Government Statistical Service

GSS/GSR Disclosure Control Guidance for Tables Produced from Surveys

October 2014

The last few years has seen continued interest from users and producers in ensuring official statistics reach their full utility whilst ensuring the risk of disclosure is minimised.

As a result of this continued interest the Statistical Suppliers and User Group, a forum for engagement between producers and users of official statistics, commissioned the GSS to update its existing disclosure control guidance published in 2007. This revised guidance on disclosure control for administrative sources has been updated to reflect the views expressed by a range of users and producers. The previous version of this guidance was approved by the GSS and GSR and hence applies to both social researchers and government statisticians.

Statistics based on survey data (both social and business surveys) support a wide range of users and uses, providing essential information for government, business, academia and the community. Many of these areas of work require detailed figures, which may raise issues about data confidentiality. Producers of such statistics must ensure that their statistics meet the needs of users by enabling relevant analysis to be carried out while at the same time protecting confidentiality.

This guidance describes the approach that data providers should follow when producing standard outputs and any ad-hoc requests based on a general framework for addressing the question of confidentiality protection. For Freedom of Information (FOI) requests this document can be used in conjunction with the exemptions in the Freedom of Information Act 2000. This is discussed further in Section 4.

The following elements of the framework are described in the guidance with more details provided in the Appendix:

- Determining user requirements
- Understanding the key characteristics of the data and outputs
- Assessing disclosure risk
- Legal and policy considerations
- Disclosure control methods
- Implementation

Relevant associated documents are

The *Code of Practice for Official Statistics* See https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/Official-Statistics-Code-of-Practice.pdf (CoP) and specifically Principle 5: Confidentiality,

National Statistician's guidance: Confidentiality of Official Statistics See https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/Confidentiality-of-Official-Statistics-National-Statisticians-Guidance.pdf (CoOS guidance) set out principles for how to protect personal data from being disclosed.

The *ICO Anonymisation Code of Practice* http://www.ico.gov.uk/for_organisations/data_protection/topic_guides/anonymisation.aspx was published in 2012 by the Information Commissioner's Office and this provides considerable support and guidance to data providers charged with generating and publishing tables (or microdata) from source microdata.

The CoP states 'Ensure that arrangements for confidentiality are sufficient to protect the privacy of individual information, but not so restrictive as to limit unduly the practical utility of official statistics.' This emphasises the dual purposes of statistical disclosure control. Detail relating to individual statistical units is to be protected but the released data must still be of high practical utility for users.

This document will be updated regularly to accommodate the latest thinking in statistical disclosure control.

Contents

1. Purpose

This is one of a pair of documents. This one concentrates on tables produced from sample surveys while the second one (also available from the same source) gives guidance on tables produced from administrative data.
ONS, the National Statistician and departmental Heads of Profession are responsible for statistical and survey standards for the production and release of National Statistics. This guidance covers the disclosure control of tables for 'common families' of surveys: social surveys, business surveys and subsamples from administrative data. The guidance is based on ONS statistical disclosure standards that set out minimum requirements to ensure confidentiality of tables for public release.

This guidance describes the approach that data providers should follow when producing standard outputs and any ad-hoc requests based on a general framework for addressing the question of confidentiality protection. For Freedom of Information (FOI) requests this document can be used in conjunction with the exemptions in the FOI.

The original guidance was approved as by the GSS and GSR and hence applies to both social researchers and government statisticians. This version has been further updated following comments from a range of interested parties.

Maximising access to GSS/GSR data and safeguarding their confidentiality are both vital to maintaining trust in the GSS/GSR. Convincing survey respondents that we take the protection of their details very seriously is a prerequisite for collecting high quality information and obtaining good response rates. The GSS/GSR must also respond to user needs and concerns, and maintain its reputation for quality service. In order to manage the risks involved with handling people's information and to ensure it provides relevant data for its stakeholders the GSS/GSR needs to adopt consistent standards and approaches.

Code of Practice for Official Statistics[1] *(CoP) and* specifically the *National Statistician's guidance: Confidentiality of Official Statistics*[2] *(CoOS guidance)* set out principles for how we protect data from being disclosed. Except in the case that respondents have given consent for information to be published (see CoOS paragraph 22), it is recommended that a guarantee should be given to survey respondents that the confidentiality of their information will be protected (see CoOS paragraph 33). For example the ONS publishes Respondent Charters both for surveys of households and individuals[3] and for business surveys[4]. The statistical disclosure control methods in the GSS/GSR guidance are designed to ensure that the GSS/GSR meets this guarantee in any tables published from the applicable surveys.

The Code of Practice states 'Ensure that arrangements for confidentiality are sufficient to protect the privacy of individual information, but not so restrictive as to limit unduly the practical utility of official statistics.' This emphasises the dual purposes of statistical disclosure control. Detail relating to individual statistical units is to be protected but the released data must be of high practical utility for users.

[1] See http://www.statisticsauthority.gov.uk/assessment/code-of-practice/index.html
[2] See http://www.statisticsauthority.gov.uk/national-statistician/guidance/index.html
[3] See http://www.ons.gov.uk/ons/about-ons/get-involved/taking-part-in-a-survey/information-for-households/respondent-charter---households-and-individuals/index.html
[4] http://www.ons.gov.uk/ons/about-ons/get-involved/taking-part-in-a-survey/information-for-businesses/respondent-charter---business-surveys/index.html

The scope, key principles and standards for this guidance are outlined along with information on implementation and evaluation and responsibilities. More guidance about disclosure control methodology for the different outputs is provided in the Appendices.

2. Scope

The guidance applies to tables derived from:

- social surveys of households, social institutions (e.g. schools, hospitals) or individuals, for example, the Labour Force Survey (LFS) or the Living Costs and Food Survey (LCFS) which rely on the knowing and voluntary participation of the respondent.
- subsamples of census or other whole population data (referred subsequently as subsamples), where only a fraction of the data source is included in the sample without the knowledge of the original respondent. This fraction is assumed to be of the order of 1% to 5% of the target population or any subgroups of the whole population. The target population is defined as the set of statistical units about which information is sought and estimates required
- business surveys including the Interdepartmental Business Register (IDBR) and sample surveys where a business is the selection unit or output unit.

Tables from these surveys include those:

- released by GSS/GSR to the public through normal publications and customer requests. This includes tables sourced outside GSS/GSR, but published in GSS/GSR reports.
- released through the Freedom of Information Act (2000) or in Scotland, the Freedom of Information (Scotland) Act (2002). This guidance can be used to help decide which exemptions in the Act are relevant, and which should be cited when withholding confidential statistical information. Whilst it is good practice to explain general policy for the withholding of information this must be done in addition to, and not in place of, the exemptions in this Act.
- derived from microdata or other non-publishable data accessed by licensed users.

Exemptions and exception arrangements for this guidance are detailed in Section 5. Guidance for data produced from administrative data is also available. Updated administrative guidance is being released at the same time as this guidance.

3. Key Steps

- **The overriding principle in the design of survey outputs should be quality and not disclosure. Historically many surveys have restricted outputs for quality reasons, such as not publishing estimates with high sampling errors. This has generally also provided confidentiality protection.**
- Confidentiality protection is needed for tabular outputs derived from surveys due to legal rights and obligations, national and international standards for statistics and in order to maintain public trust and cooperation.
- Effective disclosure control requires an understanding of the key characteristics and uses of the data.
- Disclosure risk should be assessed by considering intruder scenarios along with Case Studies and using these to identify unsafe cells in the tabular output.

- Disclosure control rules and methods should be implemented to reduce and manage disclosure risk. The choice of method must balance risk with the uses to be made of the outputs and simplicity of approach.
- Implementation of disclosure control methods should involve the application of standard tools where appropriate. When looking at published statistics, users should be aware that the dataset has been assessed for disclosure risk, and methods of protection may have been applied. For quality purposes, users should be provided with an indication of the nature and extent of any modification due to the application of disclosure control methods but the level of detail made available should not be sufficient to allow the user to recover disclosive cell counts, totals or averages.
- Different approaches may be applicable to weighted and non weighted data. Non weighted outputs will display the actual value (potentially more disclosive) whereas weighted data will be summed up to reflect the population.

Ultimately consideration of aspects relating to a particular table and not the underlying microdata should drive the level of statistical disclosure control which is applied.

Figure 1 shows the main steps which ought to be taken when considering statistical disclosure control in relation to tables created from social survey data. These steps describe the process to be followed in general terms. The steps are broadly defined so as to reflect the prevailing pressures for publishing data. For example, the current era of open data can be reflected by taking a more rigorous approach to step 3 below.

