Comments on the distributed SPSS file for British Social Attitudes 2011 (UKDS SN 7237)

John F Hall **[7 Oct 2013]**

Email: johnfhall@orange.fr
Website: www.surveyresearch.weebly.com

For many years I used data from the BSA surveys for my Survey Analysis Workshop course (hands-on, postgraduate, part-time, evening) at the then Polytechnic of North London (PNL). The course made intensive use of SPSS-X on a Vax mainframe on the Holloway campus via fast servers and visual display units in a computer lab on the Highbury campus. It was aimed at students, researchers and teachers in the social sciences with little or no previous experience of surveys, computing or statistics (some of them could not even type!). The course closed when I (early) retired in 1992, but most of my teaching and research materials from 1986 to 1992 were preserved in digital form on floppy disc or magnetic tape.

Since 1992 several more SPSS-based courses have sprung up elsewhere, but none of them have quite the same pedagogic approach: many concentrate on imparting (inferential) statistical and sampling theory rather than substantive research questions and tend to rely on equations rather than tables or charts.

Consequently, since 2009, I have been developing a website Journeys in Survey Research on which I am gradually converting, updating and greatly expanding my tutorials, examples and exercises for use with SPSS for Windows on a PC (currently release 22). Because most of my 1991-92 teaching materials were based on the BSA 1989 data, I am still using these in exercises and examples, but I have now started looking at more recent data from the BSA, European Social Survey, the General Social Survey (NORC) and similar surveys, starting with BSA 2011.

On opening the SPSS file for SN 7237 from UKDS, I have encountered a number of issues which make the data less than ideal for immediate use in my tutorials. The issues I face are less to do with the data themselves than with incorrect specification of measurement levels, incomplete declaration of missing values, unhelpful presentation of variable and value labels, and my being pedantic about spelling and the use of upper case letters at the beginning of sentences. This sort of thing is likely to put off tutors and students wishing to use the data, especially if they are doing analysis within a single wave rather than looking for trends and comparisons across time.

I have been trying to replicate on the BSA 2011 survey some analysis from the BSA1989 survey on differences in earnings between men and women from paid work, starting with zero order contingency tables for sex and earnings and proceeding to second and third order tables to test for the effects of selected control variables such as educational level, type of work, hours of work, age etc. (elaboration). The same example is used to introduce null hypothesis and the idea of statistical relationships between variables, in this case to build up an equation for chi-square, in fact the first time an equation appears. Later exercises will use the same logic for means and variances, again building up equations for statistical tests such as t-test and anova.

You can see how I have used the 1989 data in 2.3.1.6.2 Specimen answer for conditional frequencies homework [Tasks 3 and 4] and the new series of (draft) tutorials, 3.1.4.1 to 3.1.4.5, on page 3.1 Two variables (CROSSTABS) on my site. These use the 1989 earnings data because that was what I still have in my handouts and exercises, but the BSA series also has other more interesting (to my students) variables covering attitudes to abortion, welfare scroungers, “redneck” authoritarian politics etc, especially if the questions have been replicated in the ESS or the GSS.

Variable names and variable labels

Generating appropriate and accessible teaching materials from real surveys such as BSA 2011 can take an inordinately long time. I know because I once spent 19 hours creating a single exercise based on the 1989 survey. Recently I have spent many hours, days even, working through the BSA 2011 SPSS file (and composing this comment). An early problem for me was that, on the questionnaire, the names of the first variables I needed were not always the same as the names in the SPSS file (eg **NSSECG** on the questionnaire, **RNSSECG** in the SPSS file).

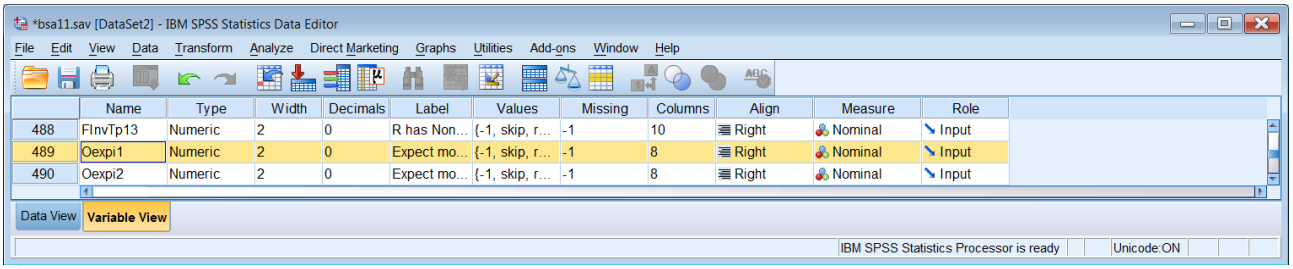
Some variables listed on the questionnaire are not even in the SPSS file (eg. the full set of 19 income groups below) **rearn** is not there, only derived variables in deciles **rearnd** or quartiles **rearnq**).

ANNUAL earnings BEFORE tax

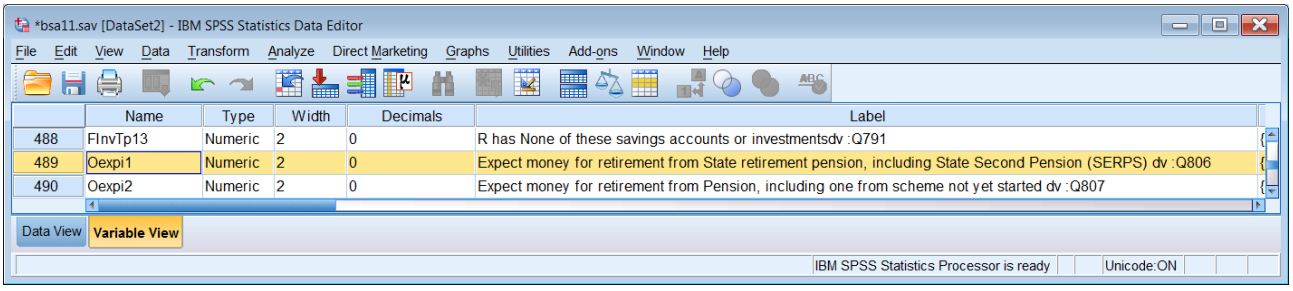
Q	Less than £3,270
T	£3,271 - 5,210
O	£5,211 - 7,130
K	£7,131 - 9,350
L	£9,351 - 11,200
B	£11,201 - 12,700
Z	£12,701 - 14,200
M	£14,201 - 15,600
F	£15,601 - 17,000
J	£17,001 - 18,600
D	£18,601 - 20,400
H	£20,401 - 22,100
A	£22,101 - 24,100
W	£24,101 - 26,500
G	£26,501 - 29,400
N	£29,401 - 32,600
X	£32,601 - 36,900
C	£36,901 - 43,200
P	£43,201 - 58,500
E	£58,501 or more

Another problem was the sheer length of some labels, eg:

Name **Label**
Oexpi1 Expect money for retirement from State retirement pension, including State Second Pension (SERPS) dv :Q806

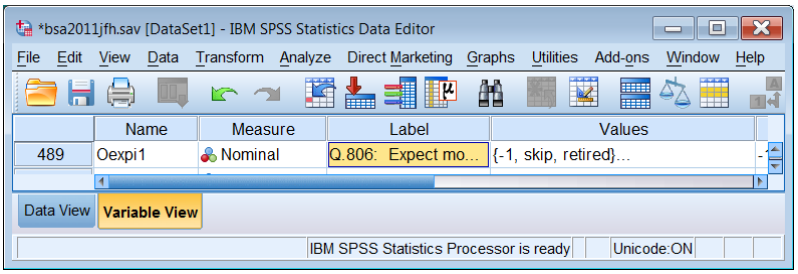


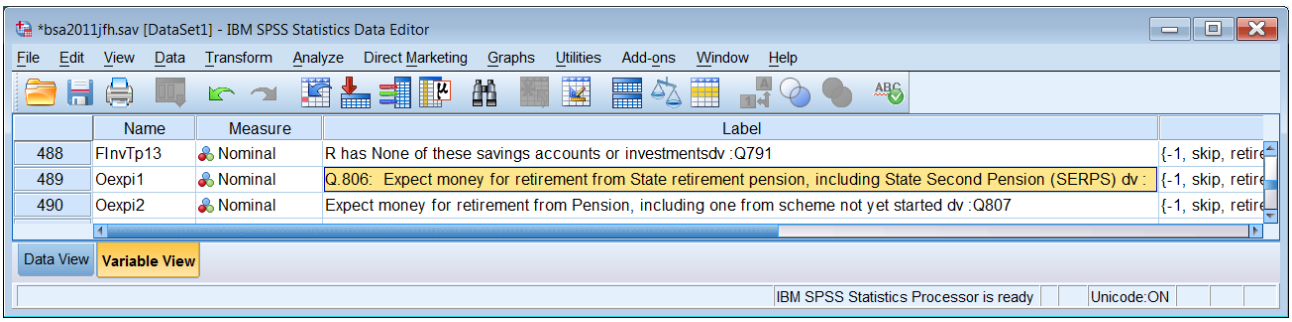
The default display from SPSS (above) is singularly uninformative, and the **Label** column has to be widened to find, right at the end, the question number to which it refers (below):



