GSS/GSR Disclosure Control Guidance for Tables Produced From Survey Data – Case Studies

Appendix B Case Studies – Tables produced from social surveys

Case Study B1. Disclosure issue: Linking tables by a single contributor

Table B1a has counts for specified ethnicities for a range of age groups. Table B1b is a request for further breakdown Chinese/Other ethnicity by income Publication of table B1b is likely to lead to self identification by the individual in the lowest Income Band., Other people in the sample from this ethnic group may also try to identify this low income individual. It may be necessary to suppress all the cells except the total in Table B1b. If the total is to be suppressed we may also need to suppress the internal cell in table B1a. The decision as to whether suppress that cell would depend on

- whether the number of sample respondents was also to be released, or could be very easily derived
- whether response knowledge (knowing a given individual is in the sample) is essential in order to find out the information
- the sensitivity of the information (risk of harm or distress to the individual)
- whether it would be possible to find out anything that the intruder would not have known, or been able to find out very easily from other sources. such as personal knowledge of their employment characteristics if for example they worked in a very 'visible' occupation.

Table B1a: Counts by Ethnicity and Age for males in a specified Local Authority. The number of sample respondents is given in brackets

	Age 16-24	Age 25-34	Age 35-55	Total
White	2,110	3,246	2,643	7,999
Mixed	124	356	..217	697
Asian / Asian British	213	267	164	644
Black / Black British	43	58	46	147
Chinese / Other	8	21	25	54

Table B1b: Income in the same Local Authority, age 16-24

	Income band Low	Income band Medium	Income band High	Total
Chinese/Other Aged 16 -24	1	3	4	8

This does raise an interesting and tricky issue of 1s in tables. If there is no attribute disclosure and self-identification is the only risk, we need to consider the sensitivity of the information. Would the individual feel exposed – and feel vulnerable to other people identifying them, or to being targeted by a journalist, for example. This is related to the Data Protection Act (1998) and considering the risk of harm or distress to the individual.

Case Study B2. Disclosure issue: Weighted percentages, unweighted base, cell size = 1, 2

Table B2a shows an example where it can be deduced with reasonable probability that the number of contributors to the cell percentage (the percentage of unemployed men aged 65+ who consulted a GP in the 14 days before the interview) is 1. This is likely although the weighting does introduce an amount of uncertainty. The decision to protect this information (effectively a cell count of 1) might hinge on the chance of someone being able to identify the individual (from response knowledge) and the sensitivity of the information. In this case, there are two variables that could be considered sensitive: the fact that the man is unemployed, and that they have consulted a GP. It is a difficult call, given the almost zero chance of being able to identify the person without already knowing all the information present, but it is not unreasonable to protect that cell in this case (there is also an issue with data quality – a percentage based on 5 people – but that is at least transparent to the user, though not in solution shown by Table B2c).

Table B2a: Percentage of unemployed men who consulted a GP in the 14 days before interview

	Age 16-44	Age 45-64	Age 65+	Total
Unemployed men	11	16	20	12
Unweighted base	*152*	*57*	*5*	*214*

The table should be protected by suppressing the unsafe cell and rounding bases to base 10 or collapsing categories as shown in table B2b.

Table B2b: Percentage of unemployed men who consulted a GP in the 14 days before interview

	Age 16-44	Age 45-64	Age 65+	Total
Unemployed men	11	16	c	12
Unweighted base (rounded)	*150*	*60*	*10*	*210*

Note that in this table the bases are no longer additive. Alternatively, instead of suppressing the percentage value, we could remove the unweighted bases as shown in table B2c. The percentage based on a small base should be italicised to indicate that.

Table B2c: Percentage of unemployed men who consulted a GP in the 14 days before interview

	Age 16-44	Age 45-64	Age 65+	Total
Unemployed men	11	16	*20*	12

Case Study B3. Disclosure issue: Cell size of one, count data

In the following example based the table has disclosed that there is only one household in the sample with six or more persons working in the sample. While this is not of much interest in itself, this information could be used in combination with other outputs to disclose further information about the household, in particular if this categorisation was cross-tabulated with any other variables (in this report it wasn't). This example shows how protection is required at the individual and the household level.

Table B3a: Characteristics of households based on weighted data (Unsuppressed and disclosive)

	Grossed number of house-holds (000s)	House-holds in sample (number)
Number of persons working in household		
No person	9,300	2,083
One person	7,090	1,606
Two persons	8,020	1,323
Three persons	1,370	371
Four persons	330	45
Five persons	30	14
Six or more persons	10	1

In this example it is unlikely that suppressing the row output for six or more persons would be sufficient since it is likely that the total number of (unweighted) households in the sample is known. If so, the suppressed figures could be easily calculated. Therefore, it would seem sensible to top code the number of persons working in the household, either to five or more persons or even, if some statistics are available elsewhere for five-person households, to four or more persons. Top coding to four or more persons removes almost all residual risk of disclosure by differencing.

Table B3b: Characteristics of households based on weighted data (Categories combined and non-disclosive)

	Grossed number of house-holds (000s)	House-holds in sample (number)
Number of persons working in household		
No person	9,300	2,083
One person	7,090	1,606
Two persons	8,020	1,323
Three persons	1,370	371
Four or more persons	370	60

Alternatively all base numbers could be rounded to base 10 provided the true figure cannot be calculated from the weighted number of households figure.

Table B3c: Characteristics of households based on weighted data (Base figures rounded and non-disclosive)

	Grossed number of house-holds (000s)	House-holds in sample (number)
Number of persons working in household		
No person	9,300	2,080
One person	7,090	1,600
Two persons	8,020	1,320
Three persons	11,370	370
Four persons	330	50
Five persons	30	10
Six or more persons	10	0

In the actual publication the table is published up to 4 or more persons with the sample numbers rounded to base 10.

