Checklist on Disclosure Potential of Proposed Data Releases

Prepared by

Interagency Confidentiality and Data Access Group
An Interest Group of the Federal Committee on Statistical Methodology

Statistical Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

July 1999

File: final_toc&body_rev2.doc
Current date: 10/1/99

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Checklist on Disclosure Potential of Proposed Data Releases

Table of Contents

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Checklist on Disclosure Potential of Proposed Data Releases

Section 1: Introduction

Federal statistical agencies and their contractors often collect data from persons, businesses, or other entities under a pledge of confidentiality. Before disseminating the results as either public-use **microdata files**[1] or tables, these agencies should apply statistical methods to protect the confidentiality of the information they collect. A review and evaluation of the statistical disclosure limitation techniques used by Federal statistical agencies can be found in the Federal Committee on Statistical Methodology's 1994 report, *Report on Statistical Disclosure Limitation Methodology* (Statistical Policy Working Paper [SPWP] # 22). In addition, SPWP # 22 contains a set of 12 recommendations to improve disclosure limitation practices.

One of the recommendations in SPWP # 22 is that agencies should centralize their review of disclosure-limited data products. In discussing this recommendation, SPWP # 22 suggests that if the number of programs is small, such a review could be handled by one individual; alternatively, if an agency has multiple or large programs, a review panel, team, or board might be needed. In this document, the term **Disclosure Review Board** is used to refer to the formally or informally designated unit or individual that handles such review. The attached document, "Checklist on Disclosure Potential of Proposed Data Releases" (called **Checklist**), is one tool that can assist agencies in reviewing disclosure-limited data products. Completed Checklists should be submitted to the Disclosure Review Board for review.

Most agency data products are intended for **public use**, with no restrictions on eligibility and intended use. Products that meet the criteria for public release may not have sufficient detail to satisfy the analytical requirements of all users. Consequently, some agencies have developed **restricted access** procedures for making more detailed microdata files and tables available to some users, subject to conditions of eligibility, location of use, purpose of use, security procedures, and other features associated with access to the data. *This Checklist is intended primarily for use in the development of public-use data products*. Some of the disclosure limitation procedures described in the Checklist may be of value in preparing data products for restricted access; however, the procedures may have to be relaxed to some degree to meet users' analytical requirements. The Interagency Confidentiality and Data Access Group (ICDAG) plans to develop additional documents (perhaps including another checklist) for use in developing arrangements for restricted releases of microdata files and tables. Pending availability of these documents, agencies may wish to consult a 1993 article by Jabine which summarizes restricted access procedures in use at that time.

The Checklist consists of a series of questions that are designed to assist an agency's Disclosure Review Board to determine the suitability of releasing either public-use microdata files or tables from data collected from individuals and/or organizations under an assurance of confidentiality.

---

[1] A **microdata file** consists of records at the respondent level. Each record contains values of variables for a person, household, establishment, or other unit.

NOTE: *Responses to the questions in this Checklist must be treated as strictly confidential.*

Section 4 pertains to microdata files that contain information from individuals or establishments, while Sections 5 and 6 refer to tabular data from individuals and establishments, respectively. This Checklist is based on one used at the U.S. Bureau of the Census. In creating its Checklist, ICDAG has liberally borrowed descriptions and definitions from SPWP # 22.

Uses of the Checklist

The Checklist was developed with following in mind:

It should be completed by a person who has appropriate statistical knowledge and is familiar with the microdata file or tabular material in question (i.e., branch chief, survey manager, statistician, or programmer). While this implies a considerable familiarity with survey and statistical terminology, those without such background will nonetheless be able to understand much of what it intends to accomplish. (Those who need a "primer" on statistical disclosure limitation methods should see Chapter 2 of SPWP # 22. Other references can be found in Section 6 of this Checklist.)

Responses to questions in the Checklist are not intended to supply all of the information required by a Disclosure Review Board before a microdata file or table is released to the public. Some additional questions may need to be answered and/or given special consideration. Nonetheless, if files and tabular material are reviewed with the aid of the Checklist early enough, the need for time-consuming and costly re-programming of the data to be released can be avoided. This allows additional time for coordination with collaborators and/or other potential users.

In addition to helping an agency's Disclosure Review Board determine the disclosure potential of proposed data releases, the Checklist has other uses:

It can serve an important educational function for program staff who complete the Checklist.

It can provide documentation when an agency is considering release of related data files and tabulations.

It can be very useful in defending legal challenges to an agency's decision to withhold certain tabular data or restrict data contained on a public-use file.

The Checklist reflects the current standards of the Census Bureau and the National Center for Health Statistics for the release of public-use data. The Checklist is not a static document but a "work in progress" that will be changed, refined, and modified as new approaches and techniques are developed. With appropriate modifications, the Checklist can be adapted by Federal agencies and other organizations and used by them to review materials of varying levels of confidentiality. ICDAG encourages agencies to modify this document to suit their particular needs.

Brief Overview of Contents

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Section 2. Cover Sheet: This asks for basic information about the proposed data release.

Section 3: Microdata Files

Most microdata files contain data collected from persons or households (referred to as **demographic data**). *Some questions in this section may not be applicable for establishment-based files.*

A major part of this section of the Checklist focuses on geographic information because it is the key factor in permitting inadvertent identification. In a demographic survey, few respondents could likely be identified within a single State, but more respondents -- especially those with rare and visible reported characteristics -- could be identified within a county or other geographic area with 100,000 or fewer persons.

The risk of inadvertent disclosure is higher with a publicly released data set that has both detailed geographic variables and a detailed, extensive set of survey variables. The risk is also often a function of the quality and quantity of "auxiliary" information (data from sources external to the data being released). This auxiliary information may be difficult to assess for its disclosure risk. "Coarsening" a data set by dropping survey variables, collapsing response categories for other variables, and/or introducing statistical perturbation, called "noise," to the data are techniques that may reduce the risk of inadvertent disclosure (Kim and Winkler, 1995).

For surveys of establishments, the issues are generally different because such entities are often selected from very skewed populations. For example, in the U.S., there are very few hospitals with 1,000 or more beds, and inadvertent disclosure in a survey of hospitals might be possible using detail on the number of beds and geographic information as large as a Census region.

Section 4: Tabular Data from Persons or Households ("Demographic Data")

This section pertains to tables based on data collected from persons or households under a pledge of confidentiality. Tables can be of two types. Tables of **frequency count data** show the number in the population with certain characteristics or, equivalently, the percent of the population with certain characteristics. Tables of **magnitude data** present the aggregate of a "quantity of interest" over all units in the cell. Equivalently, the data may be presented as an average by dividing the aggregate by the number of units in the cell. Demographic data are typically reported as frequency count data.

Section 4 of this Checklist should always be completed if the tabulations are based on a complete count or an enumeration of the target population. Its use should also be considered when:

the tabulations identify small geographic areas, e.g., areas with populations less than 100,000, or
a large sampling fraction was used, as in the case of the decennial census long-form sample, or
the tables have a large number of dimensions or cells, or
the tables cover especially sensitive topics.

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Section 5: Tabular Data from Establishments or Other Types of Organizations

This section pertains to tabular data collected from organizations under a pledge of confidentiality. As with demographic data, tables can be of two types. Tables of **frequency count data** contain the number of units in a cell. Tables of **magnitude data** present the aggregate of a "quantity of interest" over all units in the cell. Thus, a table of the number of establishments within the manufacturing sector by industrial classification group is an example of the former, whereas a table that presents the total value of shipments for the same cells is an example of the latter. Different statistical disclosure limitation methods can be used depending on the type of data being presented, although, for practical purposes, entirely rigorous definitions are not necessary.

Completing the Checklist

Users should complete the cover sheet and answer all questions for the applicable section(s). (Obviously, if the Checklist is distributed as a paper document, those who need more space for an answer can attach a continuation sheet and identify the number of the question.) The completed document should be submitted to the Disclosure Review Board.

Keeping Responses to the Checklist Confidential

Agencies need to walk a fine line between giving data users enough information so that they will have some idea of how the statistical disclosure procedures that were used might affect their analyses and giving them so much information that the data are made more vulnerable to an attacker. In order to release as much information as possible at an acceptable level of disclosure risk agencies should

describe the kind(s) of procedures used to protect the confidentiality of data but not reveal the specific value(s) of the particular disclosure limitation method.

Complete knowledge of a specific statistical disclosure method and its associated parameter value(s) could be used by an attacker to identify one or more individuals on a microdata file or infer the value of table cells that have been suppressed to protect the confidentiality of that information. Because this Checklist contains explicit details on disclosure method(s) and parameter value(s), responses to it should be kept confidential.

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Section 2: Cover Sheet

SURVEY TITLE: _____
DATE: _____

Project Manager's Name: _____
Division and/or Branch: _____
Phone: _____

Is this survey sponsored/co-sponsored by another agency?
Yes. *Please list name(s) of agency(agencies).* **_____**
_____
No.

