Block 1 - From questionnaire to data file

1.2.4: Second look at data from a major survey

[Updated 18 March 2020 from September 2010 original]

Exemplar: British Social Attitudes (1989) 1

The raw data for the 1989 British Social Attitudes survey were initially supplied by the UK Data Archive in a *.dat file which, on my previous computer, Windows interpreted as a Wordperfect file. Wordperfect displays the data in **Times New Roman proportional** font which is not only horrible to look at, but also, because the columns are not vertically aligned, is impossible to interpret visually.

```
238290109012829191203
                         511201 2 0201150450201
2382902220381511042210 311 1111121221 13113112112121132 13343132 031 3943122
23829034
               1006 33312 2 42 090106122 222 2112114410390903
2382904
        152311151
                       3
2382905
2382906
                       030114113331142122 113112 232344
2382907
2382908010407 11131210
                         1212112312 231211412111111
2382909
2382910
2382911412 4
                  2 12 4444 3 3
23829128 8 3831
                         2
                                          14122
                  1
2382913 1330342 1 1 1222 2212211211 226
                                                     222227
2382914102251 15911
                                           1
                                       069010907329604
2382915 22
            21 2
                    12
2382916 000010341 22222203 071030808327108 00061021111
                 130611112 221221230459191250789
23829172211 4
2382918
2382919432221423223 1
                      43122312223 21 3 2122111 2223 1
                                                          333422224
2382920223535525211
                                      32221122221232211511212
238292113413331223331
                                        34223222222222333
2382922
                             335333
                                        234444441311
238292323221 22322333323 225078921241064115
                                                0102654424341008373215
```

Wordperfect was not installed on my computer and Windows thought the *.dat file was a movie!!

Although SPSS can read the data in this format, provided the file name is enclosed in primes (apostrophes) i.e. **'myclass.dat'**, it is always useful to be able to inspect the original data visually.

For our purposes the data need to be displayed in a **fixed-width** font to make it easier to find our way around inside the file. On my previous machine I could change the font to **Courier New** inside Wordperfect: on this machine I had to open the file with Word and convert it.

You don't have to do this yourself, but you may find it useful to follow the steps in case you have to do something similar in future. In any case we shall be using these data in later exercises and there's a little test at the end to see if you've understood how the data relate to the questionnaire.

British Social Attitudes has its own site. Because of the General Data Protection Regulation (GDPR) the raw data file bsa89.txt for this exercise can no longer be downloaded from my site. Data files can now only be accessed via the UK Data Archive at Essex University: http://www.data-archive.ac.uk/findingData/snDescription.asp?sn=2723

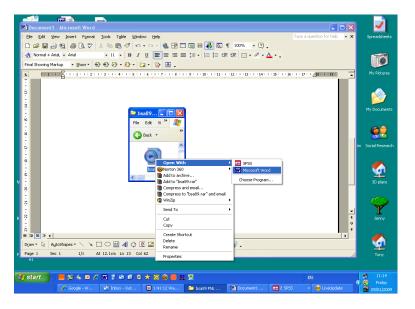
[NB: Unless you are registered with UKDS and authorised to use BSAS, you will not be able to perform the following steps. However, it is useful to be able to follow the steps I had to take.]

The original data file was supplied as a WordPerfect file **GB8901.DAT**, but I renamed it **bsa89.dat** and then converted it from **Times New Roman proportional** to **Courier New fixed-width** font in *.txt format in file **bsa89.txt** ²

Windows 2000 (2010)

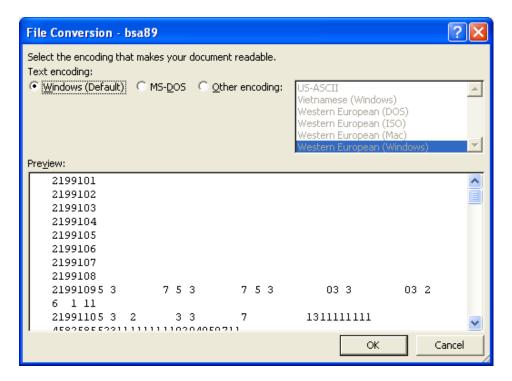


Right click the icon, click on Open With:

