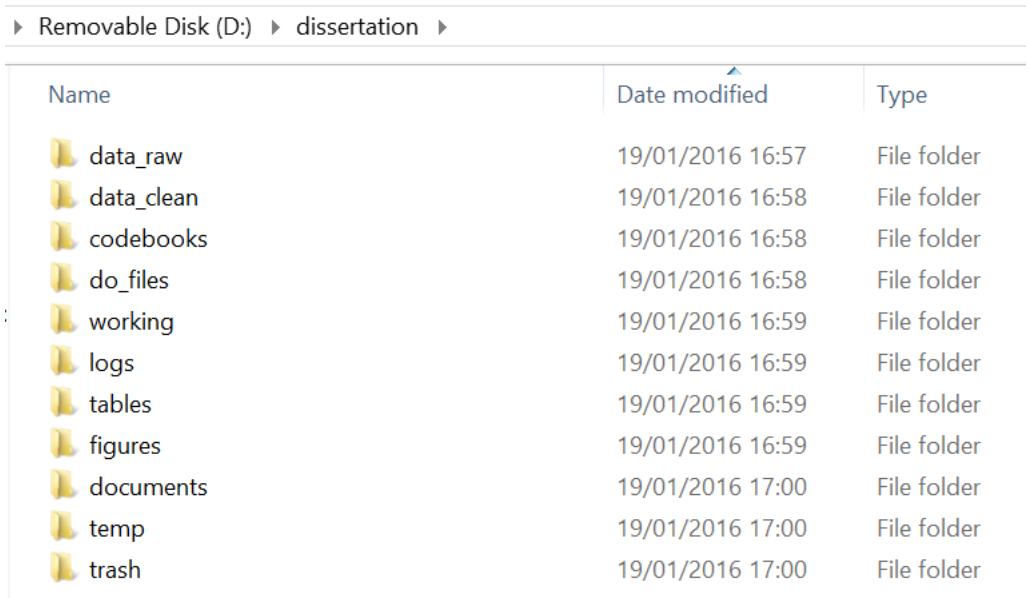


Figure 2 shows a rudimentary directory structure which would be suitable for both a student undertaking a dissertation and a more comprehensive research project. In this directory structure the original or 'raw' data files are kept separately from data files that have been amended and enabled for analysis and kept in a separate location ('data_clean'). We recommend producing

codebooks for datasets which are summaries of datasets and include information such as variable names and value labels, and some summary information. Codebooks should be stored in a separate folder ('codebooks').

We suggest that syntax or command files, i.e. files that 'command' software to undertake operations and computation, are kept separately from data files. Syntax files in Stata are known as 'do' files and have a *.do* extension. In Figure 2 we have a folder 'do_files' where our syntax files are stored separately from other materials.

A key aim of an organised workflow is to avoid messiness which easily leads to confusion which ultimately wastes time. During the analytical phase of a project it is advisable to have a working space, this is very much like a workbench in a shed. In Figure 2 this is the folder 'working'. This is an area that will not be cluttered up by data and syntax files, and where you are unlikely to mess up important files and information. Log files form a useful part of the workflow as they provide full transcripts of data analysis sessions. Log files provide an indispensable means of keeping track of actions within the research process. Once again we advocate that log files have their own storage location, and in Figure 2 this is the folder 'logs'.

In the pursuit of programming efficiency, it is good practice to use the features offered by the data analysis software to automatically (or semi-automatically) produce results. Efficiency is maximised when the software is used to produce publication-ready outputs such as statistical modelling results and graphical outputs. We recommend that results should be appropriately organised and stored in specific locations, and in Figure 2 these are the folders 'figures' and 'tables'.

Keeping track of written documents, such as draft chapters and drafts of journal articles, is as much a part of a good workflow as keeping track of data and syntax files. It is sensible to store written documents in a different location to data and syntax files. We recommend setting aside a specific location for trash, which stores junk files rather than permanently deleting them. We further recommend a location where temporary files can be stored, for example where impermanent files that are often produced during operations such as merging datasets can be located.

The directory structure that has been outlined above is by no means definitive. This is just one possible set of arrangements. As a survey data analyst you should develop a directory structure that reflects your favoured working practices. The critical feature of whatever protocol you develop is that it is simple enough to be useable and easy to adhere to at all times. In collaborative research endeavours and in research teams, what should be the simple task of establishing and sticking to a directory structure frequently proves to be challenging. In collaborative projects, having a well organised set of arrangements that team members can follow, greatly increases the chances of work being successfully completed.