What type of data are you releasing?
Public-use microdata file. *Please attach the proposed layout and content of the microdata file.*
Tables.

When were the data collected? _____

Does(Do) the reference period(s) of the data collection differ from the actual date of collection?
Yes. *Please give reference period(s).* **_____**
No.

What is the periodicity of the proposed data release?
This is a preliminary release.
This is a one-time release of a public-use microdata file from a one-time collection of data.
This is a release of a special tabulation.
This is one in a series of releases (either microdata file or tables) with substantially the same content. *Please specify the interval at which future products will be released or prior products have been released.* _____
This is the re-release of an approved product, with the addition of supplemental or previously unreleased data. *Please give the date the original product was submitted.*
_____

(NOTE. If this is a re-release of a previously approved product, then only complete those Checklist questions for which the answers are now different.)

Will there be other data release(s) (either microdata files or tables) from this survey?
Yes.
No.

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Section 3. Microdata Files

Geographic Information on the File

Identify the variables on the file used as geographic identifiers and the minimum population size for each geographic area. Generally, an agency has to balance the level of detail for non-geographic variables against the level of geographic detail. This implies that a significant increase in the former will necessitate the latter being decreased, i.e., the geographic detail would have to be coarsened.

List all geographic identifiers to be released and the minimum population of each identifier: _____
_____
_____
_____
_____

General Rule (used by the Census Bureau and National Center for Health Statistics): All geographic areas identified ***must*** have at least 100,000 persons in the sampled area (according to latest Census or Census estimate).

Have you chosen to adopt the above rule or another?
Yes, will use the rule of 100,000.
No, will use other rule. Please specify and provide rationale: _____
_____
_____

Care should be taken before releasing the primary sampling units (PSU's) that are in a sample. In addition to explicit geographic identifiers on the file, the data items, record identifiers, or file structure may provide additional geographic information by inference. Therefore, steps must be taken to avoid inadvertently identifying geographic areas that do not meet the specified minimum population criteria. Potential problem areas are discussed below.

PSU or other geographic information is often embedded in control numbers designed for internal use. For instance, consider the following two hypothetical samples from a county that had a population of one million. If an agency selected its sample from only one PSU with a population of 50,000 or, alternatively, selected its sample from several PSU's whose total population was less than 100,000, then the identity of the county could possibly be inferred. In neither case should the identity of the PSU(s) or control numbers be released.

How will this problem be avoided on the released file?
Control numbers are deleted or do not contain geographic information.
Control numbers are scrambled. Please describe: _____
_____
_____

NOTE: *Responses to the questions in this Checklist must be treated as strictly confidential.*

Other. Please describe: _____

_____

Records in many data bases are sequenced so that the first cases are in the lower numbered PSU or county that is first in alphabetic order.

Briefly, describe how the records on this file will be sequenced to avoid such geographic inferences: _____

_____

_____

Data items that imply specific geography of residence may reveal more than the explicit identifiers displayed. Some examples follow: duration of residence codes revealing State of current residence ("lifetime" or "always" where number of years is equal to the age of respondent); a migration code specifying movement from a metropolitan area to a nonmetro area when metro-nonmetro status has been excluded; residence within X miles of a nuclear reactor or an airport or health care provider when there is only one in an identified geographic area; telephone area code; or latitude and longitude coordinates; codes that indicate the existence of a particular service/utility (such as well water, septic tanks, cable TV, for example) where only a small area has (does not have) this type of service.

List all items that will be deleted for this reason: _____

_____

Identify other geographically-related variables (e.g., center city, non-center city, metropolitan area, non-metropolitan area) on the file.

List all items that you think might have geographic significance, but could not decide if they should be deleted: _____

_____

_____

_____

_____

Sampling information may also provide some geographic indicators. For example, certain sampling weights may distinguish between self-representing and nonself-representing PSU's or identify types of areas intentionally oversampled. Also, codes for "second stage units," "hit number," etc., may be related to geography.

(a) List all sampling information -- including that for variance estimation -- that will be deleted for confidentiality reasons or subsampling plans to make weights less identifying:

_____

_____

_____

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

_____

(b) List all other sampling information that you think might have geographic significance, but could not decide if it should be deleted:

_____
_____
_____
_____

File Contents Presenting an Unusual Risk of Individual Disclosure

The disclosure criteria for public-use microdata require a review of each file to determine if any of the proposed contents present an unusual risk of individual disclosure. The Disclosure Review Board has identified several measures that can reduce the possibility of identifying an individual through the characteristics available on a file. The measures are discussed below, and relevant information pertaining to the proposed file is requested to assist the Disclosure Review Board in its review.

Names, addresses, and other unique numeric identifiers such as Social Security, Medicare, or Medicaid numbers ***must*** be removed from the file.

High income is a visible characteristic of individuals or households and is considered to be a sensitive item of information. Therefore, each income figure on the file, whether for households, persons, or families, including total income and its individual components, should be **topcoded**.

There are no hard and fast rules for determining which cutoffs to use in topcoding. Decisions should be based on examination of the structure of the distribution, in combination with other key variables like race, gender, etc. For example, one rule used at the Census Bureau is to topcode at least the top ½% of the non-zero values. Note that the strict use of the same criterion could result in changing the cutoff from year to year, which would make things very difficult for data users. One suggested solution would be to change the cutoff only when there has been a substantial change in the upper tail of the distribution. Before making such a change it is important to take into account how the proposed change will affect time series analyses.

Certain special cases require more thought when rules for topcoding are being developed. For example, consider a variable with a high proportion of zero values for most of the population (such as welfare income). As the proportion of non-zero values decreases, it may be desirable to topcode in such a manner that a higher proportion of them are above the cutoff. Be aware that a data base containing rare and unusual details on race and ethnicity may be a problem, unless there is little geographic detail. In addition, data bases that contain "unusual" subgroups may need special attention (for instance, high-income persons who pay no taxes). In developing topcode rules, it might be prudent to discuss alternatives with the Disclosure Review Board well in advance of the final submission for approval to release a file.

NOTE: *Responses to the questions in this Checklist must be treated as strictly confidential.*

Please describe the topcoding rule that is used. If you have different rules for different income variables, please give details. _____

_____

_____

_____

_____

Do all income topcodes satisfy the appropriate rule(s)?

Yes.

No. If not, please specify the percent topcoded and the topcode amount. Briefly summarize discussions with the Disclosure Review Board:_____

_____

_____

_____

_____

In addition to income, certain other characteristics may make an individual more visible than others. Some examples include: unusual occupation (as revealed by coding to 3 digits); unusual health condition (e.g., as shown in highly detailed International Classification of Disease codes); very high age; value or purchase price of own property; rent or amount of mortgage. Depending on the geographic detail shown on the file, consideration should be given to topcoding (and/or collapsing) these items when they are represented as interval or ordinal variables. One rule of thumb suggested by the Census Bureau's Disclosure Review Board is that these topcode categories include at least ½ of 1 percent of the total universe (persons/households) represented on the file (weighted counts).

In a few cases, where variables apply only to very small populations, the Disclosure Review Board may consider topcode categories, including approximately 3 to 5 percent of the appropriate subpopulation. Examples of approved topcodes used at the Census Bureau include the following:

Age -- 85 years old and over. (Approximately 1.2% of all persons in the 1990 census.)
Value of property -- $500,000 or more. (Approximately 0.7% of all units, not just owner-occupied units in the 1990 census.)
Gross Rent (including utilities) -- $1,000 or more. Approximately 1.2% of <u>all</u> units, not just renter-occupied units in the 1990 census.)
Payments on mortgages -- $1,000/month (Approximately 3.0% of all mortgage holders on the 1984 Survey of Income and Program Participation file.)

In addition, some variables may require bottom-coding, such as year of birth before 1914 or large negative value for income.

(a) List all items that will be bottom- or topcoded (or collapsed) and the corresponding codes:

_____

***<u>NOTE</u>**: Responses to the questions in this Checklist must be treated as strictly confidential.*

_____
_____

(b) List all other items about which you have questions regarding the need to bottom- or topcode:

_____
_____
_____

Describe any proposed information to be released for the bottom- or topcoded data items (for example, means or medians of the coded values):

_____
_____
_____

Depending on the amount of geographic detail on the file, there are other characteristics that may make a person highly visible. These typically are represented as categorical or nonordinal variables and, therefore, cannot be topcoded . Some examples include the following: codes indicating Foreign or Indian Tribal language spoken; detailed racial identification such as Eskimo, Aleut, Guamian, or Samoan; detailed ethnic origins; codes for place of prior residence; codes for tenure in the area ("Always," "Lifetime"). In these cases, the amount of detail on the file may have to be collapsed into larger categories.