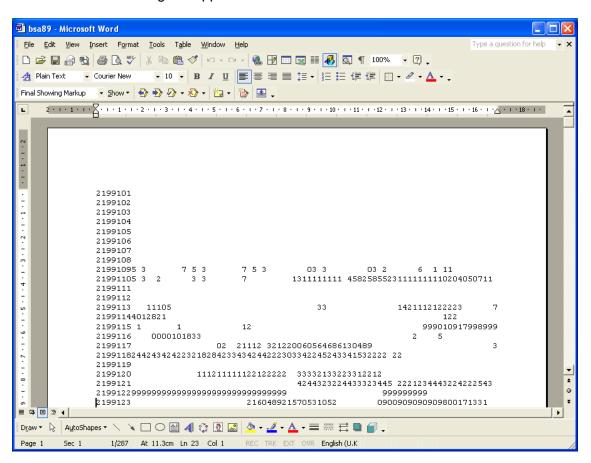


... then on Microsoft Word, which displays:

British Social Attitudes has its own site. Because of the General Data Protection Regulation (GDPR) the raw data file bsa86.txt for this exercise can no longer be downloaded from my site. Data files can now only be accessed via the UK Data Archive at Essex University http://www.data-archive.ac.uk/findingData/snDescription.asp?sn=2315.



Press **OK** and following will appear:



This is a conversion of the original data from **Times New Roman variable-width** font to plain text in **Courier New fixed-width** font. The data are displayed in 80-column lines in which all 80 columns are vertically aligned. The 1989 British Social Attitudes survey has 3024 cases and 23 lines (records) per case (that's a lot of lines!).

The first case is displayed in the screenshot above.

Windows 10 Pro (2020)