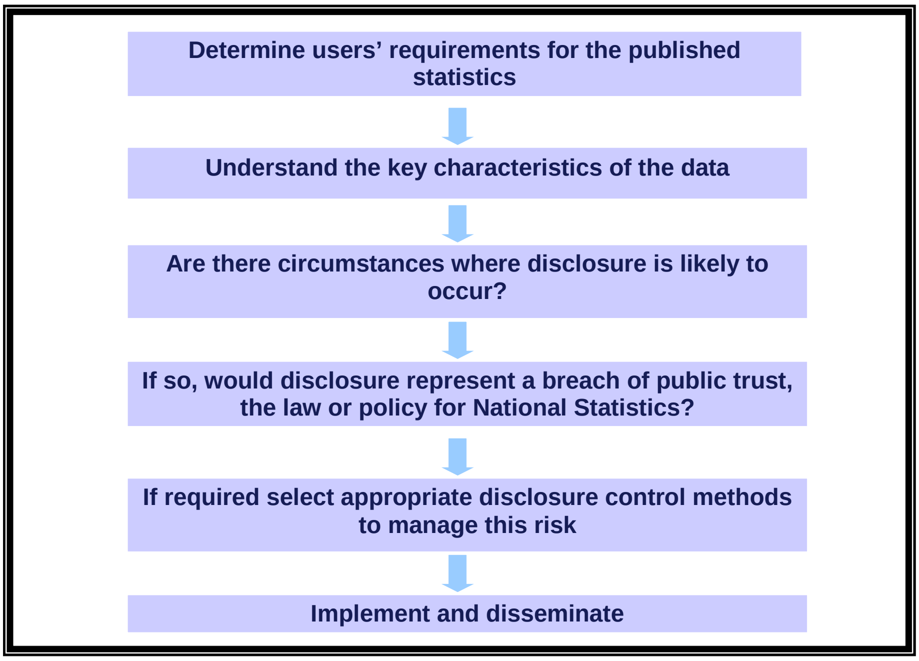
Determine users' requirements for the published statistics
Understand the key characteristics of the data
Are there circumstances where disclosure is likely to occur?
If so, would disclosure represent a breach of public trust, the law or policy for National Statistics?
If required select appropriate disclosure control methods to manage this risk
Implement and disseminate

Figure 1: Main steps to follow to ensure access to non-disclosive statistics

3.1 Intruders and Intruder Scenarios

An individual or organisation that might seek to discover the identity of a statistical unit, and/or discover additional information about a statistical unit, is referred to as an intruder. Note, however, that an intruder may be inadvertent (or not motivated) if they may spontaneously recognise a respondent in the microdata.

There will be a number of different ways in which an intruder may be able to discover information about an individual. These are referred to as intruder scenarios, as discussed in Elliot and Dale (1998, 1999). Different types of data collection and variables may make some scenarios more likely than others. Data providers need to consider the intruder scenarios most likely to apply to their dataset(s). Some general pointers are provided here for social surveys, subsamples and business surveys.

Social Surveys

Intruder scenarios:	**Unsafe cells:**
Respondents who are unique or rare in the population are identified by an intruder with some external knowledge of the population Respondents who are unique in the sample are identified by an intruder who knows the respondent is in the sample Respondents in the sample are identified by another respondent with the same characteristics as themselves	Cells based on one or two statistical units[5]

Subsamples

Intruder scenarios:	**Unsafe cells:**
Respondents who are unique or rare in the population are identified by an intruder with some external knowledge of the population	Cells based on one statistical unit

Business Surveys

Intruder scenarios:	**Unsafe cells:**
Any person or business, not a member of the cell, attempts to identify a cell respondent and deduce the exact value or a close approximation of the response A business contributing to a cell identifies another business contributing to the cell and deduces the exact value or a close approximation of the other's response	Cells with one or two contributing units Cells with one or two dominating units as defined by the p% rule as defined in Appendix D3.2.1

[5] Examples of statistical units are individuals, families, households and communal establishments

4. Guidance

This section provides a summary of the disclosure control standards for the different survey types. More details of the standards and methods are provided in the appendices.

4.1 Social Surveys

- For the majority of surveys, outputs should be for large geographical areas, e.g. Country or Government Office Region, or in some cases Local Authority District (or equivalent). The level of geography should reflect survey design.
- Suppress or combine unsafe cells, i.e. where there are one or two units contributing to the cell.
- Where the sample size of a total or sub-total is one or two, suppress the whole row or column to which the total refers, including any zero cells (or combine neighbouring categories).
- In unweighted tables, cell suppression does not provide sufficient protection. Unsafe cells should only be combined with other cells.
- If unweighted sample base numbers are essential they should be conventionally rounded to base 10.
- Percentages may be released, provided it is not possible to deduce where only one or two units have contributed to the cell.
- Units may be individuals, families or households, communal establishments or any other unit whose confidentiality should be protected.

4.2 Subsamples

- For the majority of surveys, outputs should be for large geographical areas, e.g. regions, or in some cases Local Authority District (or equivalent). The level of geography should reflect survey design.
- Table design should be used to remove all unsafe cells, i.e. where there is one unit contributing to a cell. Variable categories should be combined or variables removed until only safe cells remain.
- Percentages may be released, provided it is not possible to deduce where only one unit has contributed to the cell.
- Units may be individuals, families or households or any other unit whose confidentiality should be protected.

4.3 Business Surveys: Magnitude tables

- A cell meeting both the following criteria is safe (otherwise the cell is unsafe):

- there must be at least *n* enterprise groups in a cell (threshold rule)

- the total of the cell minus the largest *m* reporting unit(s) must be greater than or equal to p% of the value of the largest reporting unit (p% rule)

Note that the values of the p% and minimum threshold parameter *n* and m should remain confidential, since knowledge of these values reduces the protection. The choice of p, n and m would usually be decided by the Responsible Statistician. Typical examples would be 2,3,4,5 (for n), and 2,3 (for m) and 5% 10% ,15%, 20% (for p).

- Table design should be used first to reduce the number of unsafe cells in a table where this is consistent with the main uses of the data.
- Cell suppression is the standard method used to protect tables with unsafe cells. The unsafe cells are suppressed, known as *primary suppressions*. Other cells must be suppressed to prevent the values of the unsafe cells being calculated by subtraction from the marginal totals of the table. These are known as *secondary suppressions*.

Cell suppression does not generally provide protection from disclosure by differencing. Tables should be published using fixed categories to avoid disclosure by differencing. For example the same geographies and SIC codes should always be used.

4.4 Business Surveys: Count data

- Tables of count data are to be protected by redesign of the table to protect sensitive cells. If further protection is required other techniques such as controlled rounding to base 5 should be considered.
- Percentages or rates must be derived from rounded values.

5. Implementation and Evaluation

In accordance with the CoP, Principle 5: Confidentiality, Practice 4 and as outlined in the CoOS guidance paragraphs 36-40 the implementation will achieve the obligation to protect against disclosure but will through choice of disclosure method strive to give the most practical utility possible in the released statistics. Standard wording should be used in dissemination releases. No specialised software is needed to implement this guidance for social surveys and subsamples. For business surveys the Tau Argus[6] software (version 3.5.0 from this link) is available to implement cell suppression and controlled rounding. Tau Argus is the most thorough of a number of disclosure control packages and a link with SAS is under development.

Situations where this guidance may not apply, and any approval processes required, are given below.

Exemptions. Exemptions required by legislation from application of the confidentiality guarantee must be listed on the Register of Exemptions. See CoP: Principle 5, practice 5 and the CoOS guidance. Paragraphs 41-45 and Annex.

Application of non-standard methods. Other methods than the standard method described in this guidance may be used for confidentiality protection if

1. they can be shown to provide equivalent protection to the standard method

or

2. higher levels of protection are required because of special circumstances relating to the data.

Further advice may be sought from the ONS Strategy and Standards Directorate, Statistical Disclosure Control Centre.

[6] See http://neon.vb.cbs.nl/casc/tau.htm

Reduction in confidentiality protection. If none of the exemptions in CoOS Annex C apply but a data provider wishes to provide less confidentiality protection than that described in this guidance, approval is required from the Head of Profession. Further advice can also be sought from the Statistical Disclosure Control branch at ONS.

Access to disclosive tabular data. Particular procedures have to be followed when tables are potentially disclosive. CoP Principle 5, practice 2 and CoOS paragraphs 23–29 cover the principles and arrangements for access to all disclosive (i.e. identifying) data, and this includes disclosive tabular data as well as the more frequently requested microdata. At the ONS approval for release of disclosive tabular data and microdata must be obtained from the Microdata Release Panel. Each Government Department will have their own procedures when allowing access to disclosive tabular data and microdata. In the devolved administrations approval will be required from the Chief Statistician.