Natcen researchers point out that the variable labels (far too long in my opinion) are used as table titles. making for easier understanding of published tables, but even titles would be much better with the question number at the beginning rather than the end: that way the question numbers all line up in the Data Editor as well, even in the default display.

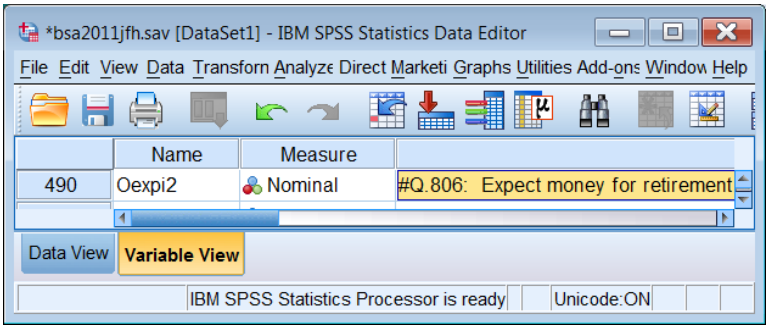
Name **Label**
Oexpi1 Q.806 Expect money for retirement from State retirement pension, including State Second Pension (SERPS) dv



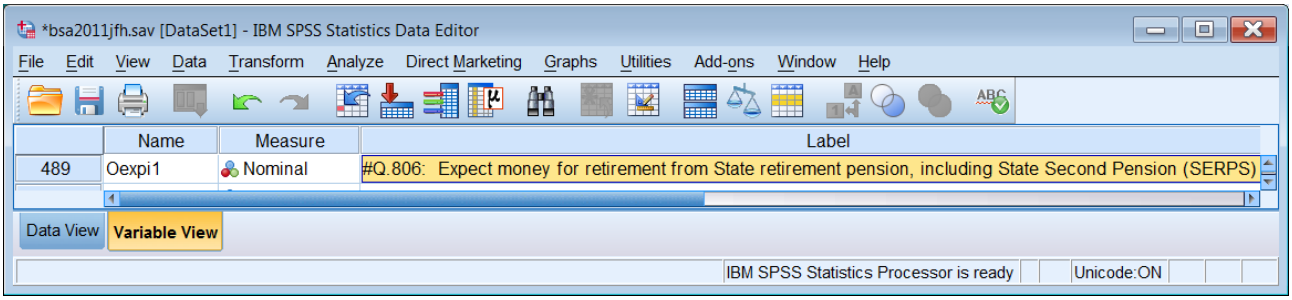


In BSA variable labels, **dv** (or **DV**) indicates a derived variable and is perhaps superfluous if the label can be amended to something like:

Name **Label**
Oexpi1 #Q.806 Expect money for retirement from State retirement pension, including State Second Pension (SERPS)



Even then, important information about including **SERPS** is right at the end of the label where it won't show up on-screen without massively widening the **Label** column:



I suspect that problems like this arise to some extent from using chunks/logic/sequence of old SPSS/Blaise setup jobs to save writing them from scratch (a leftover from the SPSS file conventions John Curtice and Ann Mair created at Strathclyde for the original surveys from 1983 onwards). One clue is the continued use of 8-character variable names, a limit which was raised years ago (although I did find one with 9 characters, probably a new variable added in later series).

I'm very aware of the history of the SPSS variable names used in BSA, and over the years had many discussions with Roger Jowell and Sharon Witherspoon about them (and also about the use of 0-10 scales in BSA and ESS). When using a single survey I find **mnemonic** names ugly, contorted, open to spelling errors when writing them in SPSS syntax (as opposed to clicking in the GUI) and impossible to find without extensive scrolling through the SPSS file and/or the questionnaire. However, I accept that keeping the names constant across survey waves was the main reason for using (and keeping) them.

The raw data files for the early BSA series were easier to use as they were stored as card images in 80-column ASCII files, 23 records per case, and the data entry location was always indicated in the right hand margin on the printed questionnaires. When reading raw data into SPSS this meant I could use variable names directly related to their data entry location. A variable entered in record 2, column 3 would be named **V203** (column locations were always two digits in the range 01 – 80) a variable entered in record 15, column 11 would be named **V1511**. Working from a printed questionnaire, a glance at the data entry information in the right hand margin (for any question) immediately indicated what the corresponding variable would be called in the SPSS file. Conversely, working from the SPSS file, even if there were no variable labels, a variable name immediately indicated where to find the corresponding question on the printed questionnaire.

This convention, which I dubbed **positional** naming, was based on the variable naming in Peter Wakeford's program SDTAB at LSE, which used 3- or 4- digit variable names such as **203** or **1511**. When SPSS became available, I simply adapted the convention by adding the prefix **VAR**, the only automatically generated names then allowed using **VAR**xxx **TO VAR**yyy: later releases allowed **V**xxx **TO V**yyy and later still, letters other than V, even in lower case, eg **q**xxx **to q**yyy. (See: 1.3.1 Conventions for Naming Variables in SPSS). Occasionally, to get more codes on a single card-column, codes for variables would be entered as letters of the alphabet: these would be prefixed **A**ddd and later converted to numeric as **V**ddd. Data from field agencies was sometimes supplied in multi-punched format (eg. answers to multiple response questions or more than one variable coded in the same column, eg.sex, class, marital status). Such data would be spread out on additional records, using Peter Wakeford's MUTOS program, before being read in to SPSS.

This convention was standard for all surveys done by the SSRC Survey Unit from 1972 onwards and, with one notorious exception, for all surveys done, or advised on, by my Survey Research Unit at PNL from 1976 to 1992. The main advantage was that it was easy to understand and apply when several people were working on the same survey, including clients, and was impervious to staff turnover and individual literary idiosyncracies. When we were asked to prepare the SPSS file and user manual for the 1982 British Crime Survey, we also used the convention **M**ddd for multiple response data.

The advent of CAPI and Blaise rendered this convention impossible, short of using the question numbers as variable names. Users (especially me) had to spend precious time scrolling up and down the labels in the SPSS file to find the variables they wanted.

As well as raw data files, early BSA surveys also came with their associated SPSS setup files, written by John Curtice and Ann Mair at Strathclyde (I have one somewhere, thousands of lines of syntax specifying variable labels, value labels, missing values laboriously one variable at a time: not a single **Vxxx TO Vyyy** in the whole file.) but they are now all distributed as SPSS *.sav files.

For teaching purposes I always changed variable names in the BSA files from mnemonic to positional to make it easier for me and my students to find variables when working from a hard copy of the questionnaire. The approach I adopted at PNL was to rename most of the variables by

their location in the raw data. This convention invariably put the question number at the beginning of the variable label: for variables without a question number (eg. for household grids) labels were prefixed with distinct characters ###. Graham Farrant (ex Natcen, now SRA) composed the SPSS RENAME VARIABLES setup file to do this for the 1989 BSA when he was one of my final year students at PNL. The modified SPSS file enabled students to work directly from facsimiles of the original printed questionnaires, and it's the one I still use for tutorials on my site.