(a) List all items that will be collapsed (or deleted) for confidentiality reasons:

_____
_____
_____

(b) List any other items about which you have questions regarding the need to collapse the detail:_____

_____
_____
_____

Contextual or Ecologic Variables

Contextual or ecologic variables are those that describe some aspect of an area, such as a State, county, census tract, or block group; percent or frequency of the area's population employed, foreign born, receiving public assistance; number of health facilities; number and specialty of physicians; local government expenditures; measures of air quality; etc.

Identify the source(s) of the contextual variables: _____

_____
_____
_____

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Identify any contextual variables and the levels at which they are coded:

_____
_____
_____

List all contextual variables that will be collapsed (or deleted) for confidentiality reasons:

_____
_____
_____

List any other contextual items about which you have questions regarding the need to collapse the detail:_____
_____
_____

Does the cross tabulation of all contextual variables and geographic variables lead to the identification or one or more geographic areas that are not supposed to be identified on this file?

Yes. If yes, some, variable(s) *must* be removed and/or collapsed. Please describe what will be done. _____
_____

No.

Don't know/did not check. Please explain. _____
_____
_____

Disclosure Risks Associated with Administrative Data and Other External Data

Efforts must be made to reduce the potential for matching microdata on this file to data on external files because external files usually contain names and addresses and, thus, can be used to identify survey respondents. Such matching may be possible if the survey contains highly specific characteristics also found on mailing lists or administrative records maintained by other agencies or organizations. For example, the inclusion of vehicle make, model, and year in conjunction with specific geographic identifiers is unacceptable because these items can be matched to automobile registration lists that contain names and addresses. These items probably could be left on the file if they were recoded into broad categories. In addition to the external files mentioned above, other potential source of such files include: manufacturer's list of purchasers of particular major durable goods (for example, airplanes); voter registration lists in some states; Federal, State, or local tax records; criminal justice system records; state hunting and fishing license registers; and membership rosters of certain trade associations.

Disclosure risk is also high if the sampling frame for a survey comes from a source outside the agency or if the file contains information obtained from other agencies. In such cases, the agency that provided the sampling frame or the auxiliary information may be able to match

survey records to its original records, particularly if survey records include data from the originating agency's files: e.g., amount of program benefit received, date of entry into program.

External Files Matchable to Proposed File

Were any of the sample cases contained in the microdata file selected from a list provided by an outside source?

Yes. Please identify the source, and describe how and by whom sample cases were selected from the list: _____
_____
_____

No.
Don't know.

Were any administrative data or data from another external sources used to "expand" the content of this microdata file (e.g., merging administrative data with survey data)?

Yes. Please identify the source, and describe the variables obtained from the external source(s): _____
_____
_____

No.
Don't know.

Are you aware of administrative records, research files, or a mailing list that contains data also included in this proposed microdata file? Such external data could be used to "compromise" your data, that is, they pose a risk because an "attacker" would use these external data sources to deliberately try to identify individual respondents and gain access to their confidential responses.

Yes. Please identify: _____
_____
_____

No.
Don't know/did not check.

Based on available information, will any data item on the public use microdata file identify residence in a particular type of institution of which there may be only one in an identified area; or for which a system of records could be obtained?

Yes. Please identify the type of institution: _____
_____

No.
Don't know.

Matching

When an external file related to the proposed file to be released exists, several steps may be taken to reduce the possibility of matching survey data to this file; for example, selected items

may be deleted or recoded, or "noise" (i.e., small amounts of random variation) may be introduced into these items.

The Disclosure Review Board cannot specify in advance exactly which steps must be taken to sufficiently reduce the potential for matching. However, it does consider several factors in determining the risk associated with releasing a file when the possibility of matching to external data bases exists: 1) the number of variables available for matching purposes: 2) the resources needed to perform the match; 3) the age of the data; 4) the accessibility, reliability, and completeness of the external file; and 5) the sensitivity or uniqueness of the data. Some factors that make matching easier are listed below, and information is requested on steps that will be taken before the file is released to reduce the matching potential. (*NOTE: This information is necessary even if you are not aware of any external files that could be used in matching.*)

Matching is easier:

...if any data item, or combination of items, isolates a small and readily identifiable population subgroup or class. The inclusion of codes that identify very small population segments should be avoided; for example, Indian tribes or detailed occupation groups in combination with the release of highly specific geographic identifiers. Normally, one has to consider more than one variable at a time if a group of variables is likely to appear together on a file or list. For example, age and sex, together with country of birth and occupation, could permit the disclosure of individual identity.

(a) List all data item(s) proposed for inclusion on the file that, in various combinations, may isolate a small, readily identifiable population:_____

_____

_____

(b) List all data item(s) that will be altered (i.e., deleted, recoded, noise added) for this reason:_____

_____

_____

...if the file includes essentially every member of a population (say $p > 0.5$). Examples include: establishments/institutions with large numbers of employees, high-income individuals, doctors, scientists of a specified type, or residents of certain types of institutions. In such instances, prior to data release, it may be appropriate to do additional subsampling.

Identify these populations, if any are on the file, and how they will be subsampled:

_____

_____

_____

...if the file contains any information obtained from records or other sources where that information could serve as a link to an external file which has individual identifiers or detailed geographic information. Examples include: fuel consumption or cost records

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

from a utility company; neighborhood, tract, or summary characteristics from a decennial census; welfare, health-related, or Social Security data from a Government agency; arrest record from a police department; benefits provided to employees, such as pensions and health insurance.

(a) List all data item(s) proposed for release on the file that were not obtained from an interview with the respondent:_____
_____
_____
_____

(b) List all data item(s) altered or deleted for this reason:_____
_____
_____
_____

...if the file includes data items frequently used for matching, such as exact date of birth, sex, and race, or if it includes other items that should be identical on both files, such as exact income amount, real estate taxes or other taxes, or date of entry or termination from a Government-sponsored program.

(a) List these data items, if any: _____
_____
_____
_____

(b) List all data item(s) altered or deleted for this reason: _____
_____
_____
_____

...if longitudinal data are being collected; i.e., if the data for the same respondents/units will be collected for several different reference periods. Primary concern relates to time series of data items potentially matchable to outside records; e.g., income tax or employment records.

If data are collected from the same respondents more than once, indicate the frequency of interview, length of time that any one unit may be in the sample, and factors affecting the likelihood of matching a sample unit from one time period to the next:
_____
_____
_____

...if highly specific geography is included on the file, such as State, Metropolitan statistical area, etc.

NOTE: *Responses to the questions in this Checklist must be treated as strictly confidential.*

List all geographic identifiers below the level of region: _____
_____

...if data collected from multiple persons in a household are linked on the released file.
Disclosure risks associated with linking of household members are well-known. For example, households can be identified because of significant difference in spouses' ages, atypical number and ages of children, a "unique" multi-racial composition of the household, etc. -- not to mention the fact that one household member, by self-identification, could look up other members' reported information.

(a) Are data collected from multiple persons in a household?
Yes.
No. *If no, skip to 3.3.3.*

(b) If yes, describe the strategy for releasing these data, and indicate whether or not the data from these will be linked: _____
_____

Describe any considerations not previously mentioned that <u>reduce</u> the ability to match this file to external data; e.g., unreliability or natural noise in the data:
_____
_____
_____

Cross-Tabulations To Identify Unique Sets of Characteristics

Were any cross-tabulations performed to identify sets of unique characteristics?
Yes.
No. *If no, skip to 3.4.*

If yes, what were the results? _____
_____
_____
_____

Will any additional steps be taken to reduce disclosure risk based on these results?_____
_____
_____
_____

The Addition of Statistical Perturbation (or "Noise")

The addition of statistical perturbation, called "noise," is another statistical disclosure limitation technique. Essentially, "noise" is defined as the addition of small amounts of random variation

to quantitative data (see Kim and Winkler, 1995). There are several methods that can be used to add noise to data.

Was any noise added to the data?
.Yes.
No. *If no, skip to 3.5.*

What procedure(s) was(were) used to add noise to the data? Please give specifics for that procedure (i.e., percent of records affected, distribution of noise, etc.). Some possibilities include the following: (*NOTE: For a description of these techniques, consult SPWP # 22 or Appendix B.)*

random noise
record swapping
rank swapping
blanking and imputation

_____
_____
_____

Was any attempt made to match back the noise-added data to the original file?
Yes.
No. *If no, skip to 3.5.*

If yes, how was it done, and what was the rate of success in matching? _____

_____
_____
_____

Other Issues

Files that include every sample case or cases in strata that are sampled at high rates ($p > 0.5$) are more likely to lead to disclosure than files containing only a subsample of cases. For example, if it were known that a certain individual participated in a particular survey, one could infer that the person's record could be found in the corresponding microdata file, assuming all sample cases were available on that file.