6. Responsibilities

Each publication has an associated Responsible Statistician who is responsible for confidentiality protection of released data, ensuring that the standard disclosure control methods are applied, and any other special circumstances are taken into account. The disclosure control method will always be signed off by, or in the name of, the Head of Profession (not always the same person as the Responsible Statistician) which for certain releases may be the National Statistician or the Chief Statistician in a devolved administration.

Day to day management of disclosure control for data release may be delegated to output managers, data managers or others responsible for the confidentiality guarantee pertaining in outputs from administrative sources, whether the data are released by GSS or by others using data from this source.

ONS SDC Methodology team are happy to help and offer advice to users and data providers where necessary. Their contact email is
Sdc.queries@ons.gsi.gov.uk

Appendix A: General Guidance

General Guidance for Disclosure Control for Tables Produced from Surveys - Does the disclosure risk identified constitute a breach of statistical obligations

The key principles as shown in Figure1 contain a section where the policy and legal consequences of releasing disclosive data are discussed. Greater detail on this section is described here for all tabular outputs from surveys while the remainder of the steps are described in Appendices B to D for outputs from Social Surveys, Subsamples and Business Surveys respectively.

A1. Legislation

All outputs from tables released by the ONS have to comply with Section 39 of the Statistics and Registration Service Act[7] (2007) (SRSA). This defines personal information as data which could reveal the identity of an individual or organisation, or any private information relating to them, through being specified in the data, by being deduced from the data, or by being deduced from the information when taken together with any other published information. Under this legislation it is an offence to disclose any personal information derived from the census, unless one of the exemptions to Section 39 applies.

ONS surveys are conducted on behalf of the UK Statistics Authority, and all outputs are subject to Section 39 of the SRSA. Surveys carried out by other departments will generally be governed by other legislation or statutory provisions. However all Official Statistics must comply with the confidentiality regulations of the Code of Practice. In addition There is a code specific to Government Social Researchers[8] (GSR) which discusses the legal and ethical aspects of research by members of this group..

For all household surveys, the public statements made about protection of confidentiality define a common law duty of confidence that the GSS/GSR must legally comply with. Common law means that we must do what we say we will do. Breach of the common law creates a right for the individual respondent to sue for damages.

Subsamples taken from whole population data must comply with any legal statutes that apply to the source data, including the case where the source data are administrative data.

Where an applicant requests personal data that relate only to themselves, this becomes a Data Protection Act (1998) matter. Section 33 of the Data Protection Act exempts the provision of access to a person's own personal information where this information is processed for statistics and research purposes only.

Business Surveys carried out by the ONS operating within the United Kingdom are governed under the Statistics of Trade Act (1947). Under the Act, participation in the survey is compulsory, and confidentiality requirements that relate to published data are specified in Section 9. This states that tables should not be published that would disclose any information relating to an individual business, unless there is expressed consent in writing from that business. Nor should data be published that would reveal the exact number of respondents contributing to a cell, if there are fewer than five contributors.

[7] See http://www.legislation.gov.uk/ukpga/2007/18/contents
[8] http://www.civilservice.gov.uk/networks/gsr/gsr-code

As well as the Statistics of Trade Act, disclosure of data received from HMRC under the Value Added Tax Act (1994) relating to individual undertakings is prohibited without the consent of that undertaking. Where data are received from HMRC under the Finance Act (1969) they cannot be disclosed at all.

Surveys of businesses not run by the ONS are governed by Acts such as the National Statistics Code of Practice and protocols, the Data Protection Act (1998) and the Freedom of Information Act (2000). In addition for many surveys there is additional EU legislation which imposes confidentiality constraints on the data.

A2. GSS/GSR Policy: Code of Practice and Protocols

It is a United Nations fundamental principle of official statistics that the records of individuals, businesses or events used to produce official statistics are kept strictly confidential[9]. The Code of Practice and CoOS guidance both conform to this principle and provide the GSS policy framework for official statistics. The CoP guarantees confidentiality to those who provide private information for the production of Official Statistics:

> Statement of Principle: Private information about individual person (including bodies corporate) compiled in the production of official statistics is confidential, and should be used for statistical purposes only.

Within the GSR Code there is a GSR Ethics Checklist[10] a Principle of which states that there should be non-disclosure of identity and personal information.

The disclosure control methods presented in this paper are judged to have met the requirements of this confidentiality guarantee GSR Ethics Checklist Principle.

A3. Statements made to respondents

The confidentiality pledge is an assurance of confidentiality given to survey respondents. The specific wording may vary between surveys but is likely to be similar to that in the respondent charter for ONS surveys of households and individuals, which states that

We will:

- recognise that you are sharing personal information with us and treat it as confidential, as directed by the Code of Practice for Official Statistics
- only use your information for statistical purposes and not publish anything that identifies an individual, household or business
- provide survey information to other organisations and approved researchers for statistical and research purposes only, subject to the same standards of protection we apply (and according to the provisions of the Data Protection Act, the Statistics and Registration Service Act and the Code of Practice for Official Statistics)
- ensure our staff are aware of their obligations to protect your confidential information;
- maintain the security of our systems and our buildings so that your data remain secure

The equivalent charter for ONS business surveys states that

[9] See http://unstats.un.org/unsd/goodprac/bpaboutpr.asp?RecId=6

[10] http://www.civilservice.gov.uk/wp-content/uploads/2011/09/gsr_ethics_checklist_tcm6-7326.pdf

We will:

- recognise that you are sharing confidential/personal information with us and treat it as confidential, as directed by the Code of Practice for Official Statistics
- only use your information for statistical purposes and not publish anything that identifies an individual business
- provide survey information to other organisations for statistical purposes only, subject to the same standards of protection we apply
- ensure our staff are aware of their obligations, and follow correct procedures, to protect your confidential information
- maintain the security of our systems and our buildings

Subsamples must comply with any statements made to respondents for the source data, including administrative data.

A4. Trust of respondents

In a very real and practical sense the GSS and GSR rely on the cooperation and goodwill of respondents (even though business surveys are compulsory under the Statistics of Trade Act) to provide the data that are the basis of our official statistics.
For GSR one of the Principles is that of informed consent, those taking part in research are informed about the major details of the project prior to their participation.

Fundamental to maintaining that trust is ensuring that identifiable information is held securely and not revealed in published outputs. In particular, some of the data collected by business surveys are commercially sensitive and recent data in particular may have high value to competitors.

If respondents do not trust us to keep their data safe, they may be reluctant to respond or may supply poor quality information**.**

Appendix B: Guidance for tables produced from Social Surveys

General guidance for disclosure control for tables produced from social surveys

B1.Determining User Requirements

A large number of tables are produced from social surveys which are used by National and Local Government for policy purposes as well as by academic researchers and the general public. It is a requirement that these tables have maximum utility for users whilst also maintaining the confidentiality of the respondents.
Standard tables are produced following discussion with users to ensure the outputs are relevant to a wide range of policy and research purposes. Disclosure control should have as little affect on these outputs as possible. Ad hoc tables can also be requested at various levels of detail and these require individual consideration with respect to statistical disclosure control.

B2. Understanding the key characteristics of the data and the required outputs

This section of the guidance relates only to social survey data derived from samples of the population. The majority of survey samples have a small sampling fraction (< 2%) while some have a medium sampling fraction (2% to 5%) or a large fraction (> 5%) There are many different surveys with different aims and different user groups, collecting a very wide range of variables, but all share certain similarities in that response is voluntary and the sampling rate is usually low. The key characteristics that relate to confidentiality are noted below.

B2.1 Table properties
Common demographic variables include age, sex, ethnicity, marital status and primary economic activity. Household variables typically include size of household, household composition and family type.

Outputs are often expressed as frequency tables (as counts and percentages), but may also include magnitude tables (for example, average household expenditure, level of unemployment).

For smaller surveys such as the ONS Living Costs and Food Survey (LCFS) tables are published at a high geographic level, usually Region. For large surveys such as the LFS, tables are produced at lower geographic levels (Local Authority District).

Publication intervals may be monthly, quarterly, annual or one-off. Ad hoc requests for non-standard tables may be received.

