CHAPTER III – Current Federal Statistical Agency Practices

This chapter provides an overview of 14 Federal agency policies, practices, and procedures for statistical disclosure limitation. Agencies are authorized or required to protect individually identifiable data by a variety of statutes, regulations or policies. Statistical disclosure limitation methods are applied by the agencies to limit the risk of disclosure of individual information when statistics are disseminated in tabular or microdata formats.

This review of agency practices is based on three sources. The first source is Jabine (1993b), a paper based in part on information provided by the statistical agencies in response to a request in 1990 by the Panel on Confidentiality and Data Access, Committee on National Statistics. Another source of agency practices was from 1991 when each statistical agency was asked to provide a description of its current disclosure practices, standards, and research plans for tabular and microdata. 12 statistical agencies responded to this request.

The third source was from 2004, when each agency was requested by the Confidentiality and Data Access Committee, a subcommittee of the Federal Committee on Statistical Methodology, to review and supplement their responses concerning current disclosure practices and standards, and to comment on any provisions for researcher access. Thus, the material in this chapter is current as of the publication date.

The first section of this chapter summarizes the disclosure limitation practices for 14 Federal statistical agencies as shown in Statistical Programs of the United States Government: Fiscal Year 2004 (Office of Management and Budget). The agency summaries are followed by an overview of the current status of statistical disclosure limitation policies, practices, and procedures based on the available information. Specific methodologies and the state of software being used are discussed to the extent they were included in the individual agencies' responses.

A. Agency Summaries

A.1. Department of Agriculture

A.1.a. Economic Research Service (ERS)

ERS disclosure limitation practices are documented in the statement of "ERS Policy on Dissemination of Statistical Information," dated September 28, 1989. This statement provides that: Estimates will not be published from sample surveys unless: (1) sufficient nonzero reports are received for the items in a given class or data cell to provide statistically valid results which are clearly free of disclosure of information about individual respondents. In all cases at least three observations must be available, although more restrictive rules may be applied to sensitive data, (2) the second condition is an application of the (n, k) concentration rule or dominance rule to insure that the unexpanded data for any one respondent does not equal a specified threshold, For each published cell value, the respondent must represent less than 60 percent of the total that is being published, except when written permission is obtained from that respondent. In this instance (n, k) = (1, 0.6). Both conditions are applied to magnitude data while the first condition also applies to counts.

Within ERS, access to unpublished, confidential data is controlled by the appropriate branch chief. Authorized users must sign confidentiality certification forms. Restrictions require that data be summarized so individual reports are not revealed.

ERS does not release public-use microdata files. ERS provides access to microdata via its "remote data center" software to authorized users. ERS will share data for statistical purposes with government agencies, universities, and other entities under cooperative agreements as described below for the National Agricultural Statistics Service (NASS). Requests of entities under cooperative agreements with ERS for tabulations of data that were originally collected by NASS are subject to NASS review.

A.1.b. National Agricultural Statistics Service (NASS)

NASS maintains a series of Policy and Standards Memoranda (PSM) which document the policies and standards established for all of the Agency's programs. PSM 12 governs the rules of attribute and inferential disclosure along with provisions for handling special cases. PSM 7 documents NASS policy on the release of unpublished summary data and estimates and access to microdata files. PSM 6 covers the use of the list sampling frame including identity disclosure. PSM 4 presents NASS's legal obligation to protect confidential information and specifies the procedures for confidentiality certification of employees and special agents.

The Agricultural Estimates program includes crop, livestock, environmental, and economic reports that NASS regularly produces through the Agricultural Statistics Board. The Agricultural Estimates program determines primary suppressions using a threshold rule of three and the (n, k) dominance rule. The values of n and k are administratively determined and, with a few exceptions, are consistent across all publications. NASS statisticians are responsible for identifying primary suppressions and their complements, and ensuring that the suppression patterns are consistent over time. Suppressions may be presented individually or as aggregates. PSM 12 allows for the use of informed consent (waivers) for the Agricultural Estimates program if it is determined to be in the interest of the industry. All parties at risk must agree to allow the estimates to be published and have the right to revoke their consent. Agreements are renewed every five years.