Does this file contain every case?
Yes, it includes every case.
No, it includes a subsample of cases. If so, specify the range of sampling rates:
_____

Project managers should be aware that confidentiality problems may arise if special tabulations are made from an internal version of the file, which includes detail omitted from the public-use file. For example, a tabulation might provide specific geography, not included on the public-use file, that is cross-tabulated by multiple data items on the file. Consult with the Disclosure Review Board if you are planning to release tabulations that

make use of detail not available on the public-use file.

Do you plan to release tables that make use of detail not available on the public-use file?
Yes.
No.

Describe the Sample Design

Briefly describe the sample design by answering these questions:

Describe the sample design, including stratification, clustering, and stages. Be certain to identify the units that were sampled at any stage with probability > 0.5:
_____
_____

Provide a brief paragraph that compares and contrasts the proposed sampling units, units of enumeration, and units of analysis in the study: _____
_____
_____

The following subset of questions pertains to how much information is (will be) publicly available about the sample design, i.e., sampling plan and estimators (including the identity of PSU's):

(a) What information is already publicly available? _____
_____
_____
(b) What information will be made public with this release? _____
_____
_____
(c) What information will be withheld? _____
_____
_____

Describe how users will estimate sampling variances, potentially identifying any proposed "nesting variables" on the proposed file layout or the design of any weights used for replication approaches: _____
_____
_____
_____

Supplements

Was this information gathered as a supplement to another survey?
Yes.
No. *If no, you are finished with this section of the Checklist.*

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Can this microdata file be linked to the file produced from the main survey?

Yes.
No.
Don't know.

If yes, what geographic information is on the main file?

_____
_____
_____

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Section 4. Tabular Data from Persons or Households (or "Demographic Tables")

The Data

A table is often referred to by its **dimensions**. If the values presented in the cells of a statistical table are aggregates over two variables, then the table is a **two-dimensional table**. An example of such a demographic table would be one that displays the counts of individuals by 5-year age category and the highest level of education (some high school, high school graduate, received General Equivalency Diploma, some college, college graduate, and post-graduate studies). Typically, categories of one variable are given in columns, and categories of the other variable are given in rows. If the values presented in the cells of a statistical table are aggregates over three variables, then the table is a **three-dimensional table**. Starting with the previous example, if the data was also presented by county for the State of Maryland, then the three dimensions would be age group, education level, and county. In a three-dimensional table, the first two dimensions are said to be presented in columns and rows, the third variable in "layers." In order for a table to be three-dimensional, all one-dimensional, two-dimensional, and three-dimensional marginal totals must be displayed, unless they are suppressed. If only the marginal totals are displayed for two of these variables, then we have a set of two-dimensional tables instead. To illustrate using our last example where we had age group, education level, and county for the State of Maryland: collapsing across age groups would result in a two-way table of county by education level; collapsing across education levels, a two-way table of county by age group; and collapsing across counties, a two-way table of age group by education level for the State of Maryland.

What data will be released and in what formats (i.e., table dimensions; variables and their detail)? One way to describe this is to use a table like the following:

Tabulations	**Cell Contents**	**Comments**
VbleA x VbleB x VbleC	ListA	
VbleA x VbleD	ListB	
VbleB x VbleD	ListB	
…		

Description of the variables:

VbleA = county
VbleB = income ranges (0; 1-9,999; 10,000-49,999;…)
VbleX = income
etc…

Description of the contents of each list (magnitude data):

ListA = unit (person/household) counts, VbleW, VbleX, VbleY, and VbleZ
ListB = unit (person/household) counts, VbleX, VbleY, and VbleZ
etc…

Use this space for additional descriptive information for item 2.1.1._____
_____
_____
_____

What is the highest dimension of a table that you are releasing? _____

What level(s) of geography is(are) released? _____
_____
_____

Do you collect written waivers from individuals or households that would enable you to publish information that might otherwise have to be suppressed or masked?
(*NOTE: For more information on waivers, see Jabine, 1993.*)
Yes.
No.

If yes, please describe: _____
_____
_____

One method of protecting the confidentiality of data is to conduct a **sample** survey rather than a census. In reporting results of large-scale sample surveys, estimates are made by multiplying a respondent's data by a sampling weight before they are aggregated.

Are the data from a sample or a census (complete count)?
sample.
census. *If census, please skip to Section 4.2.*

If data are from a sample,
…are some groups of individuals selected with certainty?
Yes.
No.

…briefly describe the sample design, including overall sampling rate:
_____
_____
_____
_____

...are weights common knowledge (or could easily be inferred) so that a cell with a value of 10, for example, could be linked to one person/household?
Yes.
No.

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

(d) If yes, please explain: _____
_____

Coordinating preliminary release(s) with final release(s).

Is this a preliminary release or a final release?
Final release.
Preliminary release. (*NOTE: Identical suppression patterns must be used in both the preliminary and final releases. If this is a preliminary release, consult with the Disclosure Review Board. Do **not** lock yourself into a disclosure method/pattern too early in the process because the disclosure limitation method(s) used for the preliminary release may restrict subsequent release(s), including the final release.)*

Disclosure Risks Associated with Administrative Data and Other External Data

The disclosure risk of a table is increased if it contains administrative data or any type of data from an outside/external source. For instance, agencies sometimes use administrative records as the sampling frame for a demographic survey. If the agency providing such administrative data uses a "stricter" set of confidentiality rules than the receiving agency, then the recipient of the data may need to apply these "stricter" rules before the data can be released.

Were any administrative data or data from an external source used as a sampling frame for this survey?
Yes.
No. *If no, please skip to 4.2.4.*

If yes, please describe: _____
_____
_____

Were any administrative data or data from an external source used to "expand" the content of this census or survey (i.e., by merging administrative data with survey data)?
Yes.
No. *If no, please skip to 4.2.4.*

If yes, please describe: _____
_____
_____

Does the agency providing these administrative/external data have different confidentiality rules than your office/program?
Yes.
No.

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

If yes, please describe rules used by the other agency and how your office/program met both sets of rules: _____

_____

_____

Are you aware of any external data sources (i.e., State data files, files released by another Federal agency) that could be used to "compromise" your data? When such files exist, there is a risk that an "attacker" would use these external data sources to deliberately try to identify individual respondents and gain access to their confidential responses.

Yes. If yes, please describe: _____

_____

_____

No.

Don't know/did not check. Please explain. _____

_____

_____

_____

Disclosure Limitation Methods

The selection of a statistical disclosure limitation technique for data presented in tables (**tabular data**) depends on whether the data represent frequencies or magnitudes.

Tables of frequency count data present the number of units of analysis in a cell. Equivalently, the data may be presented as a percent by dividing the count by the total number presented in the tables (or total in a row or column) and multiplying by 100. In a table of frequency data, each respondent contributes equally to each cell in which he/she is represented (the contribution usually being the respondent's survey weight, or 1 in an unweighted table or a census). A confidentiality problem arises for this type of table when a cell has only a few respondents and the characteristics are sufficiently distinctive. Then it may be possible for a knowledgeable user to identify individual respondents. Disclosure limitation methods are applied to cells with less than a specified number of respondents to minimize the risk that respondents can be identified -- this approach is called a **threshold definition of a sensitive cell**. Typical disclosure limitation methods include **cell suppression** and a variety of data perturbation methods (including **traditional rounding, random rounding**, **controlled rounding**, and **record swapping**).

Tables of magnitude data present the aggregate of a "quantity of interest" over all units of analysis in the cell. Equivalently, the data may be presented as an average by dividing the aggregate by the number of units in the cell. To formally distinguish **frequency count data** from **magnitude data** -- for the latter, the "quantity of interest" must measure something other than membership in the cell. In a table of magnitude data, respondents contribute unequally to each cell. This requires a more elaborate definition of sensitive cells for tables of magnitude data -- such definitions are called **linear sensitivity measures**. In a table of magnitude data the confidentiality problem is that you want to make sure that the data user

cannot use the published totals and other publicly available data to estimate a respondent's value too closely. In contrast to tables of frequency data, fewer alternative methods for statistical disclosure limitation are available for tables of magnitude data -- **cell suppression** is typically used (another possible technique is **noise addition**).

For practical purposes, entirely rigorous definitions are not necessary. Typically, demographic tables contain only frequency count data. The cell suppression technique used for magnitude data can also be used for frequency data. (*NOTE: The "noise" technique described in Evans et al., 1996, does **not** apply to frequency data.*)

Disclosure Limitation Method: Cell Suppression

One historical method of protecting sensitive cells in tables is cell suppression. This means that sensitive cells are **not published** -- they are **suppressed**. These sensitive cells are called **primary suppressions**. To make sure the primary suppressions cannot be derived by subtractions from published marginal totals, additional cells (which are nonsensitive) are selected for **complementary suppression**. Complementary suppressions are sometimes called **secondary suppressions**.