B2.2 Sample design
Sample designs vary: some are a stratified but unclustered sample of addresses (LFS); others have a stratified clustered design where clusters are sampled with probability proportional to size. Usually the clusters are areas (LCFS) but can be establishments such as schools (such as the 2003 Survey of Children's Dental Health). Most GSS/GSR samples use systematic sampling, whereby units are sampled at regular intervals through the sampling frame. The common feature essential for application of this standard is that the sample

fraction is small, defined as less than 2% of the target population and that the sampling fraction is also small for any identifiable sub-groups of the target population. There may be particular instances where the sampling fraction is greater than 2%. These could include specific Local Authorities with small populations where a larger sample is necessary. In addition extracts from administrative data sources could be larger than 5%. There is advice in Section B5.2 (High Sampling Fractions) but there are no definitive rules on sample design and the Survey Manager or equivalent will need to make the final decision.

Sample data provides specific statistical disclosure control issues. There is a degree of protection supplied by the data being a sample. An attacker may know that an individual in the population has specific characteristics but there is no guarantee that they are in the sample unless the individual has informed the attacker or the attacker has discovered this by a more circuitous route. Response knowledge is of great importance here. If an attacker has knowledge that an individual has taken part in a survey they may be tempted to search the resulting tables in an attempt to identify them or discover a related attribute. Without this knowledge an individual in the survey has another layer of protection.

B2.3 Weighting

Generally, social surveys employ post-stratification or calibration weighting to adjust for non-response and under-coverage. Weights produced through this methodology tend to be quite variable and often unique for each household. Where surveys have not adopted this approach, the survey may still use weights to adjust for selection probabilities or non-response, but the weights tend to be less variable. Where weights are not variable, perhaps used solely to adjust for selection probabilities, extra care should be taken; particularly if tables are defined by variables used as weighting classes or strata, such as Region or age groups. If an intruder can gain information about the size of the weights used, either directly from a technical appendix, or indirectly from published response rates, then the output manager should adopt the disclosure control standard for unweighted data.

B2.4 Data uses

The ways in which data are used are many and varied, ranging from central and local government for policy and planning purposes, special interest groups, academic and other researchers, to local communities and the general public.

B2.5 Microdata access

This subsection is shown for completeness as being a potential output from social survey data.. There is much more detail on microdata disclosure control in the GSS policy document 'GSS Guidance for the Disclosure Control of Microdata Derived from Social Surveys' currently being revised and scheduled for release in 2014.

There are 4 distinct levels of microdata access from ONS and these are the same, or similar, in other parts of the GSS:

1. Public Use File. This datasets would generally be released under an Open Government License (OGL) with enough protection provided to ensure that any outputs would have a negligible probability of being disclosive.
2. A file released under a Safeguarded license. This can be thought of as being similar to an End User license. Users must not attempt to identify individuals in the data nor claim that they have identified anybody. The level of protection for these data is sufficiently high to ensure that outputs will have a low risk of disclosure and can be classified as 'not personal information'.

The two other levels of access require the user to have approved researcher status allowing access to personal record level data for specific projects. For less sensitive data desktop access is allowed but for the most sensitive data, a file must be accessed via a secure data laboratory.

B3. Circumstances where disclosure is likely to occur and managing this risk

For sample surveys, some protection is provided by sampling and weighting (if there are weights). Any randomly selected unit has a probability of not being included in the sample, this probability depending on the sampling fraction. But it may still be possible to identify a respondent in the sample. This section sets out the disclosure risks that may be present in the data, which need to be protected. We first consider disclosive situations as intruder scenarios, and then show which cells in a table present a disclosure risk.

We identify three types of disclosure risk:

1. attribute disclosure - disclosure of information about a statistical unit, not already in the public domain
2. identity disclosure - identification of a statistical unit
3. self-identification

In many situations, for attribute disclosure to occur, it must be preceded by identification. This is the main reason for avoiding identity disclosure. Also, the outputs should not allow an intruder to identify a statistical unit, even if all of the details about that unit in the table are already known to the intruder. The concept of identification carries an element of uniqueness, the association of a name with a cell value.

Attribute disclosure includes inferential disclosure, where information about a statistical unit can be inferred with a high degree of confidence.

These types of disclosure risk are more likely to be present under ad-hoc requests as standard tables should be designed to ensure that disclosive outputs occur infrequently. However all tabular outputs should be checked for possible confidentiality breaches. Examples are shown as Case Studies below.

B3.1 What intruder scenarios are we protecting against?

Attribute disclosure may occur when part of the information in a table is used to identify an individual statistical unit, revealing further information. For attribute disclosure to occur, the individual must first be identified. Individuals in a sample might be identified in the following ways:

1. Sample respondents who are unique or rare in the population may be identified by an intruder with some external knowledge of the population. This becomes more likely as the size of the relevant population decreases, for example in smaller geographic areas, or a minority ethnic group (identification of *population uniques*). This scenario applies to social surveys as defined in Section 2 (Scope) of this document.

2. Respondents who are unique in the sample may be identified by an intruder who knows the respondent is in the sample (identification by *response knowledge* from *sample uniques*). This scenario applies to social surveys as defined in Section 2 (Scope) of this document.

3. Respondents in the sample may identify another respondent with the same characteristics as themselves (identification by *response knowledge* from *sample pairs*). This scenario applies to social surveys as defined in Section 2 (Scope) of this document.

Most social surveys need not consider self-identification as a disclosure risk under the terms of the CoP. But particular care should be taken if there are highly sensitive variables that might cause risk of substantial damage or distress to a respondent who was able to identify themselves. For some surveys perception of disclosure from self-identification may be an issue. So, any intruder scenario that leads to identification of a respondent presents a disclosure risk.

B3.2 What parts of the outputs pose a disclosure risk?

The intruder scenarios depend on the type of sample. Response knowledge can only occur if the intruder or respondent is aware that an individual is in the sample.

If there is a single respondent contributing to a cell, then the respondent is a sample unique for the variable categories that define the cell. They may also be a population unique, though this is much more difficult to determine accurately for sample data. In certain surveys the 'population' sampled may only be a subset of the actual population. The disclosure risk in any published table will increase as the sampled population decreases. If the sampled population is a sensitive subset, the release of any disclosive data could be especially problematic.

The uniqueness may lead to identification by an intruder. For example, there may be only one 83-year-old from a minority ethnic group in the sample, and they may be identified. For identification to occur, the variables defining the cell must be identifiable (known to third parties, and can be used to identify a unit). The identification can then lead to disclosure of further information.

Additionally, respondents in the sample may identify another respondent with the same characteristics as themselves (identification by response knowledge from sample pairs). Thus, a cell with two respondents contributing to a cell may also pose a disclosure risk.

For magnitude tables, identification, along with knowledge of the sample weight, will reveal the value supplied by the respondent. However, magnitude tables are relatively uncommon in social survey data. However there are examples such as the amount spent on particular items in LCFS. Magnitude tables can also be produced from the Wealth and Assets Survey (WAS).

If the cell is also further broken down by another variable, for example, income, then the income band of the identified respondent is revealed. This may occur in a table when a marginal sub-total has only one contributing respondent, allowing identification, and the internal cell reveals a further attribute. When there are multiple tables produced from the same microdata, *any* cell with one contributing respondent, in a particular table, may appear in another table (either previously produced or in the future) broken down by a further attribute. Thus whenever flexibility is required for multiple tables from the same microdata, to protect against attribute disclosure, all cells that lead to identification are a disclosure risk. This is the case for most social survey data.

Unsafe cells are those which pose a risk of disclosure, and are defined as cells where one or two respondents contribute to the published value. Low counts published from a social survey sample would be of low quality with much uncertainty surrounding the true value.

Unsafe cells may be present in a table, but may also be the result of differencing between two tables. Differencing is only a disclosure risk for small samples where unweighted values are provided. (see B2.3)

Cells are only unsafe if an intruder is able to determine with some certainty that one or two respondents contribute to the cell.

B3.3 What is not a disclosure risk?

We do not consider that zero cells would normally be a disclosure in the way they may be for population data. Zeros in population data allow one to say that no-one in the population has that attribute. A zero value from a small sample does not allow one to infer this, since there may be population units with those attributes who were not sampled. Thus, attribute disclosure without identification is not considered a disclosure risk.

Disclosure by differencing occurs when comparison of two or more tables reveals information that is not available from any single table. For example, the difference in counts between a table of age with age group 15-19, and one with age group 15-18 would reveal values for 19-year-olds. With weighted estimates from household surveys, differencing is of less concern since the weights should provide enough uncertainty to protect confidentiality.

Identification of two or more units, when the intruder does not have response knowledge. Small sampling fractions make it highly unlikely that an intruder could correctly identify more than one person in the population.