File Naming Protocol

Together with a structured set of directories, it is also valuable to adopt a consistent file naming protocol. A golden rule is never ever name a file *final*! It will seldom be the final version and very soon you will have another file called '*final1'* or '*finalnew'*. Unwittingly, you will have left a banana skin ripe for tripping you up sometime in the future. We recommend adopting a protocol that provides clear information on what type of file it is and what the file contains, which is easily locatable in the audit trail.

Here is an example of a simple but effective file naming protocol.

File Name = title_date_depositor's initials_version_type

For example consider a file that is named *bhpsaindresp_20140506_vg_v1.dta* .

The first part of the file name should be a meaningful title. In this example it is a file from the British Household Panel Survey (BHPS). It is a data file from wave a. The file contains 'individual responses' from the survey. Therefore the title is *bhpsaindresp*, and it has the benefit of being reasonably human-readable.

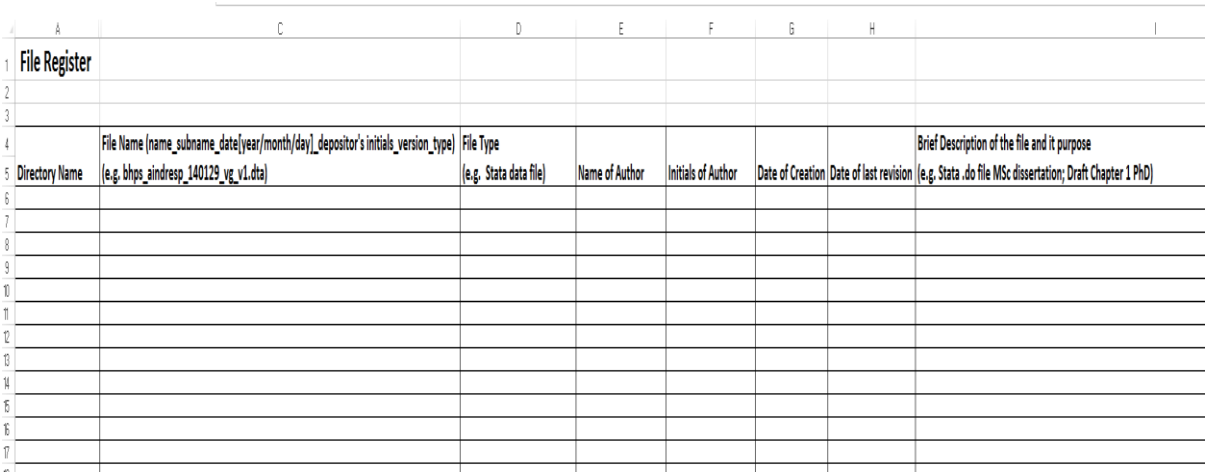
The file was created on 6th May 2014. We suggest 20140506 as a date format. This is because many files will be produced in the same year (e.g. 2014) and 20140506 can be searched far more easily than 06052014. The file was created by Vernon Gayle (vg). We suggest that a single author naming

protocol is established and maintained throughout a project especially when multiple researchers are making contributions.

This is the first version of the file (v1). The need for a systematic way to organize and control revisions has probably existed since writing began. Keeping track of revisions and updates is an important aspect of the workflow. It is especially critical when multiple researchers have access to files. Personal computers now have vast storage capacities and extra storage is now relatively cheap. Therefore there are few barriers to regularly saving updated and revised files. Having a clear and consistent numbering system is critical to locating files correctly within the audit trail. The final part of the file name will be the file type. In this example it is a data file for use in the data analysis software package Stata and therefore has the *.dta* file extension (an SPSS syntax file would have a *.sps* file extension).

In many situations it can be helpful to have a central register of all of the files associated with a project. A simple but effective way of keeping an overall record is to construct a spreadsheet and keep it up to date. Figure 3 is an example of the information that might usefully be recorded in the register.