Did you use cell suppression as a disclosure limitation method?
Yes.
No. *If cell suppression was not used, please skip to Section 4.3.2.*

(*NOTE: The following cell suppression technique is used for magnitude data.*) In some cases, a program office may have more than one primary suppression rule. For example, a cell is suppressed when: (a) it has a count of one or two; or (b) when a marginal total equals the value of that particular internal cell.

Agencies use one of three common methods for determining whether or not a cell is sensitive and, therefore, should be suppressed (called **linear sensitivity measures**). As noted above, these sensitive cells are called primary suppressions.

The most commonly used method for determining whether or not a cell is sensitive and, therefore, should be a primary suppression is the **(n,k) rule**: This rule states that if a small number (n or fewer) of the respondents contribute a large percentage (k percent or more) of the total cell value, then that cell is sensitive. Note that 'n' is usually much smaller than the total number of respondents for the cell. The (n,k) rule has also been called the **dominance rule**. In the case when n=1, this rule declares sensitive any cell in which the cell value can be used to obtain an upper bound for the largest respondent's value that is "close" to its actual value, where "close" depends on the value of k. In the case when n > 1, a cell is sensitive when n-1 collaborators in the cell can use their data together with the cell value to obtain an upper bound for the largest respondent's value that is close to its actual value.

An alternative to the **(n,k) rule** is known as the **pq rule**: This rule begins by assuming that prior information allows each respondent's value to be estimated within q percent of its actual value. If that information together with the cell value allows the value of the largest

respondent to be estimated within p percent of its actual value ($p < q$), then the cell is sensitive.

A third rule, the **p-percent rule,** is simply a special case of the pq rule with q=100.

(a) What rule(s) was(were) used to determine primary suppressions? Please check all that apply and give parameter(s).
the (n,k) rule. Give parameters: _____
the pq rule. Give parameters: _____
the p-percent rule. Give parameters: _____
other rule. Please describe this rule: _____
_____
_____

(b) Please describe how you applied this(these) rule(s). For example, if you used the (n,k) rule, describe the parameters used for each variable: _____
_____
_____
_____

(*NOTE: The following cell suppression technique is used for magnitude data.*) A simplifying procedure for cell suppression is **key item suppression**. In this method, an agency would perform primary disclosure analysis and complementary suppression on certain key data items only and then apply the same suppression pattern to other related items. Under key item suppressions, fewer agency resources are devoted to disclosure limitation, and data products are more uniform across data items. Key and related items are identified by expert judgment. They should remain stable over time.

Was(were) a key item(s) chosen in performing cell suppression?
Yes.
No.

If yes, what was this key item, and why was it selected? _____
_____
_____

(NOTE: The following cell suppression technique is used for magnitude data.) **In determining secondary suppressions, what was your goal?**
minimize the number of cells that are suppressed.
minimize the amount of information that is suppressed.
other – please describe: _____
_____
_____

(*NOTE: The following cell suppression technique is only used for frequency data.*) If a cell has only a few respondents and the characteristics are sufficiently distinctive, then it may

be possible for a knowledgeable user to identify respondents in the population. Disclosure limitation methods are sometimes applied to cells with fewer than a specified **threshold number** of respondents to minimize the risk that respondents can be identified from their data. With weighted data, a more conservative approach would be to suppress cells based on the unweighted counts rather than the weighted counts.

In addition, a disclosure risk may also occur when a cell contains all members of a domain and thereby discloses some information about them, for example, if we released a frequency count table showing that the 12 hemophiliacs in a county who are aged 25-34 are all HIV-positive or have criminal records, etc. Another example would be publishing a frequency table in which all members of a particular group (e.g., physicians in the county who graduated less than 5 years ago) are in a narrowly-defined income range.

(a) Did you suppress values that were based on less than a specified threshold number?
Yes. Please give threshold number: _____
No. If you did not use a threshold rule, please explain why not: _____
_____
_____

If yes, was it applied to weighted or unweighted data?
Weighted data.
Unweighted data.

Do any cells contain all members of a domain?
Yes.
No.
Don't know.

If yes, please describe: _____

(*NOTE: The following cell suppression technique is used for either frequency or magnitude data.*) For small tables, it is possible to manually select cells for complementary or secondary suppression and then to apply audit procedures (see Section 4.3.2 below) to guarantee that the selected cells adequately protect the sensitive cells. For large-scale survey publications with many related tables, the selection of a set of secondary suppression cells can be an extremely complex problem.

(a) Were secondary suppressions selected by hand or suppression software? (Please check all that apply, and describe what you did.)
by hand.
by suppression software developed in this agency.
by suppression software developed by an outside vendor. Please provide name of vendor and product: _____
_____

(b) Please describe your method: _____

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

_____
_____

Auditing

Once sensitive cells have been identified by a primary suppression rule, and secondary suppressions have been selected, it is important that an agency examine the resulting table to make sure that the primary suppressions are adequately protected. The proposed suppression pattern should be **audited** in order to make sure that the primary suppressions are sufficiently well-protected. There are automated methods of auditing.

Certain methods for deriving complementary suppressions are **self-auditing**. Self-auditing means that the protection provided is measured and compared to prespecified levels, thereby ensuring automatically that sufficient protection is achieved. For example, when **network flow methods** are used to derive complementary suppressions, they are self-auditing for two-dimensional tables.

Were the suppression patterns in the tables audited? That is, were all tables audited?
Yes. If yes, please explain how: _____
_____
No. If no, explain why not: _____
_____

Were any suppressions removed by hand?
Yes.
No.

If yes, please explain why: _____
_____
_____

If yes, how were suppressions chosen to replace them? _____
_____
_____

Will any additional information be released for values that were suppressed (i.e. ranges, medians, estimates, rounded values, values with noise, etc.)?
Yes.
No.

If yes, please give details: _____
_____
_____

Disclosure Limitation Method: Addition of Noise

NOTE: *Responses to the questions in this Checklist must be treated as strictly confidential.*

(*NOTE: The following method is only used for magnitude data.*) Another statistical disclosure limitation method is to add **noise** to the underlying microdata before producing tabular products. In this method, each individual/household in the sampling universe is assigned a multiplier -- called a "**noise factor**." Whenever an individual/household is canvassed in any survey or census, all of its values are multiplied by its noise factor. Within a particular survey or census, all individuals/households would have their values multiplied by their corresponding noise factors **before** the data were tabulated. Since the same multiplier is always associated with an individual/household and used wherever its data are tabulated, values would be consistent from one table to another. That is, if the same cell appeared on more than one table, it would have the same value in all tables. (For more information on this technique, see Evans et al.,1996.)

Did you add noise to the underlying microdata before creating the tables?

Yes.

No. *If no, please skip to Section 4.3.4.*

Which items received noise? _____

_____

_____

How was noise added to the data? _____

_____

_____

How much noise was added to the data?_____

_____

_____

Disclosure Limitation Method: Additional Methods For Frequency Count Data

In addition to the cell suppression techniques described above, there are other disclosure limitation methods that can be used for frequency count data. These additional methods include traditional rounding, random rounding, record swapping; blanking and imputation; and controlled rounding. (*NOTE: For a description of these other techniques, consult SPWP # 22 or Appendix B.)*

Did you use additional disclosure methods, other than cell suppression, for your frequency count data?

Yes.

No. *If no, please skip to Section 4.4.*

If yes, please describe: _____

_____

_____

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Coordination of Disclosure Limitation

There is a risk of disclosing confidential information when multiple surveys in an agency produce tables that are generated from the same data set. All suppressions must be coordinated among all tables generated from the same data set in order to ensure that disclosure limitation methods applied to one table cannot be used to identify sensitive cells in a previously released table. For example, an agency in the Department of Health and Human Services (DHHS) may contract with the National Center for Health Statistics (NCHS) to collect data in a supplement to one of NCHS's ongoing surveys. While the DHHS agency that sponsors the supplement may have plans to publish data in a certain way, it needs to carefully coordinate its data release with that done by NCHS. You want to be sure that suppression patterns in related tables do not unravel each other.

The principle of coordination is especially important when nonstandard or special tables are created or when "regular/standard" tables were previously published. Before such tabular products are released, special actions must be taken by the program staff to coordinate disclosure limitation methods.

Have these data been used for research purposes (for example, by an external researcher) or released as in a special tabulation?

Yes.
No.

If yes, please give a brief description of this use or what was released: _____
_____
_____

Has the same (or very similar) data also been released by your unit or by another division, program area, or branch?

Yes.
No. *If no, you have completed the Checklist.*

If yes, please describe: _____
_____
_____

Were disclosure limitation techniques (such as cell suppression patterns) coordinated with those previously used?