Identification of three or more units, when the intruder does have response knowledge. It is assumed that respondents are unable to identify more than one other respondent. This is reasonable for most surveys, but may not be true for some cluster designs, or when very valuable information collected in the survey leads to a highly motivated intruder.

B4. Does the disclosure risk identified constitute a breach of statistical obligations?

The relevant legislation and protocols are discussed in Appendix A.

B5. Selecting Disclosure Control Rules and Methods

B5.1 Requirements of the standard

The disclosure control method depends upon the sample design and estimation. If final sample weights used in the estimation are sufficiently variable, this provides uncertainty over how many respondents contribute to a cell, especially so if the weights are unknown. There will also be high sampling error associated with a very small number of contributors, so suppression of these cells also links in with statistical purpose and quality standards for publication. So, as long as only weighted estimates are published, sufficient protection will be obtained by suppressing estimates for unsafe cells. Secondary suppression is not required.

Publishing unweighted sample base numbers allows more information about cell sizes to be deduced, and compromises the protection. If essential unweighted sample base numbers should be modified to ensure that one could not infer only one or two contributions to a cell. They can be conventionally rounded to base 10 (this is used in preference to base 5 since a zero in the latter would indicate that a cell is a 0, 1 or 2). Probabilistic or controlled rounding could also increase the possible range of values. Alternatively cells can be combined with larger cells and one figure given for both.

If weights are known, or unweighted estimates are produced, it is necessary to disguise or remove unsafe cells. The standard method is to use table design to remove all unsafe cells. Variable categories should be combined or variables removed until only safe cells remain.

Disclosure risk increases with smaller geographies. As a guide for the majority of surveys, outputs should be for large geographical areas, e.g. regions, or in some cases Local Authority District (or equivalent). The standard reflects the fact that in the majority of cases tables produced below this level contain a high proportion of disclosive cells (containing a 1 or a 2) and that the practical utility of a table with a substantial number of suppressed cells would be diminished.

B5.2 Guidance for applying the standard

Unusual circumstances. A survey should also consider whether there are any unusual circumstances relating to their data that would mean other cells are unsafe, and apply appropriate protection. Raising the minimum cell size, and/or increasing the level of geography are the simplest methods of protection. As the number of sample contributors in a cell increases, it rapidly becomes much less likely that an individual may be identified. Examples are:

- small sub-populations for easily identifiable variables such as ethnicity, occupation, religion, country of birth, especially if the sample design includes clusters (geographic or other variables)
- other identifiers that could be used in combination with standard area classifications to identify small geographies , for example urban/rural indicators
- variables that may be considered highly sensitive and/or of high value to an intruder

Geography. A minimum population size of 50,000 should be used as a general indication of equivalence to Local Authority District geography. Note that while the City of London and Isles of Scilly are defined as Local Authorities, they must be combined with other geographies to meet the minimum size.

Extreme values. A respondent with an extreme value of a continuous variable (such as income), may be dealt with by reducing the weight for the outlying respondent. This would be appropriate for summary statistics, means, for example, but may not be sufficient for graphical output.

Percentages. Percentages may be released, provided it is not possible to deduce where only one or two units have contributed to the cell. For example, small percentages may be rounded to zero. Percentages must be rounded sufficiently, depending on the sample size on which they are based. It is essential that the underlying cell counts are not released. Output managers should ensure that the underlying count is not published in the same table and it is not available elsewhere in the same publication.

High sampling fractions. If the sample design includes high sampling fractions for some sub-groups (ethnic groups, areas and so on), then disclosure control should be applied as for a Census or other whole-population data source.

Higher level units. The statistical units on which a table is based may belong to another higher level unit, for example an establishment like a school or prison. If this higher level unit is unique in the population for a combination of characteristics, it may be identified by an intruder, e.g. it is the only private school in a local authority, for example. It may also be possible to identify the establishment in a table where counts of individuals in a cell all belong to the same establishment. Protection then needs to be applied to the higher level unit, as well as the individual.

Cluster sampling. Particular care should be taken with cluster sampling, as the clusters may make identification easier. For example:
a cluster sample design based on birth dates will be close to a simple random sample except for biological twins and higher multiple births. Then some cells of 2 or more may be a disclosure risk if they consist only of multiple births (for example, birth cohort studies).
If clusters are institutions (schools, for example), and the output unit is individuals (pupils) there is a higher risk of disclosure. Firstly, pupils are likely to know more than one other respondent in the sample because they may have completed the questionnaire in the same room. Secondly, many people other than the respondents themselves may know that the cluster was sampled, for example teachers, parents and education administrators at local education authorities. Identification of the cluster unit may be disclosive in its own right, or may assist the identification of individuals.

If all units in the cluster are sampled, for example, all persons within a household, then identification of the cluster would identify the units within the cluster.

B6. Implementation Issues and Concerns

B6.1 Impact on Quality

In accordance with CoP Principle 5: Confidentiality, Practice 4 and as outlined in the CoOS guidance paragraphs 36-40 the implementation will achieve the obligation to protect against disclosure but will through choice of disclosure method strive to give the greatest practical utility possible in the released statistics.

Because unsafe cells are also those cells that are of poorest quality (having high sample error) the disclosure control methods have little impact on the detail and quality of data that can be published. Furthermore, the importance of the small counts, in analysis and for determining policy decisions, is normally extremely small compared to that of those cells containing large numbers of cases. Historically many surveys have restricted outputs for quality reasons, not publishing estimates with high sampling errors. This has generally also provided confidentiality protection.

As a general rule, if a large proportion of the table (around 60%) is suppressed to protect confidentiality, the table should not be supplied due to quality concerns.

B6.2 Tools

No specialised software is needed to implement this standard, although Tau Argus is one package that can be used. Where the disclosure control method relies on table design it should be part of the normal production process. The primary cell suppression is easy to implement in any software, and may also be combined with suppression of poor quality estimates if these are used. Primary suppression of a table can be implemented via a program such as Tau Argus.

If conventional rounding is used for unweighted sample base numbers, any of the usual software should be suitable. (though note that Tau Argus does not implement conventional rounding)

B6.3 Standard wording

Standard statements to include with release of data are as follows

- To inform data users of the confidentiality protection method:

  'Cells in a table based on a small number of respondents are more likely to breach confidentiality. The same cells are also likely to be unreliable. Confidentiality protection is provided by releasing only weighted estimates and by suppressing the values for unsafe cells. Information on the exact number of sample respondents is restricted.'

- To inform data users of impact on quality of data:

'The effect of disclosure control on the quality of data that can be released is very small because data that appear disclosive may also be of low quality.'

- Wording of footnote to tables where suppression has been used:

  '[symbol] Cells have been suppressed to protect confidentiality.
  ‘c’ is the standard symbol used.

B7. Case Studies

Please see the associated document ‘Case Studies for tables produced from survey data’

Appendix C: Guidance for tables produced from Subsamples

General guidance for disclosure control for tables produced from subsamples

C1.Determining User Requirements

The main users of the data are researchers who access microdata either in a safe setting within ONS, or under licensed access agreements. Examples of uses from the Longitudinal Study (LS) include studies of mortality, cancer incidence and survival, fertility patterns and of change between censuses.
Outputs are produced by users from the microdata for publication or to aid research.

C2. Understanding the key characteristics of the data and the required outputs

The types of tables produced and the uses of the data depend on the source data. ONS examples are the SARs and CAMS sub-samples of 2001 Census (soon to be followed by subsamples from Census 2011), and the Longitudinal Study (LS) which is a 1% sample of census data linked between censuses and with vital events data. These all produce tables of counts, usually unweighted.

The LS and CAMS microdata files contain very detailed census data and are accessed only within ONS safe settings. All outputs are checked by ONS staff before they are released.

Subsamples from the 2001 Census are accessed as described below. Access to microdata from Census 2011 will be in line with description Appendix 2 Section 2.5.

The 2001 Household SAR is accessed under the ONS Special License through the UK Data Archive. Disclosure control rules for outputs are part of the Special License agreement, and are consistent with this standard.

The 2001 Individual and Small Area SARs files are accessed under the UK Data Archive End User License. The microdata files have been modified using disclosure protection methods to ensure that the microdata are considered non-disclosive under the conditions of the License. No further protection is required for tabular outputs.

The 2001 SARs and CAMS sample design is a stratified systematic sample, with the same or very similar selection weights across strata. The LS is a cluster design where the clusters are four birth dates (day and month).

The sampling introduces sample error that is not present in the source data. Data users are generally given sufficient information about the sample design to allow accurate calculation of sample errors. There is no additional non-response due to the sub-sampling, and, apart from sample error, data quality of the sample depends on the quality of the source data.