With the advent of CAPI and Blaise, this convention can no longer be used. Secondary users of BSA 2011 are presented with a *fait accompli* in which names and labels look like this:

13	RSex	Numeric	2	0	Person 1 SEX :Q49
14	RAge	Numeric	2	0	Person 1 age last birthday :Q50

698	REarnD	Numeric	2	0	Respondent pre-tax earnings deciles (dv) :Q1208
699	REarnQ	Numeric	2	0	Respondent pre-tax earnings quartiles (dv) :Q1209

Moving the question number to the beginning of the labels[1] helps:

13	RSex	Numeric	2	0	Q.48: Person 1 SEX
14	RAge	Numeric	2	0	Q.50: Person 1 age last birthday :Q50

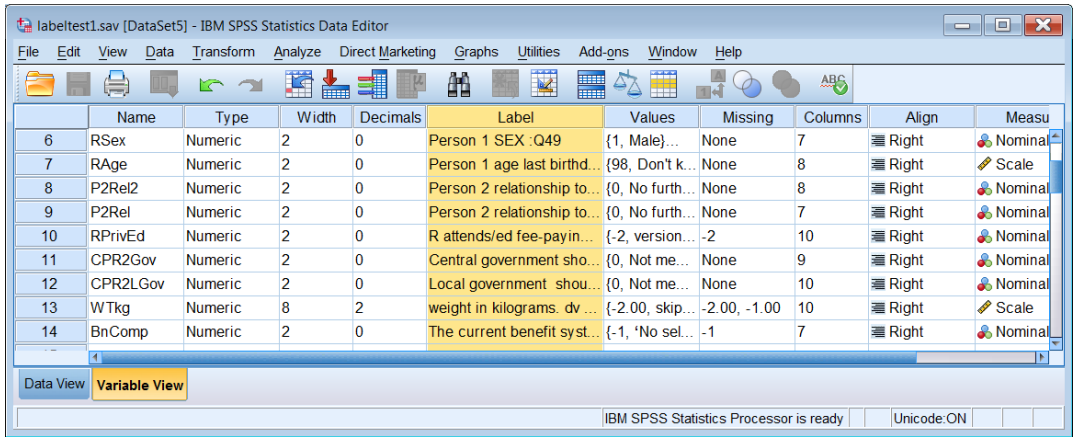
698	REarnD	Numeric	2	0	Q.1208: Respondent pre-tax earnings deciles (dv)
699	REarnQ	Numeric	2	0	Q.1208: Respondent pre-tax earnings quartiles (dv)

My own labeling preference for sex and age of respondent would be[2]:

13	RSex	Numeric	2	0	Q.48: Sex of respondent
14	RAge	Numeric	2	0	Q.50: Age of respondent (last birthday)

This would at least help when using the GUI when the SPSS settings display labels rather than names, since they would then appear in the dialog boxes and incidentally improve the presentation of output tables.

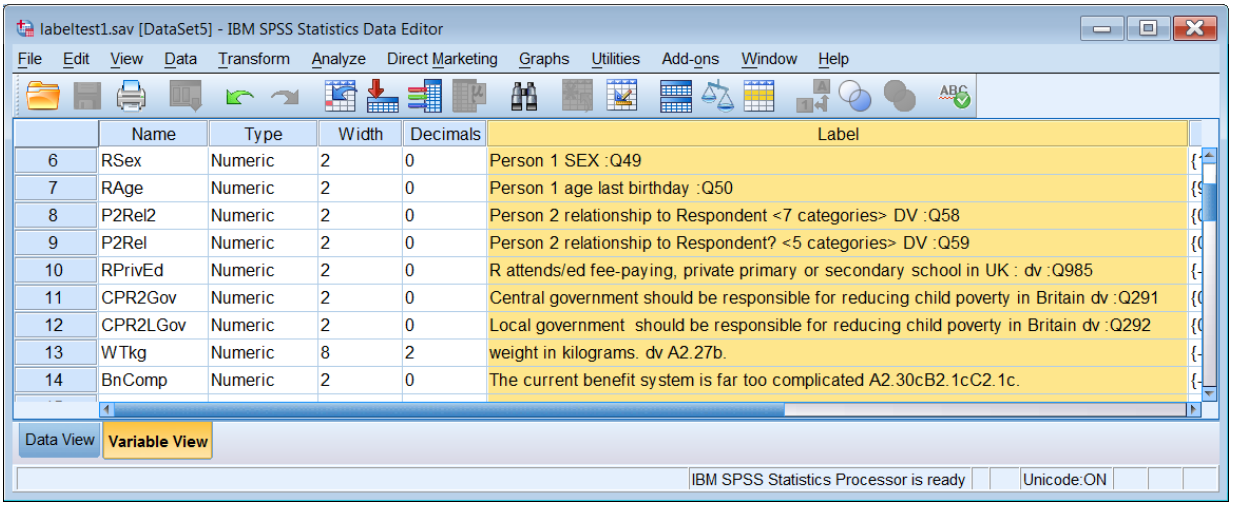
For instance, I created a small file with a few selected variables from BSA 2011:



[1] var lab rsex 'Q.48: Person 1 SEX'
/rage 'Q.50: Person 1 age last birthday :Q50'
/rearnd 'Q.1208: Respondent pre-tax earnings deciles (dv)'
/rearnq 'Q.1208: Respondent pre-tax earnings quartiles (dv)'.

[2] var lab rsex 'Q.48: Sex of respondent'
/ rage 'Q.50: Age of respondent (last birthday)'.

You can't tell to which questions the variable names relate: the edge of the **Label** column has to be dragged way out to the right to reveal the full variable labels



An alternative is to move the question numbers to the beginning of the variable labels.

6	RSex	Numeric	2	0	Q.49: Person 1 ...	{1, Male}...	None
7	RAge	Numeric	2	0	Q.50: Person 1 ...	{98, Don't k...	None
8	P2Rel2	Numeric	2	0	Q.58: Person 2 r...	{0, No furth...	None
9	P2Rel	Numeric	2	0	Q.59: Person 2 ...	{0, No furth...	None
10	RPrivEd	Numeric	2	0	Q.985: R attend...	{-2, version...	-2
11	CPR2Gov	Numeric	2	0	Q.291: Central ...	{0, Not me...	None
12	CPR2LGov	Numeric	2	0	Q.292: Local go...	{0, Not me...	None
13	WTkg	Numeric	8	2	Weight in kilogr...	{-2.00, skip...	-2.00, -1.00
14	BnComp	Numeric	2	0	The current ben...	{-1, 'No sel...	-1

This shows the question number immediately, even in the default Data Editor in Variable View, but the right edge of the **Label** column can again be dragged out to reveal the full text of the variable labels:

6	RSex	Numeric	2	0	Q.49: Person 1 sex
7	RAge	Numeric	2	0	Q.50: Person 1 age last birthday
8	P2Rel2	Numeric	2	0	Q.58: Person 2 relationship to respondent <7 categories> dv
9	P2Rel	Numeric	2	0	Q.59: Person 2 relationship to respondent? <5 categories> dv
10	RPrivEd	Numeric	2	0	Q.985: R attends/ed fee-paying, private primary or secondary school in UK dv
11	CPR2Gov	Numeric	2	0	Q.291: Central government should be responsible for reducing child poverty in Britain
12	CPR2LGov	Numeric	2	0	Q.292: Local government should be responsible for reducing child poverty in Britain
13	WTkg	Numeric	8	2	Weight in kilograms. a2.27b.
14	BnComp	Numeric	2	0	The current benefit system is far too complicated a2.30cb2.1cc2.1c.

This makes the file much easier to understand and navigate, but a further refinement would be to delete **dv** from the labels of derived variables and prefix them with a special character to make them stand out, eg #:

6	RSex	Numeric	2	0	Q.49: Person 1 sex
7	RAge	Numeric	2	0	Q.50: Person 1 age last birthday
8	P2Rel2	Numeric	2	0	# Q.58: Person 2 relationship to respondent <7 categories>
9	P2Rel	Numeric	2	0	# Q.59: Person 2 relationship to respondent? <5 categories>
10	RPrivEd	Numeric	2	0	# Q.985: R attends/ed fee-paying, private primary or secondary school in UK
11	CPR2Gov	Numeric	2	0	Q.291: Central government should be responsible for reducing child poverty in Britain
12	CPR2LGov	Numeric	2	0	Q.292: Local government should be responsible for reducing child poverty in Britain
13	WTkg	Numeric	8	2	Weight in kilograms. a2.27b.
14	BnComp	Numeric	2	0	The current benefit system is far too complicated a2.30cb2.1cc2.1c.