Yes. If yes, please explain: _____
_____
_____

No. If no, explain why you did not coordinate: _____
_____
_____

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Section 5: Tabular Data from Establishments or Other Types of Organizations

In contrast to demographic data, it is presumed that data collected from establishments or other organizations (such as hospitals, schools) often pose a higher risk for several reasons: the data are typically very skewed, the size of the universe may be small, and there are many high visibility variables. For example, in the U.S., there are only a handful or so of hospitals with 1,000 or more beds, and inadvertent disclosure in a survey of hospitals might be possible using detail on the number of beds and as large a geographic area as a Census region. In addition, the potential value of financial and other proprietary business data for identified units to outsiders may provide a strong incentive for some of them to derive such information from the released data. For these reasons, establishment data collected under a pledge of confidentiality are rarely released as public-use microdata files and are typically released in tables.

The Data

A table is often referred to by its **dimensions**. If the values presented in the cells of a statistical table are aggregates over two variables, then the table is a **two-dimensional table**. An example of such a table would be one that displays the value of construction work done during a particular period in the State of Maryland by county and by 4-digit Standard Industrial Code (SIC) groups. Typically, categories of one variable are given in columns, and categories of the other variable are given in rows. If the values presented in the cells of a statistical table are aggregates over three variables, then the table is a **three-dimensional table**. Starting with the previous example, if the data were also presented by year, then the three dimensions would be year, county, and SIC code. In a three-dimensional table, the first two dimensions are said to be presented in columns and rows, the third variable as "layers." In order for a table to be three-dimensional, all one-dimensional, two-dimensional, and three-dimensional marginal totals must be displayed, unless they are suppressed. If the marginal totals are displayed only for two of these variables, then we have a set of two-dimensional tables instead. To illustrate our last example, if the marginal totals summed all cells over county and SIC code for each single year, then we would have a set of two-dimensional tables, one for each year. However, if the marginal totals also displayed the sum over different years, then we would have a three-dimensional table.

What data will be released and in what formats (i.e., table dimensions; variables and their detail)? One way to describe this is to use a table like the following:

Tabulations	**Cell Contents**	**Comments**
VbleA x VbleB x VbleC	ListA	
VbleA x VbleD	VbleE	
VbleB x VbleD	VbleE	
…		

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Description of the variables:
VbleA = year
VbleB = county
VbleC = two-digit industrial classification
etc…
Description of the contents of each list (magnitude data):
ListA = unit (establishment) counts, VbleX, VbleY, and VbleZ
VbleE = value of shipment in range (0; 1-9,999; 10,000-49,999;…)
etc…

Use this space for additional descriptive information for item 5.1.1._____
_____
_____
_____

What is the highest dimension of a table that you are releasing? _____

What level(s) of geography is(are) released? _____
_____
_____

Do you collect written waivers from establishments or organizations that would enable you to publish information that might otherwise have to be suppressed or masked?
(*NOTE: For more information on waivers, see Jabine, 1993.*)
Yes.
No.

If yes, please describe: _____
_____
_____

Are establishment/organization counts released?
Yes.
No.

One method of protecting the confidentiality of data is to conduct a **sample** survey rather than a census. In reporting results of large-scale sample surveys, estimates are made by multiplying a respondent's data by a sampling weight before they are aggregated.

Are the data from a sample or a census (complete count)?
sample.
census. *If census, please skip to Section 5.2.*

If data are from a sample,
…are some types of establishments selected with certainty?

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Yes.
No.

…briefly describe the sample design, including overall sampling rate:
_____
_____
_____
_____

...are weights common knowledge (or could easily be inferred) so that a cell showing 10 units, for example, could be linked with only one establishment?
Yes.
No.

If yes, please explain: _____
_____

Coordinating preliminary release(s) with final release(s).

Is this a preliminary release or a final release?
Final release.
Preliminary release. (*NOTE: Identical suppression patterns must be used in both the preliminary and final releases. If this is a preliminary release, consult with the Disclosure Review Board. Do **not** lock yourself into a disclosure method/pattern too early in the process because the disclosure limitation method(s) that are used for the preliminary release may restrict subsequent release(s), including the final release.*)

Disclosure Risks Associated with Administrative Data and Other External Data

The disclosure risk of a table is increased if it contains administrative data or any type of data from an outside/external source. For instance, agencies sometimes use administrative records as the sampling frame for an establishment/organizational survey. If the agency providing such administrative data uses a "stricter" set of confidentiality rules than the receiving agency then the recipient of the data may need to apply these "stricter" rules before the data can be released.

Were any administrative data or data from an external source used as a sampling frame for this survey?
Yes.
No. *If no, please skip to 5.2.4.*

If yes, please describe: _____
_____
_____

Were any administrative data or data from an external source used to "expand" the

content of this census or survey (i.e., by merging administrative data with survey data)?

Yes.
No. *If no, please skip to 5.2.4.*

If yes, please describe: _____
_____
_____

Does the agency providing these administrative/external data have different confidentiality rules than your office/program?

Yes.
No.

If yes, please describe rules used by the other agency and how your office/program met both sets of rules._____
_____
_____

Are you aware of any external data sources (forexample, State-level data files, Dunn and Bradstreet files) that could be used to "compromise" your data? When such files exist, there is a risk that an "attacker" could use these external data sources to deliberately try to identify individual respondents and gain access to their confidential responses.

Yes. If yes, please describe: _____
_____
_____

No.
Don't know/did not check. Please explain: _____
_____
_____
_____

Disclosure Limitation Methods

The selection of a statistical disclosure limitation technique for data presented in tables (**tabular data**) depends on whether the data represent frequencies or magnitudes.

Tables of frequency count data present the number of units of analysis in a cell. Equivalently, the data may be presented as a percent by dividing the count by the total number presented in the tables (or total in a row or column) and multiplying by 100. In a table of frequency data, each respondent contributes equally to each cell in which he/she is represented (the contribution usually being the respondent's survey weight, or 1 in an unweighted table or a census). A confidentiality problem arises for this type of table when a cell has only a few respondents and the characteristics are sufficiently distinctive. Then it may be possible for a knowledgeable user to identify individual respondents. Disclosure limitation methods are

applied to cells with less than a specified number of respondents to minimize the risk that respondents can be identified -- this approach is called a **threshold definition of a sensitive cell**. Typical disclosure limitation methods include **cell suppression** and a variety of data perturbation methods (including **random rounding**, **controlled rounding**, and **record swapping**).

Tables of magnitude data present the aggregate of a "quantity of interest" over all units of analysis in the cell. Equivalently, the data may be presented as an average by dividing the aggregate by the number of units in the cell. To formally distinguish **frequency count data** from **magnitude data** -- for the latter, the "quantity of interest" must measure something other than membership in the cell. In a table of magnitude data, respondents contribute unequally to each cell. This requires a more elaborate definition of sensitive cells for tables of magnitude data -- such definitions are called **linear sensitivity measures**. In a table of magnitude data, the confidentiality problem is that you want to make sure that the data user cannot use the published totals and other publicly available data to estimate a respondent's value too closely. In contrast to tables of frequency data, fewer alternative methods for statistical disclosure limitation are available for tables of magnitude data -- **cell suppression** is typically used (another possible technique is **noise addition**).

Thus, tables of the number of establishments within the manufacturing sector by SIC group and by county-within-State are frequency count tables, whereas tables presenting total value of shipments for the same cells are tables of magnitude data. For practical purposes, entirely rigorous definitions are not necessary. The cell suppression techniques used for magnitude data can also be used for frequency data. (*NOTE: The "noise technique described in Evans et al., 1996, does **not** apply to frequency data.*)

Disclosure Limitation Method: Cell Suppression

One historical method of protecting sensitive cells in tables is cell suppression. This means that sensitive cells are **not published** -- they are **suppressed**. These sensitive cells are called **primary suppressions**. To make sure the primary suppressions cannot be derived by subtractions from published marginal totals, additional cells (which are nonsensitive) are selected for **complementary suppression**. Complementary suppressions are sometimes called **secondary suppressions**.

Did you use cell suppression as a disclosure limitation method?
Yes.
No. *If cell suppression was not used, please skip to Section 5.3.2.*

(*NOTE: The following cell suppression technique is used for magnitude data.*) In some cases, a program office may have more than one primary suppression rule. For example, consider a set of tables containing the number of on-the-job workplace injuries. For such tables, the general rule may be to suppress a cell where one establishment has 75% or more of the total of that cell. This type of table has another vulnerability in that the reporting of no data at all can also violate confidentiality. If no reporter in a given cell experienced any injury or illness cases, the estimate for that cell would legitimately be

zero. This would reveal the value (zero) that all contributors reported. A second primary suppression may need to be applied to cells that contain no injuries.