C3. Circumstances where disclosure is likely to occur and managing this risk

C3.1 What intruder scenarios are we protecting against?
As stated in Section 3, the most likely intruder scenario occurs when sample respondents who are unique or rare in the population may be identified by an intruder with some external knowledge of the population. That is, the risk of disclosure is high where the only unit in the sample with given attributes (a *sample unique*) is also a *population unique* - the only unit in the population with those attributes. This becomes more likely as the size of the relevant population decreases, for example in smaller geographic areas, or a minority ethnic group. An "intruder" may be someone deliberately attempting to identify an individual, but also includes any data user who is able to identify an individual, perhaps by spontaneously recognising someone they know.

C3.2 What parts of the outputs pose a disclosure risk?
Cells where one respondent contributes to the published value are unsafe. Population uniques, rather than sample uniques, are disclosive. However in the current applications of samples taken from census (CAMS, and Longitudinal Study), a conservative view is taken that cells of size 1 are assumed to be population uniques, unless shown otherwise. This approach is justified by the very detailed and extensive information about individuals that is made available to researchers, and the high value placed on maintaining the trust of census respondents

Unsafe cells may be present in a table, but may also be the result of differencing between two tables.

There is no extra protection offered by weighted estimates when the weights are the same or very similar for all units, or where weights are provided.

C3.3 What is not a disclosure risk?
Zero cells do not normally create a disclosure risk in small sub-samples as they may for the population data. Zeros in population data allow one to say that all others in a row or column do not fall into the category of the zero cell. A zero value from a small sample does not allow one to infer that any-one in the population has those characteristics.

Cells of size 2 do not normally constitute a disclosure risk of identification of two or more units in the population with the same characteristics. Small sampling fractions make it highly unlikely that an intruder could correctly identify more than one person in the population. However care should be taken if clusters are sampled, since knowledge of the cluster could assist with identification of all sampled units within the cluster (see the LS example below)

There is no disclosure risk from response knowledge, since membership of the sample is not known to respondents themselves or to any other persons.

C4. Does the disclosure risk identified constitute a breach of statistical obligations?

The relevant legislation and protocols are discussed in Appendix A.

C5. Selecting Disclosure Control Rules and Methods

C5.1 Requirements of the standard

- Tables must have 2 units or more contributing to any non-zero cell
- For the majority of surveys, outputs should be for large geographical areas, e.g. regions, or in some cases Local Authority District (or equivalent).
- Percentages, rates or other derived values must be based on safe cells.

Units may be individuals, families or households or any other unit whose confidentiality should be protected.

The standard method is to use table design to remove all unsafe cells. Variable categories should be combined or variables removed until only safe cells remain.

Disclosure by differencing tables may still occur. Tables published or presented publicly as a group by a single research project (or by ONS) must be checked for disclosure by differencing. It is generally not feasible to check across all released tables, and particularly when most access is for research purposes, the disclosure risks are low.

C5.2 Guidance for applying the standard

Unusual circumstances. Unusual circumstances relating to the data may mean that other cells are unsafe. Examples are:

- small sub-populations for easily identifiable variables such as ethnicity, occupation, religion, country of birth, especially if the sample design includes clusters (geographic or other variables)
- other identifiers that could be used in combination with standard area classifications to identify small geographies , e.g. urban/rural indicators, or deciles of deprivation
- variables that may be considered highly sensitive and/or of high value to an intruder

Raising the minimum cell size, and/or increasing the level of geography are the simplest methods of providing more protection. As the number of sample contributors in a cell increases, it rapidly becomes much less likely that an individual may be identified.

If the sample design includes high sample fractions for some sub-groups (e.g. ethnic groups, areas), then disclosure control should be applied as for a Census or other whole population data source.

Geography. A minimum population size of 50,000 should be used as a general indication of equivalence to Local Authority District geography. Note that while the City of London and Isles of Scilly are defined as Local Authorities, they must be combined with other geographies to meet the minimum size.

Percentages. Percentages may be released provided it is not possible to deduce where only 1 unit has contributed to the cell. For example, small percentages may be rounded to zero. Percentages must be rounded sufficiently depending on the sample size on which they are based. It is essential that the underlying cell counts are not released.

Higher level units. The statistical units on which a table is based may belong to another higher level unit, for example an establishment like a school or prison. If this higher level unit is unique in the population for a combination of characteristics it may be identified by an intruder, e.g. it is the only private school in an area. It may also be possible to identify the establishment in a table where counts of individuals in a cell all belong to the same establishment. Protection then needs to be applied to the higher level unit, as well as the individual.

Cluster sampling. Particular care should be taken with cluster sampling, as the clusters may make identification easier. For example:

- a cluster sample design based on birth dates will be close to a simple random sample except for biological twins and higher multiple births. Then some cells of 2 or more may be a disclosure risk if they consist only of multiple births (e.g. the LS).
- in addition, the LS sample design makes it critical that no sample members are identified. Identification of a very few LS sample members would lead to the discovery of the sample selection birth dates and much easier identification of all sample members
- if clusters are institutions, (e.g. schools), and the output unit is individuals (e.g. pupils) there is a higher risk of disclosure. Identification of the cluster unit may be disclosive in its own right, or may assist the identification of individuals. If all units in the cluster are sampled, for example all persons within a household, then identification of the cluster would identify the units within the cluster.

Linked data. Linked data pose particular issues. The LS is an example where a subsample from one data source (the census) is linked to data from other sources (other censuses, birth and death registers). The linkage is called 'unit record linkage' where records for the same individuals are matched across all the datasets. In longitudinally linked data, individuals are matched over time. Disclosure risks are increased for linked datasets because

- the linked data are often a very rich source of detailed information about individuals
- transitions over time are available from longitudinally linked data and can be very identifying
- if publicly available data (e.g. births) are linked with confidential information (e.g. census) then identification could be made through the publicly available data

For example, LS must provide disclosure control for birth data precisely because there is no confidentiality protection required for the published birth data. If the exact date of birth was published in a table it would make identification of an individual relatively straightforward. Once an individual has been identified in a table many attributes about that person can be determined.

C6. Implementation Issues and Concerns

C6.1 Impact on Quality

In accordance with the CoP, Principle 5: Confidentiality, Practice 4 and as outlined in the CoOS guidance paragraphs 36-40 the implementation will achieve the obligation to protect against disclosure but will through choice of disclosure method strive to give the greatest practical utility possible in the released statistics.

Designing tables to remove unsafe cells should be done considering the needs of the data user in order to minimise the impact of confidentiality protection on the usefulness of published data.

Because unsafe cells are also those cells that are of poorest quality (i.e. have high sample error) the disclosure control methods often have little impact on the detail and quality of data that can be published. The main impact is when researchers wish to remove detailed tables from the safe setting for further analysis within a research team. This situation is best handled by allowing an intermediate stage of access to some disclosive tables for researchers under a licence agreement.

C6.2 Tools

No disclosure control tools are needed.

C6.3 Standard wording

- To inform data users of the confidentiality protection method there is no need for any statements where disclosure control has had no impact on what is published. If required then

  "Tables have been designed to ensure that the confidentiality of *[source data]* respondents has been protected." followed by any necessary specific information on the variable categories combined or variables removed.

- To inform data users of impact on quality of data, state any information that has been lost due to the need to re-design tables state any information that has been lost due to the need to re-design tables.

- Wording of footnote to tables. No particular wording may be necessary. However if variable categories or some cells within the body of the table have been combined to increase cell size, the following wording may be used as appropriate:

  "Variable categories have been combined to protect confidentiality." or "Cells have been combined to protect confidentiality"

  Further detail may be given if needed.

C7. Case Studies

Please see the associated document 'Case Studies for tables produced from survey data'

Appendix D: Guidance for tables produced from Business Surveys

General guidance for disclosure control for tables produced from business surveys

D1.Determining User Requirements

As with social survey outputs tables produced from business surveys are widely used for policy purposes although there is less demand for the public for these outputs.
A wide range of standard outputs are produced along with tables created following ad hoc requests. The main differences between outputs from business surveys and social surveys is that business tables have a magnitude element (in addition to a frequency) where a cell in a table may display a total or average value such as Turnover from a number of businesses. This leads to a further element of disclosure risk which is discussed in more detail in section D3 below.