Appendix C Case Studies – Tables produced from subsamples

Case Study C1

A typical table created from a sub sample of census data for a specified Region could be a 3 way table of Age * Sex * Ethnicity.
Table C1 shows a subset of the frequency table.

Table C1: Age * Sex * Ethnicity for a Region

	Sex = Male			
	Age = 55	**Age = 56**	**Age = 57**	**Age = 58**
Ethnicity				
White	55	48	35	34
Mixed	27	21	14	17
Asian / Asian British	23	13	6	7
Black / Black British	11	8	9	3
Chinese / Other	5	3	1	3

In this sample there is a just one male aged 57 with a Chinese / Other Ethnicity. This can be defined as an unsafe cell. Disclosure could occur through self-identification if this person sees himself in the table and also within group disclosure as this individual knows no other Chinese/Other male is aged 57. One method of protecting this individual is to ensure that if this table is released, ages are grouped into 10 year age ranges. For example counts for each ethnicity for each sex will be displayed as 15 and under, 16-24, 25,34, 35-44, 45-54, 55-64, 65-74, 75+.

Case Study C2

A table of Age * Sex * Industry for a specified region is another example of a 3 way table

Table C2: Age * Sex * Ethnicity for a Region

	Sex = Male			
	Age = 65	**Age = 66**	**Age = 67**	**Age = 68**
Industry				
Agriculture, Hunting, Forestry	4	1	2	0
Fishing	5	1	1	0
Mining Quarrying	6	2	4	0
Manufacturing	4	3	3	0
Electricity, gas Water Supply	7	4	2	0
Construction	9	6	3	0
Wholesale and retail trade; repair of motor vehicles	8	3	4	0
Hotels and Restaurants	3	7	3	0
Transport storage and communication	4	4	4	0
Financial Intermediation	2	5	6	0
Real Estate: Renting and Business activities	8	3	7	0
Public administration and defence: social security	10	2	3	0
Education	6	4	2	1
Health and Social work	2	3	1	0
Other Community: Social and personal service activities	3	5	7	0
Other: Private Household with employed persons	1	6	5	0
Other: Extra territorial organisations	2	3	2	0
Total	84	62	59	1

In this table there are instances of cells containing single counts which can be protected by combining ages as in Case Study 1.

However for age 68 there is only a single Male in the column. Knowing this individual was in the sample along with their age would enable an intruder to know in which industry they work. Once again confidentiality could be maintained by combining ages although an alternative would be to not publish the Industry variable at all. A disadvantage of this is that a large amount of useful data would be lost and so an awareness of user needs is vital in assessing the ways in which the table could be protected.

Appendix D Case Studies – Tables produced from business surveys

The Labour Market statistics which are produced quarterly by the ONS display an output of Average Weekly earnings for the whole of Great Britain and for a number of broad industrial classifications..These are not disclosive but if the level of geography and/or industry was requested at a lower level there could be confidentiality issues.

Case Study D1
Region Y 5 digit SIC classification XXXXX

Average weekly earnings = £450 No of people in this category = 1. Clearly if this cell is published information about this individual could be in the public domain. Either publish the data at a higher SIC level so that there are a greater number of people in the category or suppress this cell along with other cells to avoid disclosure by differencing.

Region X 5 digit SIC classification XXXXX

Average weekly earnings = £400 No of people in this category = 2. If this cell is published each individual could calculate the earnings of the other. Either publish the data at a higher SIC level or suppress this cell along with other cells to avoid disclosure by differencing.

Case Study D2
Region W 5 digit SIC classification XXXXX

Average weekly earnings = £244. No. of people in this category = 5. (The actual values being, 1000, 70, 60, 50, 40). The second contributor can work out the total value and an estimate of the earnings of the largest contributor to the cell (total = 244* 5 = 1220; estimate = 1220 – 70 = 1150). By applying the p% rule with p = 20% this cell can be shown to be disclosive. This cell can be suppressed or the table can be redesigned.

Case Study D3
The previous case studies recommended either redesigning the table or suppressing cells in order to suppress the table. Another possibility (especially for frequency tables) is rounding. This has a number of variations but typically each cell value is changed to a multiple of a rounding base 3, 5, 10 or 100 are possible values depending on the nature of the data.
One example is from the ONS Wealth and Assets Survey (2008/10) where weighted values are rounded to the nearest 100. A subset of the table is shown in Table D3.

Table D3

Mean total household wealth and percentage of households with wealth greater than this amount; by region, Great Britain, 2008/10

Region	**All Households (unweighted)**	**All Households (weighted)**	**Mean Total Household Wealth (£)**	**Number of households with wealth greater than mean (unweighted)**	**Number of households with wealth greater than mean (weighted)**
North East	960	1,097,400	£318,900	370	348,200
North West	2,336	2,850,900	£357,000	909	915,900
Yorkshire and The Humber	1,902	2,198,700	£355,700	701	672,000
East Midlands	1,684	1,859,100	£386,100	668	592,300
West Midlands	1,800	2,222,200	£372,900	713	682,300
East of England	2,022	2,365,300	£462,300	798	747,000
London	1,890	2,967,400	£448,700	736	893,600
South East	2,761	3,372,900	£559,400	1,160	1,065,000
South West	1,703	2,185,000	£435,000	750	757,100
Wales	1,117	1,291,800	£377,400	435	395,900
Scotland	1,995	2,314,600	£350,400	758	712,400
Great Britain	20,170	24,725,300	£414,900	7,998	7,781,700