2199102 2199103 2199104 2199105 2199106 2199107 2199108 21991095						
2199101 2199102 2199103 2199104 2199106 2199106 2199107 2199108 21991095	bsa89 - Notepad					×
2199102 2199103 2199104 2199105 2199106 2199107 2199108 21991095 3	<u>File Edit Format View Help</u>					
2199103 2199104 2199105 2199106 2199107 2199108 21991095	2199101					^
2199104 2199105 2199107 2199108 21991095 3	2199102					
2199105 2199106 2199107 2199108 21991095 3	2199103					
2199106 2199107 2199108 21991095 3	2199104					
2199107 2199108 21991095 3	2199105					
2199108 21991095 3	2199106					
21991095 3	2199107					
21991105 3 2 3 3 7 1311111111 4582585523111111111102040507 2199111 2199112 2199113 11105 33 1421112122223 21991144012821 122 2199115 1 1 12 999010917998 2199116 0000101833 2 5 2199117 02 21112 321220060564686130489 2199118244243424223218284233434244222303342245243341532222 22 2199119 2199120 11121111112212222 33332133223312212 2199121 4244323224433323445 22212344432242255 21991299123 216048921570531052 0900909090909080017133 207220102002512050103 511201 2 0202150600301 1 61 00.666 2072202119150012 1 013311 1114221314 43111222221111121213342232 031 40433 2072204 3 1555555555102	2199108					
2199111 2199112 2199113 11105 33 1421112122223 21991144012821 122 2199115 1 1 12 999010917998 2199116 0000101833 2 5 2199117 02 21112 321220060564686130489 2199118244243424223218284233434244222303342245243341532222 22 2199119 2199120 111211111112212222 33332133223312212 2199121 4244323224433323445 22212344432242225 2199123 216048921570531052 0900909090909080017133 207220102002512050103 511201 2 0202150600301 1 61 00.666 2072202119150012 1 013311 1114221314 43111222221111121213342232 031 40433 2072204 3 1555555555102	21991095 3 7 5 3	7 5 3	03 3	03 2	6 1 11	
2199112 2199113 11105	21991105 3 2 3 3	7	1311111111	4582585523111	111111020	40507
2199113 11105 33 1421112122223 21991144012821 122 2199115 1 1 12 999010917998 2199116 0000101833 2 5 2199117 02 21112 321220060564686130489 2199118244243424223218284233434244222303342245243341532222 22 2199119 2199120 111211111122122222 33332133223312212 2199121 4244323224433323445 22212344432242225 2199121 4244323224433323445 22212344432242225 2199123 216048921570531052 0900909090909080017133 207220102002512050103 511201 2 0202150600301 1 61 00.666 2072202119150012 1 013311 1114221314 43111222221111121213342232 031 40433 20722034 1006 3331106 102 2 100508123 2 23 22442 2072204 3 1555555555102	2199111					
21991144012821 122 2199115 1 1 12 999010917998 2199116 0000101833 2 5 2199117 02 21112 321220060564686130489 2199118244243424223218284233434244222303342245243341532222 22 2199119 2199120 111211111122122222 33332133223312212 2199121 4244323224433323445 22212344432242225 21991229999999999999999999999999999	2199112					
2199115 1 1 1 2 999010917998 2199116 0000101833 2 5 2199117 02 21112 321220060564686130489 2199118244243424223218284233434244222303342245243341532222 22 2199119 2199120 111211111122122222 33332133223312212 2199121 4244323224433323445 22212344432242225 21991229999999999999999999999999999	2199113 11105		33	142	111212222	3
2199116 0000101833 2 5 2199117 02 21112 321220060564686130489 219911824424342423218284233434244222303342245243341532222 22 2199119 2199120 111211111122122222 33332133223312212 2199121 4244323224433323445 22212344432242225 21991229999999999999999999999999999	21991144012821				122	
2199117 02 21112 321220060564686130489 2199118244243424223218284233434244222303342245243341532222 22 2199119 2199120 111211111122122222 33332133223312212 2199121 4244323224433323445 22212344432242225 21991229999999999999999999999999999	2199115 1 1	12			9990109	17998
219911824424342423218284233434244222303342245243341532222 22 2199119 2199120	2199116 0000101833				2 5	
2199120	2199117 03	2 21112 3212	20060564686	130489		
2199120	2199118244243424223218284	2334342442223	03342245243	341532222 22		
2199121	2199119					
21991229999999999999999999999999999999	2199120 11121:	1111122122222	333321332	23312212		
2199123	2199121		424432322	4433323445 222	123444322	42225
207220102002512050103 511201 2 0202150600301 1 61 00.666 2072202119150012 1 013311 1114221314 43111222221111121213342232 031 40433 20722034 1006 3331106 102 2 100508123 2 23 22442 2072204 3 1555555555102 2072205 2072206	21991229999999999999999999	999999999999		9999999	999	
2072202119150012	2199123	21604892	1570531052	0900909	990909800	17133
20722034 1006 3331106 102 2 100508123 2 23 22442 2072204 3 1555555555102 2072205 2072206	207220102002512050103	511201 2	020215060	0301 1	61 0	0.666
2072204 3 155555555102 2072205 2072206	2072202119150012 1 01331:	1 1114221314	431112222	21111121213342	232 031	40433
2072205 2072206 <	20722034 1000	5 3331106 1	2 1005	08123 2 23	22442	
2072206	2072204 3 1555	555555102				
<	2072205					
	2072206					~
Ln 1, Col 1 100% Windows (CRLF) UTF-8	<					>
		Ln 1, Col 1	100%	Windows (CRLF)	UTF-8	