There is a remaining problem with variables in which the same question was asked, but there are three different question numbers depending on which version of the questionnaire it appeared in. The list of three different question numbers for the same variable appears at the end of labels (concatenated, with no spaces) but would look cumbersome and cluttered if moved to the beginning of the variable label.

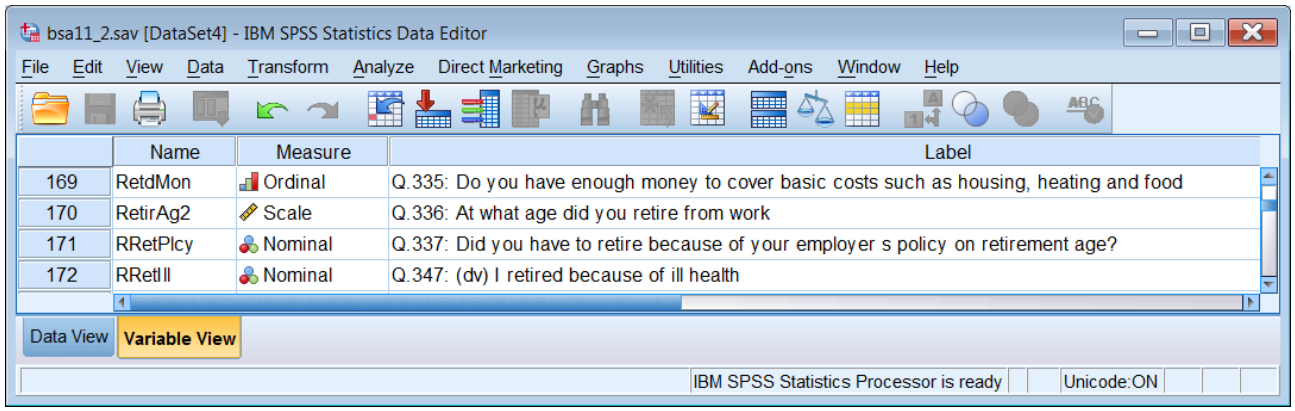
	Name	Type	Width	Decimals	Label
14	BnComp	Numeric	2	0	The current benefit system is far too complicated a2.30cb2.1cc2.1c.

One solution might be to prefix such labels with another distinct symbol, eg @

14	BnComp	Numeric	2	0	@ The current benefit system is far too complicated a2.30cb2.1cc2.1c.

Many variable labels also start with a lower case letter, offending my sense of grammar. They all need changing to upper case. Making these changes manually one label at a time in the Data Editor or in SPSS syntax would be extremely tedious and time-consuming, so I posted the problem to the SPSS-X mailing list.

Jon Peck of IBM/SPSS has kindly provided a Python program which automatically moves question numbers starting with Q to the beginning of the labels, inserts a full stop between Q and the number (**Q.**) and a colon and two spaces after the question number (**:□□**). It also moves free-standing **dv** (or **DV**) to before the main text (in brackets just after the question number, retains upper case letters and changes the first character of the main label to upper case, even if it was originally lower case. Apart from the major improvement in aesthetics, the file is now much easier to navigate.



The above extract is from a file in which I have also reset all the measurement levels.

Jon is still working on the labels for questions which have different numbers depending on which version carried them: all of these start with **a.** , **b**. or in one case **c.** However some of these refer to a single variable (version A only), some to two (versions A and B or B and C) and some to all three. We think the best solution for now is to denote such variables by a special character @ and two spaces at the beginning of the label:

Name	**Label**
Redistrb	@ Government should redistribute income from the better-off to less well-off A2.48aB2.25aC2.24a

Measurement levels

Almost all of the measurement levels in BSA 2011 are set incorrectly at **Nominal**. Some are truly **Scale**: others, whilst technically **Scale**, are actually guesses or pseudo-logarithmic and at the higher values tend to cluster around multiples of 5 or 10 (eg the questions about how many beans to allocate to social policy targets, how many hours are spent watching tv, on internet etc., or how long R has been in current relationship).

Many of the other variables set at **Nominal** should be **Ordinal** but some are actually **Scale**.(eg **NCh415** to **RCh1617**, numbers of children).

Scale variables: to check (File position)

Rage 14 WWWHrsWk 113 CPWNone 167 RetirAg2 170
HipOpA 191 HmHelpA 192 WtHelpA 194
HipOpB 198 HmHelpB 199 HipOpB 210 HmHelpB 202 CnslHelp 203 WtHelp 204
AirTrvl 226 AfDied 320 IqDied 329 LATlenb 379
SEmpNum 440 WkJbHrsI 447 EJbHrsX 448 RetExpb 472
PenExp 505
Tea2 600 HTcm 757 WTkg 758 ShrtJrn 779 leftrigh 852 libauth 853 welfare2 854

Scale variables, should be Ordinal

RAgecat3 73

Nominal variables: should be Ordinal: (File position)

RAgeCat RAgeCat2 71-72
TVNews WebNews 106 – 107
Idstrng 120 Politics 121 SocTrust 122 UKSpnGBE 124
SocSpnd1 to SocSpnd6 129 - 134
TaxSpend 136 HIncDif4 137
UBprobs 139 UBreq 140 RetdMon 169
NHSSat to NHS5yrs 179 – 185
TRFPB6U to TrfConc3 208 – 213 DRIVMIL 215
CycConf 217 CycDang 218 TRAVEL1 to TRAVEL6 221-225
CliCar CliPlane 232 233
CCASpe to CCAPLANE 251-254
MiL10Yrs to MiCultur 256-259 MiGdBad 264
MiGroup1 271 MiGroup2 276 MiGroup3 281
ASAStay 287 ASWork 290 ASApplic 292
ASGdBad 297 OpAF 313 AfOpChg 314 DefSpend 315
AfAccept to AfSupMem 321 – 325
IqAccept to IqSupMem 330 – 334
ESCompND ESCompDi ESCompHo 374-376
REmpWork 441 REmpWrk2 442
RNSEG to RNSSECG 459 -465
PenKnow2 to MPsTrust 506 -509 LoseTch to GovComp 510-515 TrstParl to TrstGov 516-519
QuitRule QuitBdJb 534 535 Lords11 538 CLLRTXTR to PCPOLINF 540-546 SRPrej 570
LETIN 636 S2Class to S2NSSECG 652 – 654 HHIncD to REarnQ 696-699
RUHappy2 to AltVisit 702 -740 BestTrt to LTSGnHth 745 – 755
PrivMCov to RThDsPrd 760-772 HlthGap to Carbike2 775-778
PlnAllow to MobDLaw 780-806
RelMarFC RelMarMu 807 -808 MigWorPS to MediaEx 809-818
FreqElec 823 CTAXREF2 to HLExpert 828-831
ChCandte 832 WelfHelp to Censor 833 – 851

Nominal, should be Scale: (File position)

NCh415 to RCh1617 83-92 CarNum 216

HipOpA to Total 191 – 205
[allocation of beans, but these cluster round 5s and 10s with little discrimination between]

WtFactor OldWt Househld Rage RAgecat3

Partially ranked

[Main responses ranked, but some off-rank answers: 7 = Other etc]

DBwork 143 RetExp 471 NatIdB 568 PrejNow 569 ResPres 573
ChAttend 578 Tea 601 HEdQual 634 HEdQual2 635