Figure 3 Example of a Simple File Register



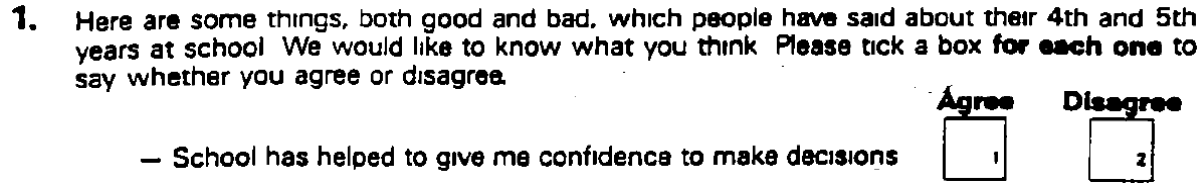
	A	C	D	E	F	G	H	I
1	File Register							
2								
3								
4		File Name (name_subname_date[year/month/day]_depositor's initials_version_type)	File Type					Brief Description of the file and it purpose
5	Directory Name	(e.g. bhps_aindresp_140129_vg_v1.dta)	(e.g. Stata data file)	Name of Author	Initials of Author	Date of Creation	Date of last revision	(e.g. Stata .do file MSc dissertation; Draft Chapter 1 PhD)
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								

Managing Variables

Large-scale social surveys will tend to be well organised and appropriately curated to facilitate research by a wide range of data analysts. The data providers will tend to stick to clearly defined variable naming conventions. For example in wave 1 of UKHLS (UK Data Archive Study Number 6614) *a_sex* is the gender variable, and in wave 2 of the study *b_sex* is the gender variable.

In the Youth Cohort Time Series for England, Wales and Scotland, 1984-2002 (UK Data Archive Study Number 5765) the variable *t0schtyp* is the type of school that the pupil attended in Year 11 (at time point *t0* in the study). By contrast in the less well curated Youth Cohort Study of England and Wales Cohort 4 (UK Data Archive Study Number 3107) the opaque variable *dx11_a* relates to the survey question in Figure 4.

Figure 4 Survey Question From the Youth Cohort Study of England and Wales Cohort 4 (UK Data Archive Study Number 3107)


1. Here are some things, both good and bad, which people have said about their 4th and 5th years at school We would like to know what you think Please tick a box for each one to say whether you agree or disagree.

	Agree	Disagree
– School has helped to give me confidence to make decisions	1	2

Having human-readable variable names increases transparency within the audit trail and also helps the researcher undertaking the data analysis, and other colleagues who might be reading intermediate outputs that are not yet fully worked up into a publication-ready state. In the process of enabling survey data for analyses researchers will routinely need to copy, reorganise and recode variables. In order to achieve an accurate, efficient, transparent and reproducible workflow with a suitable audit trail it is paramount to adopt a clear and consistent approach to naming and coding variables. The creation of new variables and alterations to existing variables must always be accompanied by a permanently documented note or comment.

Syntax Files

We advise against undertaking survey data analyses using software packages in interactive modes, for example through a graphical user interface (gui) such as drop-down menus. If you use a 'gui' then one day you will inevitably end up in a sticky mess. The complete analytical cycle from data downloading through to archiving should always be undertaken using software syntax. Syntax files send commands to the data analysis software. Clear and consistently well organised and annotated syntax files are central to accurate, efficient, reproducible, and transparent survey data analysis.

In this section we discuss the organisation of syntax files using Stata related examples, but the issues are equally pertinent when using other data analysis software packages. Syntax files in Stata are known as 'do' files and have a *.do* extension. Figure 5 provides an example of a template for a blank *.do* file. It is possible to save a file called 'template.do' into your home Stata folder so that a blank *.do* file that is pre-populated with organisational information is automatically generated when you open Stata. This is easily achieved by adding a line to your Stata profile file (*profile.do*) which points to the template, for example using the following line of syntax

doedit "C:\Program Files (x86)\Stata14\template.do"

Figure 5 A Stata .do File Template (part 1)

```
template* ×
 1  STOP
 2
 3  /**
 4
 5  ********************************************************************************
 6
 7  Next Actions:
 8
 9
10
11
12  Author:
13
14
15  Project:
16
17
18  Sub-project:
19
20
21  Date of Next Meeting (or supervision):
22
23
24  Latest Update:
25
26
27  Previous Updates:
28
29
30
31  Useful information:
32  http://www.samaritans.org/ (08457 90 90 90)
33
34
35  ********************************************************************************
```

We often start our *.do* files with the word *STOP* because this is not a valid command in Stata. This little trick ensures that the whole *.do* file can never accidentally be completely executed in a single step. This provides some insurance against accidentally running the complete file which could have potentially damaging consequences.

It is difficult to overemphasise the importance of annotation and making comments in *.do* files. Notes and comments are the building blocks of the audit trail. We recommend that every *.do* file begins with some 'meta-information' that helps to give a reader an overview and to locate its production in time and space. We advocate beginning each incomplete *.do* file with a note on *Next Actions*. This reminds the data analyst of which actions they must undertake next to move the work closer to completion. In practice it is much more convenient to place this at the start of the *.do* file, rather than having it buried somewhere within the file, or at the end of the file which can easily be

several hundred lines long. Having information on the project and authorship is critical for keeping track. Supplementary information such as the date of the next meeting or supervision also provides helpful information in the audit trail.