Agencies use one of three common methods for determining whether or not a cell is sensitive and, therefore, should be suppressed (called **linear sensitivity measures**). As noted above, these sensitive cells are called primary suppressions:

The most commonly used method for determining whether or not a cell is sensitive and, therefore, should be a primary suppression is the **(n,k) rule**: This rule states that if a small number (n or fewer) of the respondents contribute a large percentage (k percent or more) of the total cell value, then that cell is sensitive. Note that 'n' is usually much smaller than the total number of respondents for the cell. The (n,k) rule has also been called the **dominance rule**. In the case when n=1, this rule declares sensitive any cell in which the cell value can be used to obtain an upper bound for the largest respondent's value that is "close" to its actual value, where "close" depends on the value of k. In the case when n > 1, a cell is sensitive when n-1 collaborators in the cell can use their data together with the cell value to obtain an upper bound for the largest respondent's value that is close to its actual value.

An alternative to the **(n,k) rule** is known as the **pq rule**: This rule begins by assuming that prior information allows each respondent's value to be estimated within q percent of its actual value. If that information together with the cell value allows the value of the largest respondent to be estimated within p percent of its actual value (p < q), then the cell is sensitive.

A third rule, the **p-percent rule,** is simply a special case of the pq rule with q=100.

What rule(s) was(were) used to determine primary suppressions? Please check all that apply and give parameter(s).
the (n,k) rule. Give parameters: _____
the pq rule. Give parameters: _____
the p-percent rule. Give parameters: _____
other rule. Please describe this rule: _____
_____
_____

Please describe how you applied this(these) rule(s). For example, if you used the (n,k) rule, describe the parameters used for each variable: _____
_____
_____
_____

(*NOTE: The following cell suppression technique is used for magnitude data.*) A simplifying procedure for cell suppression is **key item suppression**. In several economic censuses, the Census Bureau uses key item suppression, that is, performing primary disclosure analysis and complementary suppression on certain key data items only and then applying the same

suppression pattern to other related items. Under key item suppressions, fewer agency resources are devoted to disclosure limitation, and data products are more uniform across data items. Key and related items are identified by expert judgment. They should remain stable over time.

Was(were) a key item(s) chosen in performing cell suppression?
Yes.
No.

If yes, what was this key item, and why was it selected? _____
_____
_____

(*NOTE: The following cell suppression technique is used for magnitude data.*) **In determining secondary suppressions, what was your goal?**
minimize the number of cells that are suppressed.
minimize the amount of information that is suppressed.
other – please describe: _____
_____
_____

(*NOTE: The following cell suppression technique is only used for frequency data.*) If a cell has only a few respondents and the characteristics are sufficiently distinctive, then it may be possible for a knowledgeable user to identify respondents in the population. Disclosure limitation methods are sometimes applied to cells with fewer than a specified **threshold number** of respondents to minimize the risk that respondents can be identified from their data. With weighted data, a more conservative approach would be to suppress cells based on the unweighted counts rather that the weighted counts.

In addition, a disclosure risk also may also occur when a cell contains all members of a domain and thereby discloses some information about them, for example, if we released a frequency count table showing that the 12 restaurants in a particular county which had been in business for 6-10 years all have a certain characteristic.

Did you suppress values that were based on less than a specified threshold number?
Yes. Please give threshold number: _____
No. If you did not use a threshold rule, please explain why not. _____
_____
_____

If yes, was it applied to weighted or unweighted data?
Weighted data.
Unweighted data.

Do any cells contain all members of a domain?
Yes.

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

No.
Don't know.

If yes, please describe: _____

(*NOTE: The following cell suppression technique is used for either frequency or magnitude data.*) For small tables, it is possible to manually select cells for complementary or secondary suppression and then to apply audit procedures (see Section 5.3.2 below) to guarantee that the selected cells adequately protect the sensitive cells. For large-scale survey publications with many related tables, the selection of a set of secondary suppression cells can be an extremely complex problem.

Were secondary suppressions selected by hand or suppression software? (Please check all that apply, and describe what you did.)
by hand.
by suppression software developed in this agency.
by suppression software developed by an outside vendor. Please provide name of vendor and product: _____
_____

Please describe your method: _____
_____
_____

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Auditing

Once sensitive cells have been identified by a primary suppression rule, and secondary suppressions have been selected, it is important that agencies examine the resulting table to make sure that the primary suppressions are adequately protected. The proposed suppression pattern should be **audited** in order to make sure that the primary suppressions are sufficiently well-protected. There are automated methods of auditing.

Certain methods for deriving complementary suppressions are **self-auditing**. Self-auditing means that the protection provided is measured and compared to prespecified levels, thereby ensuring automatically that sufficient protection is achieved. For example, when **network flow methods** are used to derive complementary suppressions, they are self-auditing for two-dimensional tables.

Were the suppression patterns in the tables audited? That is, were all tables audited?
Yes. If yes, please explain how: **_____**
_____
No. If no, explain why not: _____
_____

Were any suppressions removed by hand?
Yes.
No.

If yes, please explain why: _____
_____
_____

If yes, how were suppressions chosen to replace them? _____
_____
_____

Will any additional information be released for values that were suppressed (i.e. ranges, medians, estimates, rounded values, values with noise, etc.)?
Yes.
No.

If yes, please give details: _____
_____
_____

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Disclosure Limitation Method: Addition of Noise

(*NOTE: The following method is only used for magnitude data.*) Another statistical disclosure limitation method is to add **noise** to the underlying microdata before producing tabular products. In this method, each establishment/organization in the sampling universe is assigned a multiplier -- called a "**noise factor**." Whenever an establishment/organization is canvassed in any survey or census, all of its values are multiplied by its noise factor. Within a particular survey or census, all establishments/organizations would have their values multiplied by their corresponding noise factors **before** the data were tabulated. Since the same multiplier is always associated with an establishment/organization and used wherever its data are tabulated, values would be consistent from one table to another. That is, if the same cell appeared on more than one table, it would have the same value in all tables. (For more information on this technique, see Evans et al.,1996.)

Did you add noise to the underlying microdata before creating the tables?
Yes.
No. *If no, please skip to Section 5.3.4.*

Which items received noise? _____
_____
_____

How was noise added to the data? _____
_____
_____

How much noise was added to the data?_____
_____
_____

Disclosure Limitation Method: Additonal Methods For Frequency Count Data

In addition to the cell suppression techniques described above, there are other disclosure limitation methods that can be used for frequency count data. These additional methods include traditional rounding, random rounding, controlled rounding, or record swapping. (*NOTE: For a description of these other techniques, consult SPWP # 22 or Appendix B.)*

Other than cell suppression, editing, and noise, were any other disclosure limitation techniques used for this data?
Yes.
No. *If no, please skip to Section 5.3.5.*

If yes, please describe: _____
_____
_____

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Treatment of Special Types of Magnitude Data

Some data require special treatment in terms of applying disclosure limitation techniques. Some possibilities follow (*NOTE: For a description of these other techniques, consult SPWP # 22.)*:

Tables that report negative values: If all reported values are negative, suppression rules can be applied directly by taking the absolute value of the reported data.

Tables that report net changes (that is, the difference between values reported at different times): If either of the values used to calculate net change were suppressed in the original publication, then net change must also be suppressed.

Tables where differences between positive values are reported: If the published item is the difference between two positive quantities reported for the same time period (e.g., net production equals gross production minus inputs), then the suppression rule that is applied depends on whether the resulting difference is generally positive or generally negative. (For suggestions on appropriate suppression procedure, consult SPWP # 22, page 90.)

Tables reporting weighted averages: If a published item is the weighted average of two positive reported quantities, such as volume weighted price, apply the suppression procedure to the weighting variable (volume in this example).

Are you planning to release tables that report negative values, net changes, differences, or weighted averages?

Yes.
No.

Did any data require special treatment?

Yes.
No.

If yes to either 5.3.5.1 or 5.3.5.2, please describe the data element(s) and what special disclosure procedures were done: _____

_____

_____

Collecting Data that Naturally Fall into Clusters or Groups

An agency collects data at the establishment or organizational level. However, it may be aware that certain establishments can be grouped into "clusters." For example, a firm may have multiple locations, and all locations have a uniform pricing policy. Therefore, in a survey of pricing policies, all of these establishments should be considered to be one sampling unit when disclosure limitation methods were applied. As another example, multiple establishments owned by one company may share confidential data among themselves.

Does the data that you collect fall into natural "clusters?"

Yes.
No. *If no, please skip to Section 5.5.*

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

If yes, please describe these clusters: _____
_____
_____

Do you consider such clusters when you apply disclosure limitation methods to tables?
Yes. If yes, please explain what was done: _____
_____
No. If no, please explain why the clusters were not taken into consideration:
_____
_____
_____

Coordination of Disclosure Limitation

There is a risk of disclosing confidential information when multiple surveys in an agency produce tables that are generated from the same data set. All suppressions must be coordinated among all tables generated from the same data set in order to ensure that disclosure limitation methods applied to one table cannot be used to identify sensitive cells in a previously released table. For example, in the Bureau of Labor Statistics (BLS), the Universe Database (UDB) is used as the sampling frame for almost all of BLS's establishment surveys. So, hypothetically, in BLS, it is possible that two program areas which use the UDB could want to publish the same two-way table with identical primary suppressions but different secondary suppressions. An agency wants to be sure that suppression patterns in related tables do not unravel each other.