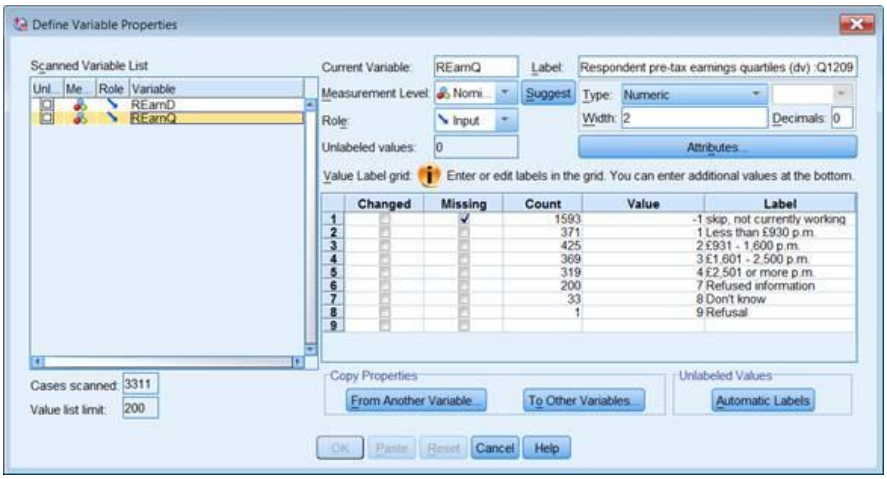
AFWork to ESNoHome 337 – 364 AfOpChg childopsh smokops2 ageopsh2
[currently Nominal, but could be recoded (2=3)(3=2) to yield Ordinal]

Missing values

Another problem is that not all missing values have been declared, apart from negative values (and not always then). Whole swathes of values in the range 7, 97, 997, 9997 (Uncodable) 8, 98, 998, 9998 (Don't know) and 9, 99, 999,9999 (Refusal) have not been so declared, even though they are explained in the user documentation.

In many cases there are up to 5 values (and in one case 8) which can be, or need to be, treated as missing, but SPSS only allows for three discrete values. My solution would be to recode all positive missing values to negative and declare missing values in the range **(LO thru -1)** or **(LO thru 0)** . As well as "Don't know", "Refused" some variables have codes for 'Other' 'Uncodeable' 'Can't decide' etc which can/should be treated as missing.

SPSS only allows three discrete missing values to be entered, but In the example below, for variable **REarnQ**, there are four values to be treated as missing.



This is not a problem since two of the allowed values can be used to indicate the lower and upper limits of a range of values, eg:

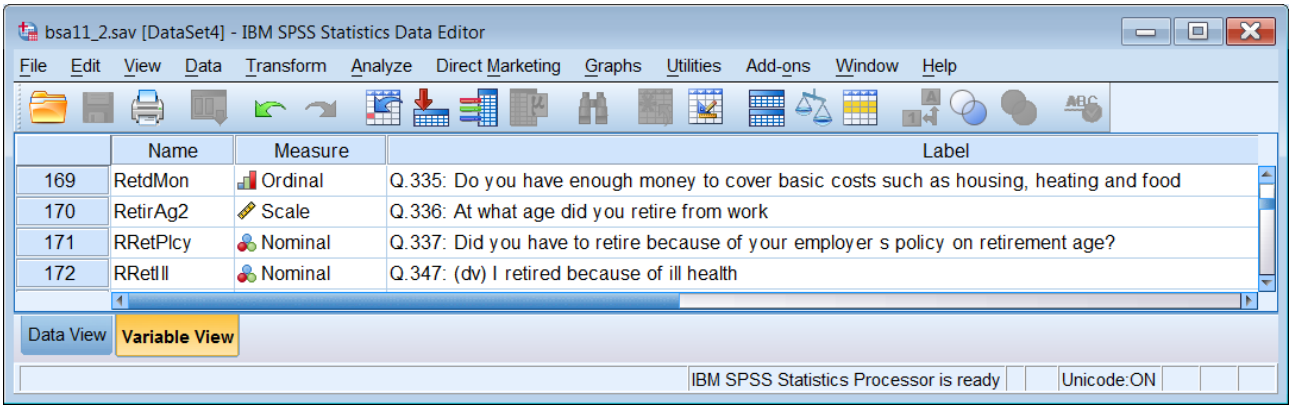
missing values rearnq (-1, 7 thru 9).

Checking variable properties

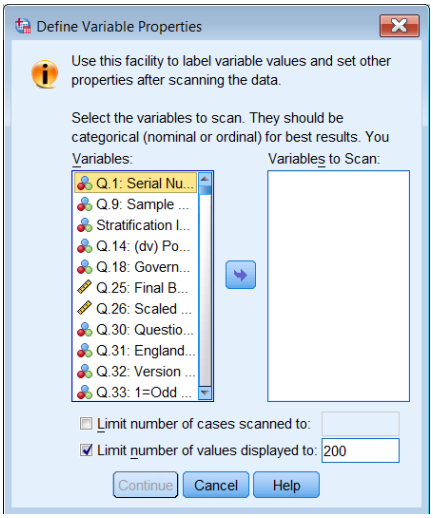
A very useful SPSS tool (only available as a drop-down menu from the GUI) is:

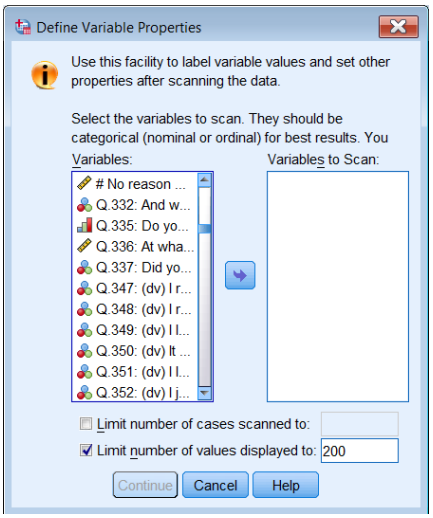
Data >> Define Variable Properties

It is there for those who prefer GUI over syntax (or have never heard of syntax) to use for entering properties or importing/exporting them from/to other variables, but I only ever use it to check complete SPSS *.sav files once I have already created them using syntax (or, in this case, downloaded them from elsewhere). It is particularly useful for checking missing values and labels. It displays either variable **names** or variable **labels** in either alphabetical or questionnaire order, or measurement **level** in alphabetical order only.

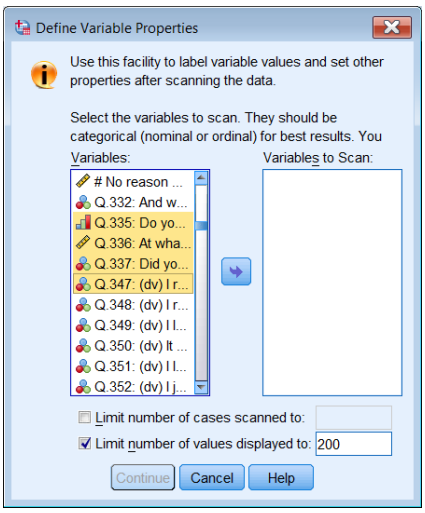


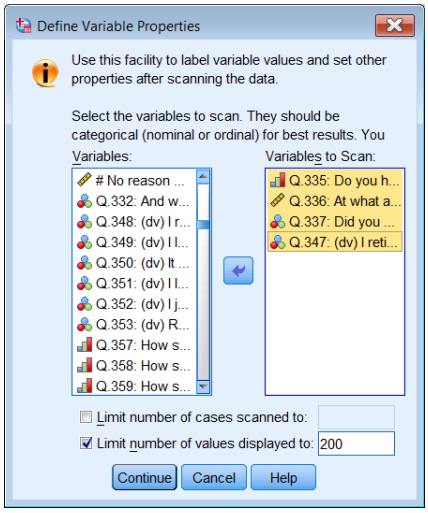
The examples below display variable **labels** in **questionnaire order** and show the clear advantage of having question numbers at the beginning of the labels. The display always opens at the top of the SPSS file, but you can scroll down to find the variables you want:



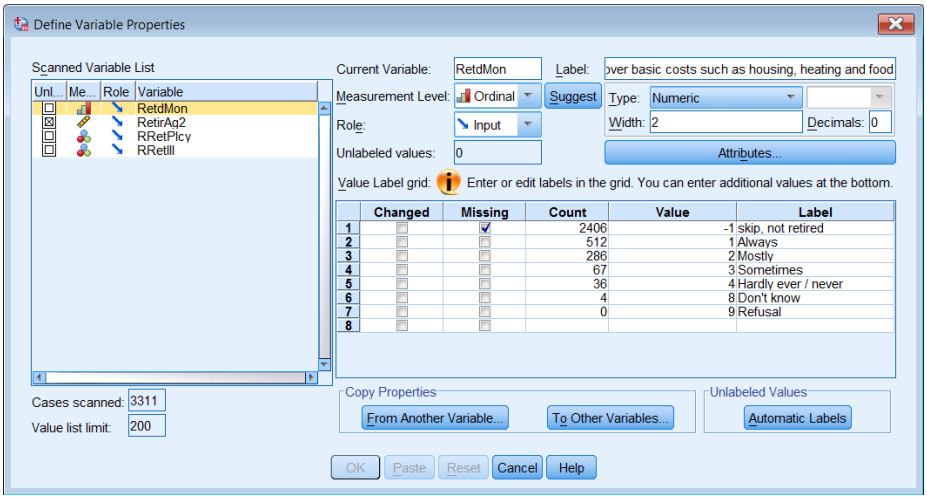


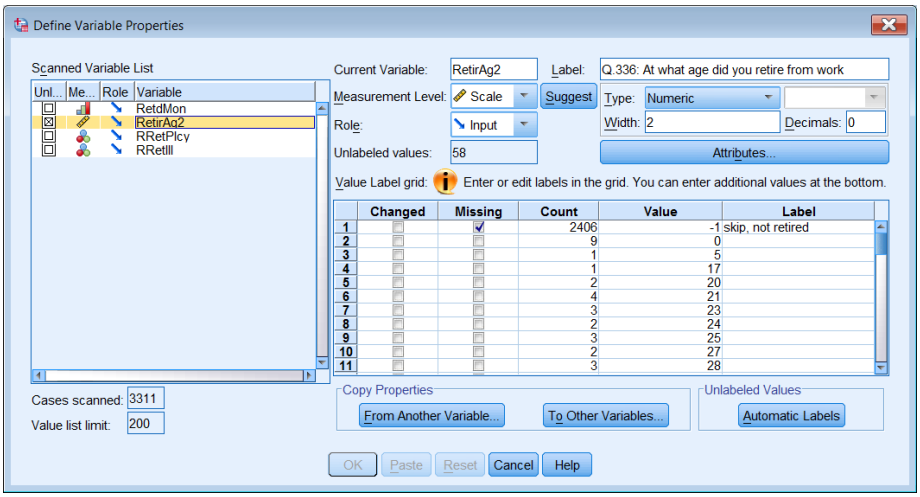
By highlighting the variables of interest and clicking on the blue arrow to transfer them from the Variables box across to the Variables to scan box:





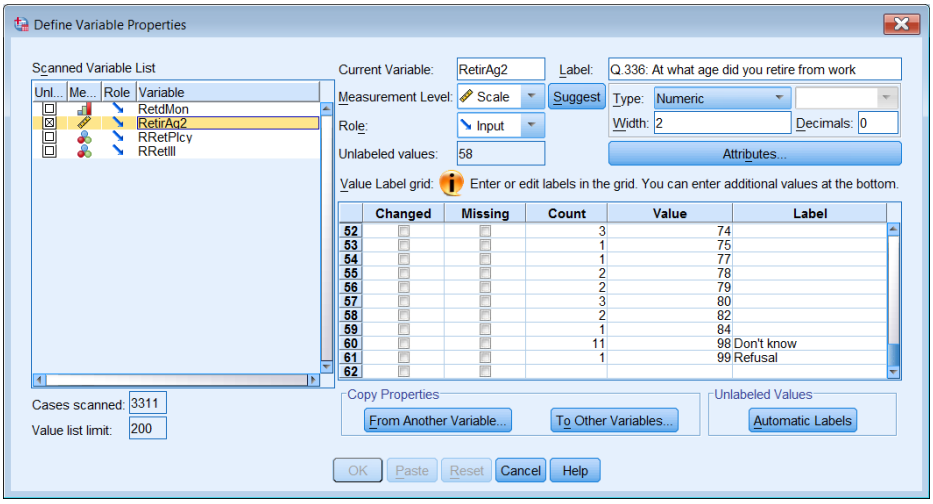
Clicking on Continue opens new window in which you can easily check the properties of any variable by clicking on it in the left pane:





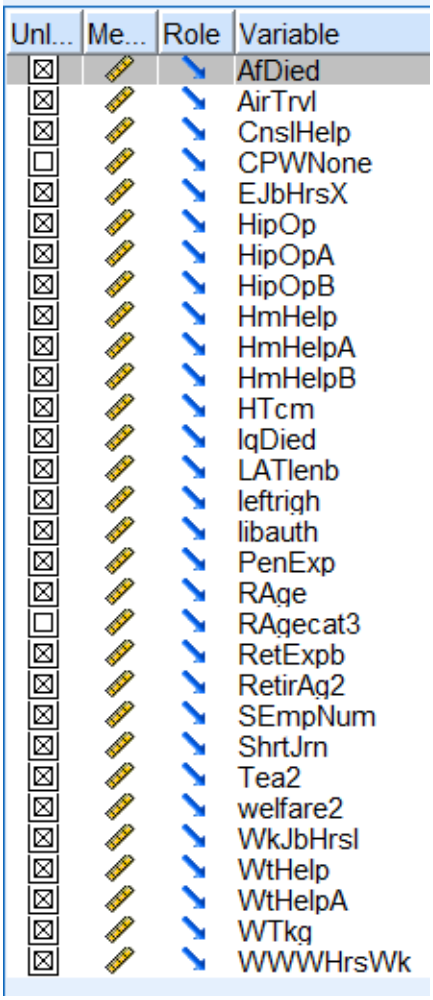
In the displays above, variables **RetdMon** and **RetirAg2** both have **-1** declared as missing. **RetdMon** has values **8** and **9** clearly labelled as "Don't know" and "Refusal" but the boxes are not checked under **Missing**.

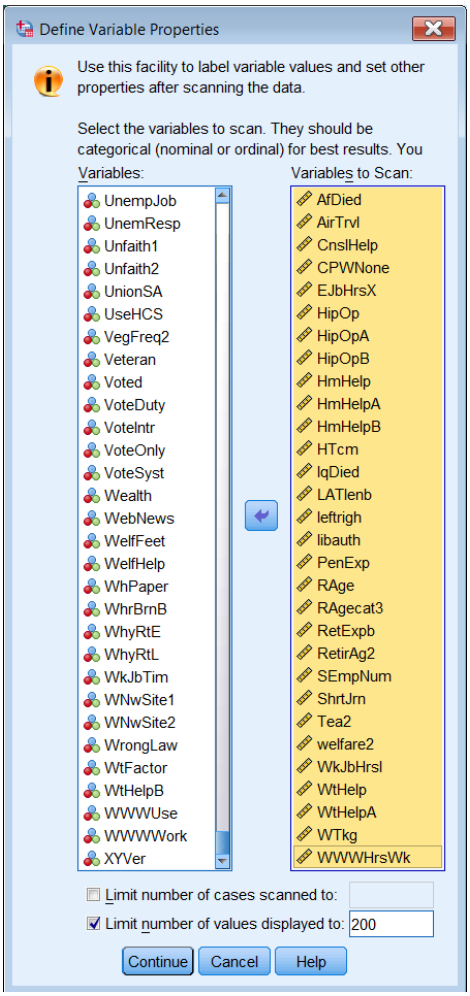
Sometimes you need to scroll down the right pane to see the maximum values used (and their value labels, if any):

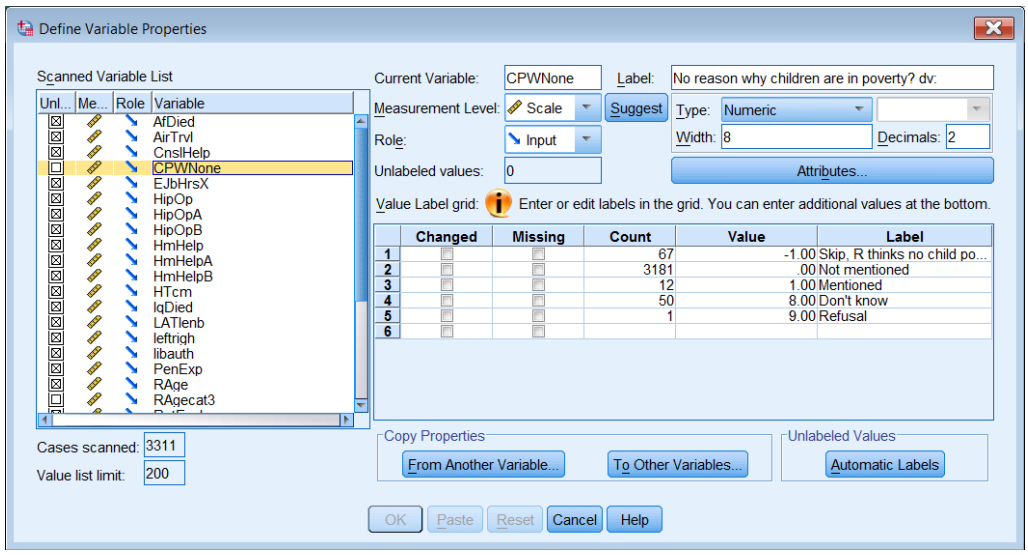


RetirAg2 has values 98 and 99 clearly labelled as "Don't know" and "Refusal" but again the boxes are not checked under **Missing**. OIt's perfectly possible to work your way through the file checking the relevant boxes under **Missing** but that doesn't save any syntax for later use and anyway it's quicker to use **MISSING VALUES** in syntax.