The actual data for the first case are shown below:

```
2199101
2199102
2199103
2199104
2199105
2199106
2199107
2199108
21991095 3 7 5 3
                   753 033
                                   03 2 6 1 11
21991105 3 2 3 3 7
                        1311111111 45825855231111111110204050711
2199111
2199112
                                    1421112122223
2199113 11105
                           33
21991144012821
                                        122
                                     999010917998999
2199115 1
                  12
2199116 0000101833
                                       2 5
               02 21112 321220060564686130489
2199118244243424223218284233434244222303342245243341532222 22
2199119
2199120
             111211111122122222 33332133223312212
2199121
                       4244323224433323445 2221234443224222543
2199122999999999999999999999999999
                                              99999999
2199123
                  216048921570531052
                                       0900909090909800171331
```

Second exercise on a major survey

British Social Attitudes 1989 - Version B, questionnaire page 44

Below is a facsimile of part of the questionnaire dealing with household information:

	eboli			•	44 -						Col./ Code	Skip to
900a)	Can I just chec	k your own	ı marit							90 mel 58 1	1408	
	Can I just check your own marital status? At present are you READ OUT married,							ed,	101233			
140 F	CODE FIRST TO A	APPLY					liv	ing as	marri	ed,	2	
						S	eparat	ed or	divorc	ed,	3	
									widow	ed,	4	
							or	- not	marri	ed?	5	
	Including yours as members of t CHECK INTERVIEW	his househ	old?			SC GMA	01.589	1 2300		HTAV S		
81.74	OF HOUSEHOLD IF	NECESSARY	1916 1917 1918	REA _	(90P. 0	ONDE G	WRIT	M CA W	OV CRA] 1	+09-10	
901.		necessary ask for a	few d	etail:	s about	t each what wa	person	n in	OV CRA	14	+09-10	
901.	OF HOUSEHOLD IF Now I'd like to your household.	ask for a	few d	etails yours	s about	what wa	person	n in	DV GRA (D G S V S S R A B A B A B A B A B A B A B A B A B A B A B A B A B A B A B A B B		+09-10 	
901.	Now I'd like to your household. last birthday? WORK DOWN COLUM MEMBER.	ask for a	few dg with	etails yours	s about	what wa	person	n in	G GARAGE AND	114	+09-10 10	
901. a)	Now I'd like to your household. last birthday?	ask for a Startin	few dg with	etails yours	s about	what wa	person	n in	8 45		BARR C BRARL DOA: (JJG RSB RSB RSB RSB RSB RSB RSB RSB RSB RSB	
	Now I'd like to your household. last birthday? WORK DOWN COLUM MEMBER.	ask for a Startin	few dg with	etails yours ACH HO	s about self, t	what wa	person	n in age		9	2222 9	
	Now I'd like to your household. last birthday? WORK DOWN COLUM MEMBER.	ask for a Startin NS OF GRID Respondent	few dg with	etails yours ACH HO	s about self, to DUSEHOL	what was	person as your	n in cage	45	9 50	10	
	Now I'd like to your household. last birthday? WORK DOWN COLUM MEMBER. Sex:	ask for a Startin NS OF GRID Respondent 11 1e 1 1e 2	few dg with FOR E	etails yours ACH HO 3 20 1 2	s about self, v	5 30 1	person s your 6 6 35 1 2	7 40 1 2	45 1	9 50 1	10 55 1 2	

Question 900a (Marital status of respondent) is **precoded** and the interviewer will circle the appropriate code (1 to 5) in the box numbered 1408 in the right hand column. For question 900b (Number of people in the household) the interviewer will write the number (right justified or with leading zero) in the pair of boxes numbered 1409 - 10.

Question 901 is a typical **grid** for collecting information about members of the respondent's household. The layout is slightly more complex, but quite straightforward. Question 901a (Sex of respondent) is precoded and again the interviewer will circle the appropriate code (1 for Male, 2 for Female). For question 901b (Age of respondent last birthday) the interviewer will write in the age in the first pair of boxes on the line (numbered 12 - 13).