The principle of coordination is especially important when nonstandard or special tables are created or when "regular/standard" tables were previously published. Before such tabular products are released, special actions must be taken by the program staff to coordinate disclosure limitation methods.

Have these data been used for research purposes (for example, by an external researcher) or released as in a special tabulation?
Yes.
No.

Please give a brief description of this use or what was released: _____
_____
_____

Has the same (or very similar) data also been released by your unit or another division, program area, or branch?
Yes.
No. *If no, you have completed the Checklist.*

Were disclosure limitation techniques (such as cell suppression patterns) coordinated with those previously used?
Yes. If yes, please explain: _____

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

_____
_____

No. If no, explain why you did not coordinate: _____

_____
_____

***NOTE**: Responses to the questions in this Checklist must be treated as strictly confidential.*

Section 6. Selected References

Statistical Methods to Limit Disclosure:

Evans, T., Zayatz, L., and Slanta, J. (August 1996). "Using Noise for Disclosure Limitation of Tabular Data," *Proceedings of the 1996 Annual Research Conference and Technology Interchange*. Washington, DC: U.S. Department of Commerce, Bureau of the Census, pp. 65-86.

Federal Committee on Statistical Methodology. (May 1978). *Report on Statistical Disclosure and Disclosure-Avoidance Techniques*. (Statistical Policy Working Paper # 2). Washington, DC: U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Federal Committee on Statistical Methodology. (May 1994). *Report on Statistical Disclosure Limitation Methodology*. (Statistical Policy Working Paper # 22). Washington, DC: Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office.

Journal of Official Statistics (vol. 14, no. 4, 1998) special issue entitled "Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data." The Table of Contents follows:

Fienberg, S.E. and Willenborg, L.C.R.J. "Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data," pp. 337-346.

Keller-McNulty, S. and Unger, E.A. "A Database System Prototype for Remote Access to Information Based on Confidential Data," pp. 347-360.

Skinner, C.J. and Holmes, D.J. "Estimating the Re-identification Risk Per Record in Microdata," pp. 361-372.

Samuels, S.M. "A Bayesian Species-Sampling-Inspired Approach to the Uniques Problem in Microdata Risk Assessment," pp. 373-384.

Fienberg, S.E. and Makov, U.E. "Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data," pp. 385-398.

Pannekoek, J. and de Waal, T. "Synthetic and Combined Estimators in Statistical Disclosure Control," pp. 399-410.

Zaslavsky, A.M. and Horton, N.J. " Balancing Disclosure Risk Against the Loss of Nonpublication," pp. 411-420.

de Waal, A.G. and Willenborg, L.C.R.J. "Optimal Local Suppression in Microdata," pp. 421-436.

Hurkens, C.A.J. and Tiourine, S.R. "Models and Methods for the Microdata Protection Problem," pp. 437-448.

Defays, D. and Anwar, M.N. "Masking Microdata Using Micro-Aggregation," pp. 449-462.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., and de Wolf, P.-P. "Post Randomisation for Statistical Disclosure Control: Theory and Implementation," pp. 463-468.

Sande, G. "Comment," pp.479-484.

Fienberg, S.E., Makov, U.E., and Steel, R.J. "Disclosure Limitation Using Perturbation and Related Methods for Categorical Data," pp. 485-502.
Kooiman, P. "Comment," pp. 503-508.
Fienberg, S.E., Makov, U.E., and Steel, R.J. "Rejoinder," pp. 509-512.
Kirkendall, N. and Sande, G. "Comparison of Systems Implementing Automated Cell Suppression for Economic Statistics," pp. 513-536.
Evans, T., Zayatz, L., and Slanta, J. "Using Noise for Disclosure Limitation Establishment Tabular Data," pp. 537-552.
Fischetti, M. and Salazar-Gonzalez, J.-J. "Experiments with Controlled Rounding for Statistical Disclousre Control in Tabular Data with Linear Constraints," pp. 553-566.

Kim, J.J. and Winkler, W. E. (1995). "Masking Microdata Files," *American Statistical Association, 1995 Proceedings of the Section on Survey Research Methods*, pp. 114-119.
Manual on Disclosure Control Methods. (1996). (Catalogue #: CA-94-96-283-EN-C). Luxembourg: Eurostat.
Willenborg, L., and de Waal, T. (1995). *Statistical Disclosure Control in Practice.* (Lecture Notes in Statistics 111). NY: Springer-Verlag, Inc.

Restricted Access Procedures:

Jabine, T. B. (1993). "Procedures for Restricted Access," *Journal of Official Statistics*, 9(2), pp. 537-590.
Reznek, A. P., Cooper, J. M. R., and Jensen, J. B. (1997). "Increasing Access to Longitudinal Survey Microdata: The Census Bureau's Research Data Center Program," *American Statistical Association, 1997 Proceedings of the Government Statistics Section and the Social Statistics Section*, pp. 243-248.

Appendix A: Statistical Disclosure Limitation Method by Three Types of Release (Microdata Files, Demographic Tables, and Economic Tables)

Statistical Disclosure Limitation Method	Type of Release		
	Microdata File	**Demographic Tables**	**Economic Tables**
Record swapping	Yes	Yes	
Blanking and imputing	Yes	Yes	
Rank swapping	Yes		
Traditional rounding		Yes	
Controlled rounding		Yes	
Random rounding		Yes	
Noise	Yes		Yes
Cell suppression		Yes	Yes
Local suppression	Yes		
Recoding into broader categories (includes top-coding, bottom-coding, and geographic restrictions	Yes	Yes	Yes
Blurring	Yes		
Microaggregation	Yes		
Multiple imputation	Yes		
Data modification		Yes	

Appendix B: Definitions of Selected Statistical Disclosure Limitation Methods[2]

Blanking and imputation: Blanking and imputation involve selecting a sample of respondent records from internal files and blanking a subset of the values on those records and then using imputation techniques to fill in the blanked values.

Data modification (data blurring): Data modification (or data blurring) involves adding uncertainty to all cell values in a table or to cell values that are less than a prescribed threshold (e.g., 10). Modifications to cell values (e.g., -1, 0, or +1) are added with prescribed probabilities. For example, with these modifications, a published value of 4 could represent an actual value of 3, 4, or 5. To the extent that tables have actual cell values below a prescribed threshold, there will be rounding error and inconsistencies within tables and among tables.

Data swapping: Data swapping involves selecting a sample of respondent records from internal files and interchanging data for these respondents with other respondents that have identical characteristics on a set of key variables.

Rounding: Rounding involves adjusting the values in all cells in a table to a specified base so as to create uncertainty about the real value for any cell while adding a small but acceptable amount of distortion to the data. There are several ways in which the rounding technique can be implemented:

Controlled rounding: In controlled rounding, an attempt is made to round the values in cells of a table so as to preserve summation to subtotals and table totals. With controlled rounding, there may be inconsistencies among tables. For example, the total population in a table showing data on age may be different from the total population in a table showing data on race.

Fixed (or traditional) rounding: In fixed (or traditional) rounding, each cell value is rounded independently in a prescribed manner, which produces nonadditivity within tables but consistency among tables. An example of fixed rounding is to round values ending in 8, 9, 0, 1, or 2 to a value ending in 0 (e.g., a value of 18, 19, 20, 21, or 22 would be rounded to 20). In the example of tables on age and on race in the preceding paragraph, the total population with fixed rounding will be the same in the two tables, but the sum of cell values by age or by race may not add to the total populations in the respective tables.

Random rounding: In random rounding, each cell value is rounded independently in a random manner. For example, values of 6, 7, 8, or 9 could be rounded to 5 or 10 based on assigned probabilities. With random rounding, there may be inconsistency in data within tables and among tables.

---

[2] For more information consult the Federal Committee on Statistical Methodology's *Report on Statistical Disclosure Limitation Methodology* (May 1994).

Note # 1. With controlled rounding, fixed rounding, and random rounding, there is a choice of the base for rounding, the most common choices being 3 and 5. All rounded values (other than zeros) are multiples of 3 or 5, respectively. Base 3 rounding introduces less distortion than base 5 rounding, but since multiples of 3 can end in any digit from 0 to 9, the use of base 3 rounding is less obvious to data users than the use of base 5 rounding, which produces values ending only in 0 or 5.

Note # 2. The techniques of rounding and data modification are similar in that they introduce uncertainty into the published data. The primary difference is that the rounding technique uses a specified base (usually 3 or 5) for rounding values in table cells whereas data modification does not use a specific base.