Programming efficiency is one of the goals of a well-planned and well-organised *.do* file. There are a number of preliminary commands that are useful in the set-up of any data analysis session using Stata.

Figure 6 A Stata .do File Template (part 2)

```
39
40  ********************************************************************************
41
42
43  **/
44
45  clear all
46
47  macro drop _all
48
49  set more off
50
51
52  * set paths *
53
54  /**
55
56  global path1 "f:\working\"
57  global path2 "f:\do_files\"
58  global path3 "f:\data_raw\"
59  global path4 "f:\data_clean\"
60  global path5 "f:\logs\"
61  global path6 "f:\codebooks\"
62  global path7 "f:\temp\"
63
64  **/
65
66
67
68  * log file *
69
70  capture log close
71
72  capture log using $path5\keywords_log_yearmonthday_author_v1.txt, replace text
73
74
75  * data file *
76
77  use $path2\???.dta, clear
78
79
80
81
82  ********************************************************************************
83  /**
84
85
86  Professor Vernon Gayle, University of Edinburgh (vernon.gayle@ed.ac.uk)
87
88  ********************************************************************************
```

Figure 6 shows more of the Stata *.do* file template. It is a good practice to clear the memory before undertaking a new session of work (*clear all*). It is also advisable to clear away any macros (*macro drop _all*), which are useful small sub-programs in Stata. By default Stata pauses after showing a screen-full of output. For many data analysis tasks it is worth turning this feature off (*set more off*). Setting paths is an extremely useful practice because it aids efficient working in projects with multiple users or when moving between computers (e.g. from an office machine to your laptop). Macros are sub-programs in Stata, and there are two types of macros in Stata. In this instance we will be using the global type. A macro is analogous to a kit bag. You can pack a load of things into the kit bag and then give it a name, and it is ready for use later on. When it is required you can tell Stata to go and get the kit bag and unpack its contents.

We specify a series of global macros. The first is called *path1* and points to the directory that contains our raw data. The second macro is called *path2* and points to the directory that contains our clean data. The neat aspect of this practice is that we can get a file from a long path with directories and many subdirectories simply by typing *use $path2\file.dta* which is the path followed by the file name[9]. This is especially critical when working in collaboration with colleagues. This is because the global macro for the path only has to be changed once (i.e. at the start of the file) and not every time a file is opened or saved. This is an example of a small practice that greatly helps to improve efficiency because it utilises the programming capability of the software. We advise that survey data analysts should set their paths so that they are congenial to the directory structure that they have organised for their project.

A log file has been created. The log file is a complete record of the data analysis session and echoes all of the output that appears in Stata's results window. The log file should be a central component of the audit trail. We recommend that data analysts first close any existing log files. This is

[9] The notation shown is technically a shorthand for

"${path1}\file.dta"

(including the quote marks). In most situations the curly brackets and the quotation marks are unnecessary unless the path includes non-standard features (e.g. spaces).

undertaken by the command *capture log close*. The command *log close* will close any existing log files. We prefix this with the *capture* command because this suppresses any error messages, for example if there are no log files to be closed.

A new log file should be started. The command *capture log using* will begin a new file. We can use the global macro to send the file to the correct location *$path6*\ . If the file naming protocol that we suggested above was being used the log file should be

k*eywords_log_yearmonthday_author_version*.

We prefer log files that are in plain text format so we specify a *.txt* extension and the *text* option. This is because files in text format can easily be read by a number of text editors. Stata is a very careful piece of software and does not allow you to overwrite files unless you explicitly tell Stata to *replace* them.

The data file, which in Stata is a *.dta* file, can be acquired from the correct directory using the global macro *$path2*\. The *clear* option simply clears the memory before loading in the data file.

The work undertaken in the session will form the main body of the *.do* file. We recommend keeping things neat and tidy and always indicating where the end of the file is in a comment such as ** End of file **.

Figure 7 provides an example of the start of a genuine *.do* file that Vernon Gayle wrote a couple of years ago. It illustrates how much annotation should be included in a file for even the most simple data analysis activity. Once again we wish to stress the importance of making notes and comments in order to establish accurate, efficient, transparent and ultimately reproducible analyses.