The grid continues (not shown) with questions on the relationship to the respondent of each additional person in the household and whether that person has legal responsibility for the accommodation. For now, this doesn't concern us, but if there are any other people in the household, the interviewer will, for each additional person, circle the code for sex and write in the

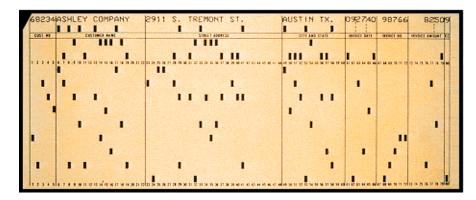
age in the appropriate pair(s) of boxes, starting with the second person and coding sex in the box numbered 15 (at the top) and age in the boxes numbered 16-17. How do the responses to these questions get into the computer?

The data from this survey were originally keyed in as fixed format 80-column lines (card images), using the data layout template printed in the right-hand margins of the questionnaire. This format derives from the days before visual display units (VDUs) and keyboards, when all data (and also programs) had to be punched on 80-column Hollerith cards and fed into the computer via a card reader.

80-column Hollerith card

From Computer Desktop Encyclopedia

© 2000 The Computer Language Co. Inc.

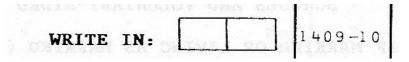


Have another look at the questionnaire.

In the top right-hand corner, for question 900a (Marital status), you can see the number **1408** in smaller type in the margin and above the response codes 1 - 5.

	Col./ Code
ma people you've jbar sam:	1408
married,	1201
living as married,	2
separated or divorced,	3
widowed,	4
or - not married?	5 -5

...and further down, for question 900b a pair of boxes for the interviewer to write in the number of people in the household, next to **1409 -10** (in the margin).



These are special numbers used at the data preparation stage. In the 1970s and 1980s, before computer-assisted personal interviews (CAPI) data from completed questionnaires were usually punched on to 80-column Hollerith cards (sometimes 60- or 120- columns), which were then fed into the computer via a card reader.

The data could then be printed out on a line printer. Later developments enabled data to be keyed directly into the computer as fixed-format card-images and displayed on a computer screen, one line for each card. In other words, the data for each response code were always punched in the same column(s) on the same card for every respondent.

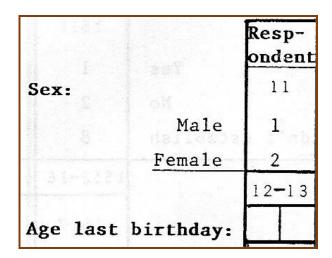
The 1989 British Social Attitudes survey consisted of two versions, A and B, with separate questionnaires and a separate self-completion questionnaire for each version. Each version carried a core of common questions plus a different a set of questions on key topics which were split between versions. Altogether there were 3,024 respondents, of whom 1,508 were administered version A and 1,516 version B. Each questionnaire takes up 23 lines of data and each line is numbered from 01 to 23.

On the computer, when you open the raw data file, each record is displayed as a line on the screen. There are 23 lines of data for each respondent, but, apart from the serial number and record number, the lines for the data from versions A or B ³ will be blank depending on which version the respondent completed. In each of the 23 records for each of the 3024 cases (i.e. every record in the entire data set) columns 1-5 contain a unique serial number and columns 6-7 the record number (from 01 to 23).

In this example, the data for questions 900 and 901 were punched on card 14. We know this because the data layout convention used on the printed questionnaire is indicated by the four-digit number **1408** in small type at the top of the right hand margin (see above). In this convention the first pair of digits **14** indicate the card or line number (from 01 to 23) within each case and the second pair **08** the **field** (column(s)) within each card (from 01 to 80). Thus **1408** indicates that the data for Q.900a (Marital status) are to be punched on card 14 column 8: similarly data for Q.900b (Number in household) are on card 14 columns 9 and 10, as indicated by **1409-10** (two columns are needed for numbers greater than 9).

In SPSS terminology each card, card-image or data-line is now called a **record**.