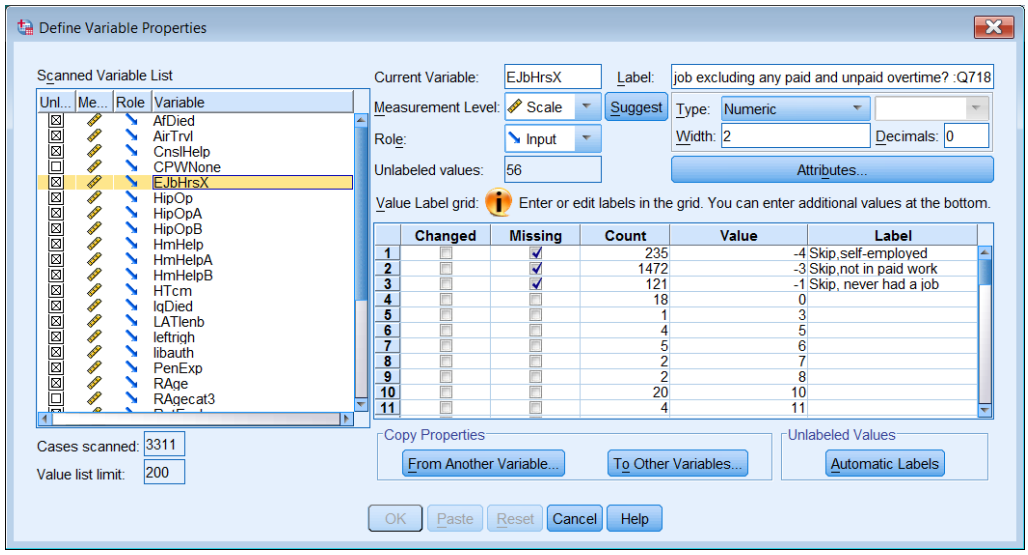
The order of **Scale** variables below is in **alphabetical** order of variable **name**, not in questionnaire order.



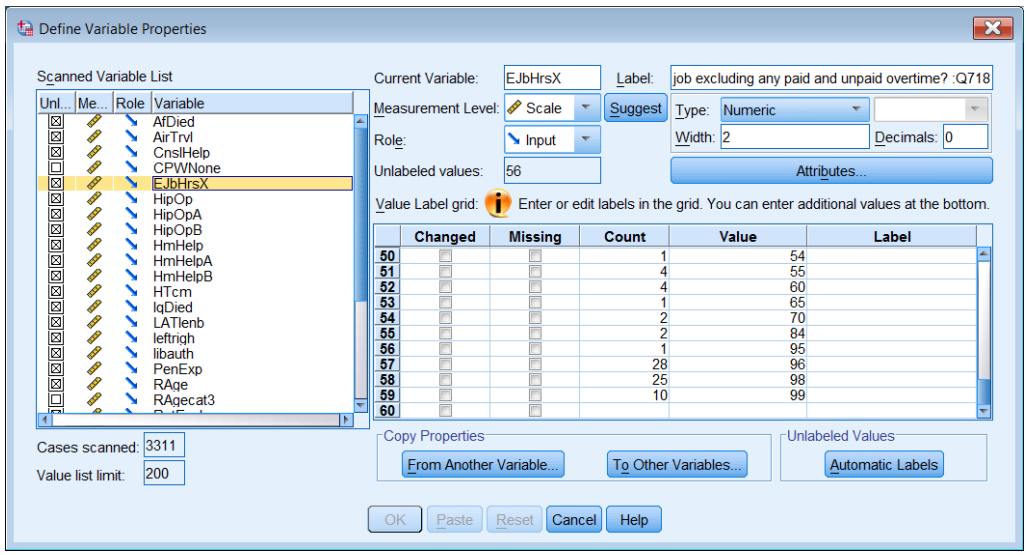




In the display above, variable **CPWNone** has no missing values declared at all.



Variable **EjbHrsX** has -4, -3 and -2 declared as missing, but not 98 or 99: 96 may be a maximum or missing and will have to be checked against the questionnaire. (It's actually "Varies too much to say")



Here are some tables showing the problem with the way missing values are coded in the SPSS file for the BSA 2011 survey (three for **remplyee**, four for **remploye** and five for **esrjbtim**).

The following SPSS output has been edited to show in **red** values which are already declared as missing and in **pink** values which are not, or should be.

REmplyee In your (main) job (are you/ were you/ will you be) employee or self-employed :Q698

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-1 skip, never had a job	121	3.7	3.7	3.7
	1 ... an employee,	2783	84.1	84.1	87.7
	2 or self-employed?	382	11.5	11.5	99.2
	8 Don't know	8	.2	.2	99.5
	9 Refusal	17	.5	.5	100.0
	Total	3311	100.0	100.0	

REmploye Respondent currently employee or self-employed dv :Q699

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Employee	1473	44.5	85.7	85.7
	2 self-employed	235	7.1	13.7	99.4
	8 Don't know	2	.1	.1	99.5
	9 Refusal	8	.2	.5	100.0
	Total	1718	51.9	100.0	
Missing	-3 Skip,not in paid work	1472	44.5		
	-1 Skip,never had a job	121	3.7		
	Total	1593	48.1		
Total		3311	100.0		

ESrJbTim In present job, working full-time[if emp]. DV :Q715

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Full-time,	1111	33.6	74.2	74.2
	2 Part-time?	361	10.9	24.1	98.3
	8 Don't know	1	.0	.1	98.3
	9 Refusal	25	.8	1.7	100.0
	Total	1498	45.2	100.0	
Missing	-4 Skip,self-employed	382	11.5		
	-3 Skip,not in paid work	1310	39.6		
	-1 Skip, never had a job	121	3.7		
	Total	1813	54.8		
Total		3311	100.0		

It would be so much easier if they were all coded as negative numbers and missing values declared as a range **(LOWEST thru -1)**. Some variables use **7** and/or **0** as missing, so recoding **7** to **-7** and using **(LO thru 0)** would work for these. Note that **-1** is not declared as missing for **remplyee**, but negative values are declared as missing for the other two. This is inconsistent..

To get round this problem, I recoded all positive missing values to negative and declared missing values as a range[3] using **(-99 thru -1)**.