Figure 7 A Genuine Stata .do File

```
*******************************************************************************

/**

Quasi Variance in Stata(qv is a post-estimation command)

Professor Vernon Gayle,
School of Social and Political Science,
University of Edinburgh


        ********************************************
        * IT IS IMPORTANT THAT YOU READ THIS HANDOUT *
        * AND FOLLOW THE STATA.DO FILE LINE BY LINE! *
        ********************************************

A picture paints a thousand words!

In my experience, with the exception of weighting, graphing data is one of the most
troublesome aspects of data analysis. Plotting results from statistical models can
take time and effort.

Here is a brief introduction to the Quasi Variance (qv) function and
related graphs in Stata.

The qv command is a post-estimation command.
It uses e(V) from the most recently fit model to compute quasi variances for
all categories in a multi-category variable.

Aspen Chen's (aspen.chen@uconn.edu) qv package is available from

http://econpapers.repec.org/software/bocbocode/s457831.htm

The package used David Firth's methodology (see Firth 2003).

Gayle and Lambert (2007) provide an explication of this method
with a variety of sociology examples.

Further resources are available at http://www.restore.ac.uk/Longitudinal/qv/


Updated Tuesday 12th August 2014
(Sad news - actor and comedian Robin Williams dies age 63)

**/
```

Conclusions

Analysing survey data without a planned and organised workflow can be compared to drinking and driving. In both situations it doesn't matter how careful you are, it is still highly likely to end in a wreck![10] Therefore just like drinking and driving, we strongly warn against not having a systematic workflow. In our view no researcher should ever undertake any serious survey data analyses without a planned and organised workflow. Long (2009) reminds us that there is an unavoidable tension

[10] We are grateful to Professor Philip Stark, University of California Berkeley, for this useful and clear analogy.

between undertaking work carefully and completing work, but also makes the additional point that the production of incorrect results injures both the researcher and the research field.

This working paper has focussed on the workflow and the very practical aspects of organising the analysis of social surveys. In practice these issues are also pertinent to the analysis of administrative social science datasets which very often are in similar formats to surveys (Connelly et al., 2016). If forced to sum up what lies at the heart of a good workflow we would have to adapt a forthright comment from the great physicist David Mermin[11], and say "shut up and document!".

Using software syntax is central to a well organised workflow. The much reported case of the error in Rogoff and Reinhart (2010a) and Rogoff and Reinhart (2010b) that was detected by Thomas Herndon, provides a beacon that warns against both using software interactively, and not having a documented workflow (see Herndon et al., 2014). Much of the time spent analysing social survey data involves undertaking repetitive tasks. As Long (2009) helpfully reminds readers, repetition invites errors and automation is both faster and less error prone.

Time invested in data enabling is seldom wasted and will always pay dividends in the longer term. The French culinary term '*mise en place*' refers to the setting up that is required before cooking commences (the phrase roughly translates to 'everything in its place'). This might provide a suitably more exotic image for readers who imagine that data enabling is a mundane activity. Unless you are certain that your programming skills are at an advanced level then the 'fives times' guideline is worth being mindful of. In all cases we strongly advise you to ensure that you build in adequate time for enabling your data, and try to remember that data enabling tasks are usually very time consuming.

Large-scale social survey datasets are seldom delivered to the data analyst in a format that makes them immediately ready for comprehensive analyses. This is typically because there are several different data files supplied. There are often a number of different measures (e.g. different socioeconomic classifications and measures of income) in the dataset that could be used, and often

[11] See http://www.gnm.cl/emenendez/uploads/Cursos/callate-y-calcula.pdf accessed 20.06.16.

metadata are also supplied with the dataset. Developing good data enabling skills pays important dividends because they allow researchers to take fuller advantage of the richer data resources available. Good data enabling skills will free researchers from the shackles of having to analyse only the most readily available information.

The most obvious payoff for having an effective workflow is that accurate results can be produced systematically and efficiently. There are other extrinsic rewards that emerge from establishing a good workflow. For example a number of high quality academic journals such as *Science*, *American Economic Review*, *Econometrica* and the *Review of Economic Studies* now require supporting computer code that is involved in the creation and analysis of research data (McCullough et al., 2008, Hanson et al., 2011). Over five hundred journals across a range of academic disciplines are now signatories of the Transparency and Openness Promotion (TOP) Guidelines, which encourage transparency, open sharing and reproducibility[12].