Sex is coded as 1 or 2 and is to be punched as a single digit on record 14, column 11. **Age** is to be punched as a two-digit number on record 14, columns 12 - 13.



³ The code for which version was admiinistered is in column 8 of record 2 (viz, 208)

Here are the actual data ⁴ for a single case from a respondent who completed version B of the questionnaire.

```
238290109012829191203
                           511201 2 0201150450201
2382902220381511042210 311 1111121221 13113112112121132 13343132 031 3943122
                    1006 33312 2 42 090106122 222 2112114410390903
           152311151
2382904
2382905
2382906
2382907
                                      030114113331142122 113112
                                                                  232344
2382908010407 11131210
                           1212112312 231211412111111
2382909
2382910
2382911412 4
                         2 12 4444 3 3
                     1
23829128 8 3831
                                   2
                                                                 14122
                    1 1 1222 2212211211 226
2382913 1330342
                                                                     2222227
2382914102251 15911
             21 2
                                                             069010907329604
2382915 22
                           12
2382916 000010341 22222203
                                071030808327108
                                                  00061021111
                                                                         1
                                                                           3
23829172211
           4 130611112 221221230459191250789
2382918
                        43122312223 21
                                           3 2122111
                                                      2223 1
2382919432221423223 1
                                                                  333422224
                                                      32221122221232211511212
2382920223535525211
238292113413331223331
                                                         34223222222222333
                                                335333
2382922
                                         1
                                                              234444441311
238292323221 22322333323
                            225078921241064115
                                                    0102654424341008373215
```

This is a conversion of the original data from **Times New Roman variable-width** font to plain text in **Courier New fixed-width** font. The data are displayed in 80-column lines in which all 80 columns are vertically aligned.

Question: What are the sex and age of this respondent?

Try to answer this yourself before reading the next page.

Because of the General Data Protection Regulation (GDPR) the raw data file bsa89.txt for this exercise can no longer be downloaded from my site. Data files can now only be accessed via the UK Data Archive at Essex University: http://www.data-archive.ac.uk/findingData/snDescription.asp?sn=2723

The record number is in columns 6-7, highlighted in pink. Count downwards to find record 14 (highlighted in light turquoise) then count along to find 2 (sex) in column 11 and 51 (age) in columns 12-13.

```
511201 2 0201150450201
238290109012829191203
                                                              1
                                                                   61 00.6666
2382902220381511042210 311
                           1111121221 13113112112121132 13343132 031 3943122
23829034
                     1006 33312 2 42 090106122 222 2112114410390903
23829<mark>04</mark>
            152311151
                               3
2382905
2382906
2382907
                                        030114113331142122 113112
                                                                      232344
2382908010407 11131210
                             1212112312
                                         231211412111111
2382909
2382910
2382911412 4
                           2 12 4444 3 3
23829128 8 3831
                                                                     14122
2382913
         1330342
                      1 1 1222 2212211211 226
                                                                         222227
23829<mark>14</mark>102<mark>2</mark>51 15911
2382915 22
               21 2
                            12
                                                                 069010907329604
2382916 000010341 22222203
                                  071030808327108
                                                     00061021111
                                                                             1
23829172211 4
                       130611112 221221230459191250789
                                                                               3
2382918
2382919432221423223 1
                          43122312223
                                       21
                                              3 2122111
                                                         2223 1
                                                                      333422224
2382920223535525211
                                                         32221122221232211511212
238292113413331223331
                                                            34223222222222333
2382922
                                                   335333
                                                                 234444441311
                                            1
238292323221 22322333323
                             225078921241064115
                                                        0102654424341008373215
```

This respondent is therefore a **woman** aged **51**. What is her **marital status**? How many people live in the accommodation? Which version of the questionnaire did she complete? Find the answers yourself!!

Now go to section 1.3: Reading raw data into SPSS

Next step: 1.3.3.1 Preparing the ground

[Back to Block 1 menu]