D2. Understanding the key characteristics of the data and the required outputs

D2.1 Table properties

At the ONS most tabular outputs from business surveys consist of magnitude tables of financial variables (*e.g.* turnover, capital expenditure, sales) or employment. However some financial variables are a net value of two components and may have negative values (*e.g*. capital expenditure = acquisitions- disposals). Outputs can also be tables of counts (number of businesses) which are produced from the whole population data on the Business Register. Categories in these commonly include geography, industry, sector, employment size and product code. The Industry SIC variable may be at a 2, 3 or 4 digit level while the level of geography varies with the size of the survey, some producing only UK level data, through to outputs from the Business Register at small area level. A number of surveys are used to create Indices such as the Average Earnings Index and publication may be Monthly, Quarterly or Annual. Surveys vary in their relationship with the customer with some supporting ad hoc customer requests for tables, while others do not.
As can be seen above many tables are obtained from business surveys carried out by the ONS. In addition Government Departments also carry out business surveys from which tables are published.

D2.2 Sample Design

The Interdepartmental Business Register (IDBR) covers almost the entire population of businesses in the UK. Most business surveys have a similar sample design using the IDBR as the sampling frame. They are generally based on a stratified sample design that includes a full-coverage stratum for the larger businesses. Other strata have differing sample fractions. The businesses in the full-coverage strata are often well known and easily identifiable. Some surveys have different sampling frames. For example the ASHE (Annual Survey of Hours and Earnings) sample is 1% of the working population with the sample coming from Inland Revenue PAYE records. There are also a number of very small surveys targeted to specific industry groups such as financial services.

D2.3 Data Uses

The data are used by the ONS in the production of National Accounts and Balance of Payments and may be provided to Eurostat for combined EU tables. Other published tables have a wide range of uses such as for government policy formulation, allocation of funding, local body planning, by industry groups and academic researchers. Key data users include the Bank of England, Treasury and government departments.

D2.4 Data Quality

Because business surveys carried out for the ONS are compulsory under the Statistics of Trade Act, response rates are generally high for ONS business surveys especially for large businesses.

D2.5 Microdata Access

Government departments (for example Department for Business Innovation and Skills (BIS) and local authorities may be provided with identifiable business survey microdata under specified conditions. Any other researchers may access business survey microdata only within a safe setting environment at the relevant business microdata laboratory.

D3. Circumstances where disclosure is likely to occur and managing this risk

Common variables used to define tables such as geography and industry may allow the identification of prominent businesses, at least in the case of full-coverage strata. Identification could then lead to the values supplied by respondents being revealed, known as *attribute* disclosure. For the magnitude tables typically released by business surveys, attribute disclosure means revealing either the exact values or a close approximation.
The Statistics of Trade Act 9 (5) (a) is interpreted as requiring us not to reveal the exact number of respondents contributing to a cell if that number is less than 5.

D3.1 What intruder scenarios are we protecting against?

There are two types of intruders as defined in section 3- businesses who also contribute to the cell value (scenario 2), and individuals or businesses not contributing to the cell (scenario 1). One obvious motivation for a business attempting to discover a respondent's value is to gain a commercial advantage. In this case it can be assumed that the intruder is generally well-informed on the situation in that sector of the economy and is able to identify the largest contributors to a cell. The intruder scenarios are as follows:

1. Any person or business, not a member of the cell, attempts to identify a cell respondent and deduce the exact value or a close approximation of the response

2. A business contributing to a cell identifies another business contributing to the cell and deduces the exact value or a close approximation of the other's response

D3.2 What parts of the outputs pose a disclosure risk?

D3.2.1 Threshold and p% rules

We need to protect cells where there are either a small number of contributors that would allow exact values to be revealed, or where there are dominating contributors whose values could be revealed to a close approximation. We first assume full coverage

The *threshold* rule states that for a cell to be safe:

there must be a minimum of n contributing units. The Statistics of Trade Act (1947)[11] states that 'no such report, summary or communication shall disclose the number of returns received with respect to the production' of any article if that number is less than five'[9]

However this is a necessary but not sufficient condition. We wish also to prevent a unit contributing to the cell from finding out the value of another unit in the cell to within a certain approximation. The approximation is defined as p% of the true value. This leads to the definition of unsafe cells using the *p% rule* for scenario 2. The p% rule also provides protection against scenario 1 since in this situation less information is available to the intruder. The p% rule states that for a cell to be safe,

> the total of the cell minus the largest *m* contributor(s) must be greater than or equal to p% of the value of the largest.

D3.2.2 Samples.

For samples, the cell total, *T*, in the p% rule is a weighted sum of responses. The values of the two largest contributors remain the original unweighted values.

D3.2.3 Units.

Units for business surveys may be Enterprise Groups, Enterprises, Reporting Units or Local Units.

An Enterprise Group is a group of legal units under common ownership. Each group will have their own decision making procedures and can comprise a number of Reporting Units (a grouping of local units). An example could be a large supermarket group (Enterprise Group) made up of a number of smaller units (food, electrical goods, insurance).

The enterprise is the smallest combination of legal units that is an organisational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making,

The threshold rule is applied at Enterprise Group level to prevent disclosure of the exact value supplied by the whole Enterprise Group. This also protects the responses of all lower units belonging to the same Enterprise Group.

The p% rule is applied at Reporting Unit level and prevents one Reporting Unit contributing to the cell finding out the value of another Reporting Unit in the cell to within p%.

D3.2.4 Other information available

The CoP, Principle 5 implies that we must take account of other information likely to be available to third parties. Information about businesses is often widely available through advertising, and many other means. The Companies House data is a list of businesses available to the public which is used as a secondary source for compiling the IDBR. (The main sources for compiling the IDBR, the VAT trader system and PAYE employer system, are held securely by HM Revenue and Customs.) This extensive public knowledge of the existence and general characteristics of businesses means that identification is often likely to be correct. Another main source of information is other tables released from the same data source. Other tables may lead to disclosure in several ways:

> **Averages and percentages.** If tables are released as averages, rates or percentages the numbers by themselves may not be disclosive. However when combined with other tables it may be possible to recover the totals on which they are based, which may be disclosive. An average, rate or percentage calculated from an unsafe cell is considered unsafe if the original total can be recovered.

[11] See http://www.legislation.gov.uk/ukpga/Geo6/10-11/39/contents

Disclosure by differencing. Disclosure by differencing can occur if two similar tables are released and subtracting the values of one from the other would reveal disclosive information. For example consider a table produced that includes SIC Subsection DA: manufacture of food products, beverages and tobacco. If a second table is released for Subsection DA, but excluding Class 15: manufacture of food products and beverages, the value of respondents in Class 15 would be found by differencing the two tables. If this cell would have failed one of the standard safety rules then disclosive data will have been inadvertently released.

Linked tables. Linked tables are tables derived from the same microdata where some of the cells are in common. For example a table of geography and industry will have the same area totals as a table of geography by size. Care must be taken to ensure that the protection provided in one table is not "undone" by common cells released in a linked table.

D3.2.5 Counts

There is a requirement under the Statistics of Trade Act to not reveal the exact number of respondents contributing to a cell if that number is fewer than 5. Thus cell counts of fewer than five are unsafe in tables of counts of businesses, and for magnitude tables counts of the number of contributing respondents are also unsafe if fewer than five. In both cases, any counts of the number of businesses may be unsafe because knowledge of the exact number of respondents can make it easier to determine respondent values in magnitude tables.

D3.3 What is not a disclosure risk?

The disclosure control rules recognise links within Enterprise Groups, since these are defined on the IDBR, but no other less formal relationships such as franchises. The risk of businesses other than those within Enterprise Groups combining their information is considered to be low, and is an acceptable disclosure risk. No further protection is necessary.

Zero cells do not normally create a disclosure risk in sample data as they may for the population data. Zeros in population data allow one to say that all others in a row or column do not fall into the category of the zero cell. A zero value from a small sample does not allow one to infer that no one in the population has those characteristics. If the sample is large or there is complete coverage of the strata then some zeros could potentially be disclosive. However this risk is considered low.

D4. Does the disclosure risk identified constitute a breach of statistical obligations?

The relevant legislation and protocols are discussed in Appendix A.

D5. Selecting Disclosure Control Rules and Methods

D5.1 Requirements of the standard

Magnitude tables. For magnitude tables, the threshold and p% rules define unsafe cells. A cell meeting both the following criteria is non-disclosive.

1. There must be at least *n* enterprise groups in a cell (threshold rule).

2. The total of the cell minus the largest *m* reporting unit(s) must be greater than or equal to p% of the value of the largest reporting unit (p% rule)

Note that the values of the p% and minimum threshold parameter *n* should remain confidential, since knowledge of these values reduces the protection.