A well organised workflow is what social psychologists call a positive valance concept, by which we mean that it is intrinsically attractive. Academic life is measured by outputs and structured by deadlines, and an organised workflow can make an essential contribution to making progress. Oliveira and Stewart (2006) conclude that if your program (in our case the syntax file) is not correct nothing else matters, for without correctness we cannot expect any useful results. Long (2009) reminds us that getting the correct answer is the *sine qua non* (i.e. the essential and indispensable action) of a good workflow.

[12] See https://cos.io/top/ accessed 25.02.16.

References

CONNELLY, R., PLAYFORD, C. J., GAYLE, V. & DIBBEN, C. 2016. The role of administrative data in the big data revolution in social science research. *Social Science Research*.

HANSON, B., SUGDEN, A. & ALBERTS, B. 2011. Making Data Maximally Available. *Science,* 331**,** 649.

HERNDON, T., ASH, M. & POLLIN, R. 2014. Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge journal of economics,* 38**,** 257-279.

IHAKA, R. 1998. R: Past and future history. *COMPUTING SCIENCE AND STATISTICS***,** 392-396.

KOHLER, H. P. & KREUTER, F. 2012. *Data Analysis using Stata,* College Station, Tx, Stata Press.

LAMBERT, P. S., BROWNE, W. J. & MICHAELIDES, D. T. 2015. Contemporary developments in statistical software for social scientists. *In:* PROCTE, R. & HALFPENNY, P. (eds.) *Innovations in Digital Research Methods.* London: Sage.

LONG, J. D. 2011. *Longitudinal data analysis for the behavioral sciences using R*, Sage.

LONG, J. S. 2009. *The Workflow of Data Analysis Using Stata,* College Station, Stata Press.

MCCULLOUGH, B. D., MCGEARY, K. A. & HARRISON, T. D. 2008. Do Economics Journal Archives Promote Replicable Research? *The Canadian Journal of Economics / Revue canadienne d'Economique,* 41**,** 1406-1420.

NIE, N. H. 1983. SPSSX user's guide. *SPSS Inc, North Michigan, USA*.

OLIVEIRA, S. & STEWART, D. E. 2006. *Writing Scientific Software: A Guide to Good Style*, Cambridge University Press.

PEVALIN, D. & ROBSON, K. 2009. *The Stata survival manual*, McGraw-Hill Education (UK).

PINZON, E. 2015. *Thirty Years with Stata: A Retrospective*, Stata Press.

ROGOFF, K. & REINHART, C. 2010a. Growth in a Time of Debt. *National Bureau of Economic Research Working Paper No. 15639*.

ROGOFF, K. & REINHART, C. 2010b. Growth in a Time of Debt. *American Economic Review,* 100**,** 573-8.

TREIMAN, D. J. 2009. *Quantitative Data Analysis: Doing Social Research to Test Ideas,* San Francisco, John Wiley & Sons.

Data Citations

Courtnay, G. (1993). *Youth Cohort Study of England and Wales, 1989-1991 Cohort Four, Sweep One to Three.* [data collection]. UK Data Service. SN:3107, http://dx.doi.org/10.5255/UKDA-SN-3107-1.

Croxford, L., Iannelli, C., Shapira, M. (2007). *Youth Cohort Time Series for England, Wales and Scotland, 1984-2002*. [data collection]. UK Data Service. SN: 5765, http://dx.doi.org/10.5255/UKDA-SN-5765-1.

University of Essex. Institute for Social and Economic Research. (2010). *British Household Panel Survey: Waves 1-18, 1991-2009*. [data collection]. *7th Edition.* UK Data Service. SN: 5151, http://dx.doi.org/10.5255/UKDA-SN-5151-1.

University of Essex. Institute for Social and Economic Research, NatCen Social Research. (2015). *Understanding Society: Waves 1-5, 2009-2014*. [data collection]. *7th Edition.* UK Data Service. SN: 6614, http://dx.doi.org/10.5255/UKDA-SN-6614-7.

Acknowledgements

We would especially like to thanks Roxanne Connelly and Sarah Stopforth for their insightful comments and editorial advice. We would also like to thank Hannah Buchanan-Smith, Chris Playford, Kevin Ralston, Audrey Thomas and Malcolm Quon their comments and their helpful suggestions. Over the last two decades we have delivered a wide range of workshops and training events. This working paper has greatly benefitted from comments and questions from numerous participants who have attended these events. Vernon would also like to thank U.C. Berkeley for providing a pleasant and engaging environment in which to think about the social science workflow and producing accurate, efficient, transparent and reproducible social science data analyses.