Notes on SPSS files for European Social Survey 2002

From John Hall 8 Nov 2004 to ESS team

Dear Colleague(s)

I've just received the ESS Bulletin #1 and thought rest of the team might be interested, including overseas colleagues, in my comments.

Please find below an extract (blue) from a private mail I sent Roger Jowell on 21 June.

I read the O'Shea, Bryson and Jowell paper and it took me back more than 30 years to the original QoL proposals I did with Mark Abrams, Angus Campbell, Norman Bradburn and Aubrey McKennell as well as an application I put in to ESRC in 1984 with Dave Phillips for the British end of the Values and Social Problem Indicators in Contemporary Europe (Chaired by Walter Rüegg, but driven by Louis Guttman's facet theory). I also downloaded the 2002 data from Norway, extracted the British subset and played around with it for a bit.

I see you've stuck to the same variable naming convention you inherited from John Curtice's early efforts on BSA, and I'm sure you'll have the same reasons for doing so. However, as with BSA, unless a user is only interested in repeat measures over time (not sure these are advisable on single item measures anyway) it is difficult if not impossible to find his way around the data even for an experienced and expert user like me, let alone the sort of students and clients we used to get at PNL.

Based on our experiences at SSRC/LSE, but especially for teaching purposes at PNL, I always used to create special SPSS setup files for the BSA and other surveys using **rename** to change all the variable names to our positional convention (ie **v362** would be the variable containing the data (starting) from card 3 col 62). This made it far simpler and easier to work directly from your questionnaires, which had record (card) and column locations printed in the margins, and is much less prone to run errors caused when users try to remember all the variable names. (Another setup file was always available to change them all back again: Graham Farrant did these for the 1989 survey.) Your ESS questionnaire doesn't even have record and column locations (presumably owing to modern data capture methods!) but most if not all of it can be done using question numbers as variable names.

As well as the, to me, somewhat tortuous variable naming, I also find problematic some of the labelling used in the SPSS file for the ESS2002. Some of the variable labels with repeated text are very long, with key information at the end, so for some procedures where SPSS has limits of 40 characters (eg charts and multiple response tables), the key information is chopped on the output. It's the sort of thing I used to get from novice or inexperienced researchers or students (for which they would lose assessment marks). One or two liberties have also been taken with the question wording when condensing to variable labels.

SPSS has many strides in the 33 years I've been using it (not all to the good). Thus while frequency counts and contingency table presentation have improved and can handle longer value labels, the multiple response procedures still produce output identical with that from 30 years ago. This means that column headers are printed out in 2 groups of (up to) 8 characters, and this looks odd on the output.

For my purposes on the 2002 data, I've changed most variable names to tally with the questionnaire (mostly question numbers, but some slightly modified for complex questions) and put the question number at the beginning of each variable label. With these changes I can now work directly from my SPSS file, and consult the questionnaire if I get stuck. I appreciate that question numbers and content may change with successive waves, but I still think my system is far easier to use and less error prone. For analysis over time it's a simple matter to produce conversion tables

from my style of variable names from each wave to the names you currently use. In any case, over time, I would have thought you would be producing standard derived variables based on scaling, and there's no reason why these shouldn't remain constant over all waves. After all, **anomy** is much easier to understand than **pplhlp**! When I've sorted some examples out, I'll send you them.

On another, but related matter, I've also now completed the restoration of the SSRC Survey Unit Quality of Life surveys and one I did with Mark on attitudes and opinions of senior girls at a public school. I'm attaching flysheets giving summary descriptions of the questionnaire coverage, which has significant overlap in content with ESS, but different response scales. This is a shame, but that's life. If you're interested in the actual questionnaires and portable SPSS data files, you can get them from the archive at Essex (much quicker) or I can send them to you (the SPSS **por** and Adobe **pdf** files, not the published ones for 1973 and 1975 as I only have one copy of each).

As expected, Roger has stuck to his original method for variable names, but I have been experimenting with my method on an extracted data for UK only. I attach the file **ess02rename.sps** which renames your variables to my convention and produces a file which is far easier to use directly from the questionnaire. The rationale behind my naming convention is explained in the attached teaching note **names.doc**. My convention is also easier to use with (the abhorrent) point and click in SPSS as at least you have some idea where you are in the file. However I prefer to work in syntax mode.

Below is a copy (red) of a note to myself when working on your data: this is self-explanatory.

European Social Survey 2002

Variable names and labels a mess as supplied. Need something to enable analysis with questionnaire in front, so have added question number to labels. Have also renamed variables where these are common to all countries. OK I know the questions will change next year, but with a set of renaming tables for pedants this should not present insuperable problems. Whoever did the var labels was inexperienced. The most important bit is often at the end and therefore likely to get chopped on output, especially in multiple response: they've also taken a few liberties with wording.

I also attach my version of the UK file **essuk02jfh.por** which has the modified variable labels, but I have attached a version with your original names as well **essuk02.por**. Variables relating to non-UK coding (eg political parties) have been left with your original names and labels. I still can't get my head round SPSS pivot tables (haven't found out how to get past the red triangle which truncates the output, or probably did once and have forgotten!) otherwise I could have sent you the output from the **disp lab.** command instead. It used to be much easier on the Vax mainframe when you could pull vertical strips off and paste them back on somewhere else.

Best wishes from an old survey addict.

John F Hall (retd)

Reply from Caroline Roberts on 2 Dec 2004 to John Hall

Dear John,

Please accept my apologies for not having replied to you sooner.

I had a lengthy discussion with my colleagues at City and with Roger about your comments and suggestions. We all agreed that the current 'mnemonic' system for naming variables on ESS datasets is not the most user-friendly, and for the purposes of cross-sectional analysis, it does make sense to have variable names and labels that correspond more readily with the questionnaire. Indeed, somebody else contacted us earlier this year and raised points similar to those that you made, and I am sure that many others would agree. However, from the point of view of archiving time-series data, it seems that the solution that has been adopted is much safer than using say, question numbers, since it makes it much easier to handle changes to the questionnaire over time. In fact, similar time series such as the US GSS and ONS surveys also use mnemonic variable names for the same reason. In addition, in an albeit very small number of cases, certain countries have been permitted to insert additional questions into the core of the questionnaire, and this could also lead to confusion over question numbers, were such an approach to variable naming adopted.

So, as you suspect, we are sticking with the present system for variable naming and labelling (although we will pass on your comments about long variable labels to our colleagues at the ESS data archive).

From my own point of view, I found your syntax for renaming and labelling variables most helpful and I am certain that other users would also. There are plans to set up a forum for ESS data users to share and exchange ideas, analyses, helpful tips and syntax, etc. on the NSD ESS data archive website and we all agreed that this would be an ideal way for you to disseminate your syntax to other users. We will, of course, inform you and other data users via our news bulletin when this 'user forum' gets off the ground. When it is launched, I do hope you will consider passing on your syntax to others who may also find it useful.

With kind regards,

Caroline