Table design should be used first to reduce the number of unsafe cells in a table where this is consistent with the main uses of the data.
Cell suppression is the standard method used to protect tables with unsafe cells. The unsafe cells are suppressed, known as *primary suppressions*. Other cells must be suppressed to prevent the values of the unsafe cells being calculated by subtraction from the marginal totals of the table. These are known as *secondary suppressions*. Cell suppression could be implemented using the Tau Argus software (see Section 5 Implementation and Evaluation) or similar software.

Cell suppression does not generally provide protection from disclosure by differencing. Tables should be published using fixed categories to avoid disclosure by differencing. For example the same standard geographies and SIC codes should always be used.

Count tables. Tables of count data are to be protected by controlled rounding to base 5. Controlled rounding introduces ambiguity to prevent exact counts being derived because a rounded value of 0 could be any value in the range [0,5], and a rounded value of 5 could be any value in the range [1,9]. Wider ranges can be achieved under particular conditions. Controlled rounding to base 5 provides good protection against disclosure by differencing and for disclosure by comparison of common cells in linked tables. The uncertainty for a true value of small count may be reduced in these cases to [1,4] but this may still provide enough uncertainty for confidentiality protection.
Controlled rounding may be implemented using Tau Argus or other similar programs.

D5.2 Guidance for applying the standard

The level of p. The standard level of the p% value used protects against an intruder who attempts a straightforward calculation of a competitor's response as described above. An intruder who is able and willing to use more sophisticated techniques could potentially estimate a respondent's value to closer than the nominal p% protection. One motivation could be that the value of a competitor's response would provide the intruder with a large commercial advantage. If a survey considers that the risk of sophisticated methods being used to obtain individual respondent data outweighs the value of the data to legitimate users, then a higher value of p may be used.

Ad hoc requests and other linked tables. Cell suppression is applied to standard published tables. If there are common cells between standard published tables, then secondary suppressions should be consistent. A cell used as a secondary suppression in one published table should also be suppressed if it appears in another published table.

When additional ad hoc tables are released to customers, it is difficult to ensure that cell suppressions are consistent with all previously released tables. Where common cells exist in linked tables, the same primary suppressions will result, but secondary suppression patterns may be different. New ad hoc releases should be checked against the standard published tables to ensure secondary suppressions are not revealed. Consistency with other ad hoc releases should be considered unless the resources required are unreasonable.

Units for the p% rule. Application of the threshold and p% rules at higher levels also guarantees protection at lower levels. Thus the threshold rule applied at Enterprise Group level also protects the values of all Reporting Units and local units. However the reverse is not true. A cell that is disclosive at Reporting Unit level will also be disclosive if the p% rule is applied at Enterprise Group level, but it is not true that a safe cell at Reporting Unit level will always be safe at Enterprise Group level. There remains the possibility that Reporting Units from the same Enterprise Group could combine their results and obtain accurate estimates of another Enterprise Group from the same cell. If this is considered to be a serious disclosure risk, then the p% rule could be applied at Enterprise Group level.

Derived values. Derived data such as averages, percentages and rates should be based on protected data. Thus an average must be suppressed if the total from which it is derived has been suppressed.
For count tables, percentages or rates must be derived from rounded values. No further protection is needed.

Negative numbers. These can occur in some magnitude tables. An example are tables of Foreign and Direct Investment where negative numbers can be reported if the value of imports from a country exceed the value of exports to that country   Where negative numbers appear in magnitude tables, methods for identifying disclosure risks and protecting the data have to be modified. Please contact the ONS Statistical Disclosure Control branch, Strategy and Standards Directorate for more information.

D6. Implementation Issues and Concerns

D6.1 Impact on Quality

In accordance with the CoP, Principle 5: Confidentiality, Practice 4 and as outlined in the CoOS guidance paragraphs 36-40 the implementation will achieve the obligation to protect against disclosure but will through choice of disclosure method strive to give the greatest practical utility possible in the released statistics.

Protection of confidentiality through cell suppression can lead to a high loss of information. Up to three secondary suppressions are needed to protect one unsafe primary suppression and can lead to many suppressed cells. Large businesses that dominate an industry are often a disclosure risk and are suppressed, but also contribute to high quality estimates when they are part of full coverage strata.

Loss of information can be minimised by

- design of tables to reduce the number of unsafe cells by combining categories.
- optimising the cell suppression to minimise the total value of cells suppressed
- gaining written expressed consent from businesses to allow publication of critical cells

D6.2 Tools

The Tau Argus software is one possibility for applying both primary and (near optimal) secondary suppression along with controlled rounding. A link with SAS is under development.

D6.3 Standard wording

To inform data users of the confidentiality protection method:

Cell suppression. "Cells have been suppressed to protect confidentiality. The confidentiality of respondent information is protected by suppressing cells that are unsafe, known as *primary suppressions*. Other cells must also be suppressed to prevent the values of the unsafe cells being calculated by subtraction from the marginal totals of the table. These are known as *secondary suppressions*. "

Controlled rounding. "Cells have been rounded to protect confidentiality. Controlled rounding to base 5 has been used. Controlled rounding means that cells are rounded up or down to the adjacent multiples of 5 in a way that retains the additivity of tables. For example, an original value of 23 will be rounded to either 20 or 25, and rounded values in a row or column always add up to the rounded row/column total. Original cell values of zero or multiples of the base are unchanged. Values may be rounded down to zero and so all zeros are not necessarily true zeros."

To inform data users of the impact on quality of data:

The impact on the data quality will vary depending on the table. Avoiding unsafe cells by the design of tables may reduce the detail available to data users, for example by combining variable categories or restricting the level of geography or industry. Protection of unsafe cells through cell suppression results in the complete loss of the information in the suppressed cells. Other published cells remain unchanged. If Tau Argus is used for secondary suppression, cells are chosen for secondary suppression in a way that minimise a cost function chosen by the user, e.g. the total value lost.

Controlled rounding perturbs the original cell values. A cell value will change by between 0 and 4, either added or subtracted to the original value. The table remains additive, and the error on the total is controlled so that totals also do not change by more than 4 from the original values. For large cell values the relative difference between original and rounded values is very small. Rounding has most impact on data quality for small original cell values where the relative change may be large. Controlled rounding is optimised to achieve the smallest change from all original values while still retaining additivity. Occasionally on very large or difficult tables, some cells may change value by more than 4 to achieve this additivity.

Wording of footnote to tables where suppression has been used:

Cell suppression: " [symbol] Cells have been suppressed to protect confidentiality"
‘c’ is the standard symbol used.

Controlled rounding: "Cells have been rounded to base 5 to protect confidentiality. The rounding is controlled so that the table remains additive."

D7. Case Studies

Please see the associated document ‘Case Studies for tables produced from survey data’

Appendix E. Summary

This guidance outlines the issues concerned with protecting the confidentiality of statistics resulting from social surveys, subsamples and business surveys and describes an approach for ensuring that the public interest in the use of the figures is met while managing data disclosure risks. It also spells out the main steps that a data provider will consider in order to develop specific confidentiality rules for different types of statistics.
If any problems arise when applying statistical disclosure control to tabular outputs from administrative data please contact
Sdc.queries@ons.gsi.gov.uk

References

Elliot, M. J., and Dale, A. (1998) Disclosure risk for microdata: Workpackage DM1.1 What is a key variable? *Report to the European Union ESP/204 62/DG III*

Elliot, M. J., and Dale, A. (1999) Scenarios of attack: The data intruder's perspective on statistical disclosure risk. *Netherlands Official Statistics. Vol 14, Spring 1999, 6-10.*

UK Statistics Authority Code of Practice for Official Statistics
http://www.statisticsauthority.gov.uk/assessment/code-of-practice/index.html

National Statistician's guidance: Confidentiality of Official Statistics
http://www.statisticsauthority.gov.uk/national-statistician/guidance/index.html

Respondent Charter for Individual and Household Surveys
http://www.ons.gov.uk/ons/about-ons/get-involved/taking-part-in-a-survey/information-for-households/respondent-charter---households-and-individuals/index.html

Respondent Charter for Business Surveys
http://www.ons.gov.uk/ons/about-ons/get-involved/taking-part-in-a-survey/information-for-businesses/respondent-charter---business-surveys/index.html

Data Protection Act (1998)
http://www.legislation.gov.uk/ukpga/1998/29/contents

Tau Argus link
http://neon.vb.cbs.nl/casc/tau.htm

Statistics and Registration Service Act (2007)
http://www.legislation.gov.uk/ukpga/2007/18/contents

United Nations Guidance on the Production of Official Statistics
http://unstats.un.org/unsd/goodprac/bpaboutpr.asp?RecId=6

Statistics of Trade Act (1947)
http://www.legislation.gov.uk/ukpga/Geo6/10-11/39/contents