[3]
```
missing values
        rearnq rearnd remplyee remploye esrjbtim ExPrtFul  EJbHrCax rNSocCl rClassGp rNSSECG
          (-99 thru -1).
freq rearnq rearnd esrjbtim.
```

Thus the table for **ESrJbTim** above becomes:

ESrJbTim In present job, working full-time[if emp]. DV :Q715

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Full-time,	1111	33.6	75.5	75.5
	2 Part-time?	361	10.9	24.5	100.0
	Total	1472	44.5	100.0	
Missing	-9 Refusal	25	.8		
	-8 Don't know	1	.0		
	-4 Self-employed	382	11.5		
	-3 Not in paid work	1310	39.6		
	-1 Never had job	121	3.7		
	Total	1839	55.5		
Total		3311	100.0		

The table for the derived variable **REarnD**

REarnD Q.1208: (dv) Respondent pre-tax earnings deciles (

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Less than £430 p.m.	131	4.0	7.6	7.6
	2 £431 - 780 p.m.	151	4.6	8.8	16.4
	3 £781 - 1,100 p.m.	162	4.9	9.4	25.8
	4 £1,101 - 1,300 p.m.	177	5.3	10.3	36.1
	5 £1,301 - 1,600 p.m.	175	5.3	10.2	46.3
	6 £1,601 - 1,800 p.m.	151	4.6	8.8	55.1
	7 £1,801 - 2,200 p.m.	143	4.3	8.3	63.4
	8 £2,201 - 2,700 p.m.	132	4.0	7.7	71.1
	9 £2,701 - 3,600 p.m.	120	3.6	7.0	78.1
	10 £3,601 or more p.m.	142	4.3	8.3	86.4
	97 Refused information	200	6.0	11.6	98.0
	98 Don't know	33	1.0	1.9	99.9
	99 Refusal	1	.0	.1	100.0
	Total	1718	51.9	100.0	
Missing	-1 skip, not currently working	1593	48.1		
Total		3311	100.0		

. . becomes:

REarnD Respondent pre-tax earnings deciles (dv) :Q1208

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Less than £430 p.m.	131	4.0	8.8	8.8
	2 £431 - 780 p.m.	151	4.6	10.2	19.0
	3 £781 - 1,100 p.m.	162	4.9	10.9	29.9
	4 £1,101 - 1,300 p.m.	177	5.3	11.9	41.8
	5 £1,301 - 1,600 p.m.	175	5.3	11.8	53.6
	6 £1,601 - 1,800 p.m.	151	4.6	10.2	63.8
	7 £1,801 - 2,200 p.m.	143	4.3	9.6	73.5
	8 £2,201 - 2,700 p.m.	132	4.0	8.9	82.3
	9 £2,701 - 3,600 p.m.	120	3.6	8.1	90.4
	10 £3,601 or more p.m.	142	4.3	9.6	100.0
	Total	1484	44.8	100.0	
Missing	-99 Refusal	1	.0		
	-98 Don't know	33	1.0		
	-97 Refused information	200	6.0		
	-1 Skip	1593	48.1		
	Total	1827	55.2		
Total		3311	100.0		

Unique combinations of missing values

Missing values seem to have been declared only for items which were skipped or not applicable, usually as negative numbers, but the logic is not always apparent and may be inconsistent. When other values normally treated as missing are taken into account the range of possible combinations is bewildering to say the least. These are the 32 combinations I've found so far.

Declared	+ Not declared	Notes
(-1,	20)	[20 = inadequately described]
(-1,	8)	[? 7 = Armed forces, 8 = inadequately described]
(-1,	8)	[8 = Not classified/classifiable]
(-1,	8, 9)	
(-1,	9)	[average, not score: format f5.4!!]
(-1,	16, 17)	[16 = inadequately described, 17 = Unclassifiable]
(-1,	96, 97,98, 99)	
(-1,	97 98 99)	
(-1,	98, 99)	
(-1,	998, 999)	
(-1,	99998 99999)	
(-2,	8, 9)	
(-2,	7, 8, 9)	
(-2,	98, 99)	
(-2	8 99)	
(-2,	6, 8 9)	[? spontaneous 6] ccacar
(-2,	8, 9)	
(-2 -1	8, 9)	
(-2 -1	97 99)	
(-2, -1,	98, 99)	
(-2 -1	998 999)	[format f4 .2!]
(-2, -1,	8, 9, 6)	[6 = Rarely or never use]
(-2, -1,	9)	
(-3,-1,	8, 9)	
(-3 -1	98 99)	
(-3, -1,	8)	
(-3, -2, -1,	8, 9)	
(-4 -3 -1	)	[?combined with 95 96 97 98 99?]
(-6, -4,	8 9)	
None	(8)	[8 = Not classified/classifiable; format f2.1]
None	(97 98 99)	[? 95 96 ? plus someone with TEA aged 1!]
None	(998 999)	

Missing values only partially declared

wwwhrswk penexp
(-1, 998, 999)
idstrng drivmil travel1 carppub to carpnoch sempnum rempwkfw
(-3,-1, 8, 9)
politics soctrust NHSSat to NHS5yrs govtwork to mpstrust
(-2, 8, 9)

ukspngbe ruhappy2 to choicedr drinkfr to ltsgnhth hlthgap pinallow to mobdlaw
ctaxref2 to chcandte
(-2,-1, 8, 9)
hlncdif4
(-2, 7, 8, 9)
cpwsocbn to cpwnone retdmon afaccept to iqsupmem rempwrk2 penknow2 rearnq
migworps to mediaex
(-1, 8, 9)
cpvwhym
(-1, 96 thru 99)
retirag2 latenb
(-1, 98, 99)
hipop
(-2, 98, 99)
hipopa to hmhelpb [no variable labels]
brhssat to smokday2 relmarfc relmarmu
(-2, -1, 98, 99)
rretill to rretoth
(-6, -4, 8, 9) afdied iqdied
(-1, 99998 99999) airtrvl
(998, 999)
shrtjrn
(-2 -1 97 99)
cnshelp
(98, 99)
ejbhrsx
(-4 -3 -1 98 99) [? 95 = 95 or more; 96 = varies]
htcm wtkg
(-2, -1, 998, 999) [format f4 .2!]
leftrigh libauth welfare2
(-1, 9) [average, not score: format f5.4!!]
retexpb
(-1, 998, 999)
tea2
(97, 98, 99) [? 95 96 ? plus someoned aged 1!]
wrkjbhrs1
(-3, -1, 98, 99)
wthelp
(-2, 98, 99)
wthelpa
(-2, -1, 98, 99)
rnseg
(-1, 20) [20 = inadequately described]
rnseggrp rnsoccl
(-1, 8) [7 = Armed forces, 8 = inadequately described]
ropcat
(-1, 16, 17) [16 = inadequately described, 17 = Unclassifiable]
rclass rclassgp
(-1, 8) [8 = Not classified/classifiable]
rnssecg s2nsseg
(8) [8 = Not classified/classifiable; format f2.1]
losetch to govcomp
(-2, 6, 8, 9) [? spontaneous 6]
trstparl to srprej
(-2, 8, 9)
s2class s2classg
(-3, -1, 8)

rearnd
(-1, 97, 98, 99)
privmcov
(-3, -2, -1, 8, 9)
bnlowp to falcatch
(-1, 8, 9)
rthdsw2 to rthdsprd welfhelp to censor
(-1, 9)
carwalk2 to carbike2
(-2, -1, 8, 9, 6) [6 = Rarely or never use]
freqelec
(-2, -1, 9)

Missing values not declared at all

rage mil10yrs to migroup3 asgdbad (98 99)
RAgeCat (8)
RAgeCat2 RAgecat3 (9)
tvnews webnews (98,99)
socspend1 to socspend6 TRFPB6U to TrfConc3 travel2 to travel6 clicar cliplane
causdfor to causabrd asastay asapplic defspend escompnd escompdi escompho hhincq (8,9)
opaf (8 9) [? spontaneous 6]
afopchg (8 9) [reorder?]
taxspend (4, 8, 9)
ubprobs ubreq cpr2gov to cpr2nocp cycconf cycdang letin (8,9)
hhincd (97, 98, 99)

Oddities

One or two dubious codes to check on TEA2 (eg: Completed full time education at age 1 !!)

Dichotomies

Binary for Mult Response (File position)

As well as variables with Yes/No answers there are variables with [Mentioned / Not mentioned] These can be treated as dichotomies or recoded as 1 to n for use with **MULT RESPONSE**:

Name	**Position**
CPR2Gov to CPR2NoCP	144 - 150
CPWSocBn to CPWNone	151 – 167
RRetIll to RRetOth	172 - 178
CarPPub to CarPNoCh	227-231
CausDfor to CausAbrd	234-247
FInvTp1 to Oexpi15	476 – 503
DoneMP to DoneNone	524-532
BAbrNone to BAbrFP	547-552
brnInd to brnOth	553-566
NatBrit to NatNone	579- 590
EdQual1 to EdQual39	604-633
BenefOAP to BenFNone	659 – 694