For the Census of Agriculture, the Puerto Rico Census of Agriculture, the census follow-on programs including the Farm and Ranch Irrigation Survey, and the Census of Aquaculture, NASS uses the p-percent rule to identify sensitive data cells at risk of disclosure. The threshold rule is also applied to all magnitude data to ensure that a minimum number of farms are represented in each published cell. All magnitude data associated with cells with less than three farms are also suppressed. Complementary suppressions are chosen using network flow methodology. Frequency count data are not considered sensitive and not subject to suppression. Also, NASS does not allow the use of informed consent from respondents for the Census of Agriculture and its follow-on programs.

While it is NASS policy not to release microdata files, NASS operates a Data Lab within its Washington headquarters. Individual researchers may submit a research proposal and request

permission to run specialized models or tabulations on certain microdata files within the lab. Requests are addressed and approved or disapproved on a case-by-case basis by the Associate Administrator. NASS staff monitors the lab and all materials leaving the lab are subject to disclosure review. Individuals using the data lab sign confidentiality forms as NASS agents and are bound by the statutes restricting unlawful use and disclosure of data. NASS will arrange for a data lab in any of its 46 field offices, when needed. Data users may also request special tabulations through the Data Lab. These tabulations are performed by NASS staff and eliminate the need for access to microdata files. The results of each tabulation are considered public domain and are available to any data user.

NASS and the Economic Research Service cooperatively provide an interactive web tool with built-in disclosure review and filtering, that allows individual researchers to run tabulations and special analysis against microdata from the Agricultural Resource Management Survey. Access procedures mirror those of the Data Lab. Individual researchers may submit a research proposal and request an authenticated access ID. Data confidentiality is protected by applying a noise-based approach to the underlying microdata before the tabular data are generated. The parameters used for the noise creation are kept confidential. The p-percent rule is also applied to the aggregates to test a table cell for dominance from a single establishment.

NASS conducts a number of reimbursable surveys for government or academic organizations, and has developed special confidentiality procedures for these surveys. In these situations, NASS will clearly identify the sponsoring organization and purpose of the survey to respondents prior to collecting their voluntary responses. In these situations NASS may provide a microdata file, stripped of identifiers, to the sponsoring organization for their analyses. The microdata file must reside in a physically secure site under security measures approved by NASS. All individuals who will have access to the file must sign confidentiality forms as NASS agents and are bound by the statutes restricting unlawful use and disclosure of data.

In February 1993, USDA's Office of the General Counsel (OGC) reviewed the laws and regulations pertaining to the disclosure of confidential NASS data. In summary, OGC's interpretation of the statutes allows data sharing to other agencies, universities, and private entities as long as it enhances the mission of USDA and is through a cooperative agreement, cost-reimbursement agreement, contract, or memorandum of understanding. Such entities or individuals receiving the data are also bound by the statutes restricting unlawful use and disclosure of the data. NASS's current policy is that data sharing for statistical purposes will occur on a case-by-case basis, as needed, to address an approved specified USDA or public need, and under the specialized situations described above.

To the extent future uses of data are known at the time of data collection, they are explained to the respondent and permission is requested to permit the data to be shared among various users. This permission is requested in writing with a release form signed by each respondent

A.2. Department of Commerce
A.2.a. Bureau of Economic Analysis (BEA)

BEA's disclosure limitation activities pertain mainly to data that it collects on international direct investment and trade in services. These data are collected from U.S. business enterprises—both U.S.-owned and foreign-owned—in mandatory surveys conducted under authority of the International Investment and Trade in Services Survey Act (P.L. 94-472, as amended). Surveys of trade in financial services also are authorized by the Omnibus Trade and Competitiveness Act of 1988. As required by the Survey Act, the data collected are held confidential and are published in a manner that precludes the identification of individual responses. Disclosure limitation activities also are conducted for certain data on regional economic activity that are obtained from the Bureau of Labor Statistics. BLS conducts the disclosure limitation activities for its own purposes and provides a copy of the results to BEA.

With regard to the data on direct investment and trade in services, the general rule for primary suppression involves looking at the data for the top reporter, the second reporter, and all other reporters in a given cell. If the data for all but the top two reporters add up to no more than a certain percent of the top reporter's data, the cell is a primary suppression. This is an application of the p-percent rule.

This rule protects the top reporter from the second reporter, protects the second reporter from the top reporter, and automatically suppresses information in any cell with only one or two reporters. On very rare occasions, respondents may, upon request by BEA, grant a waiver of confidentiality.

When applying the general rule, absolute values are used if the data item can be negative (for example, net income). If a reporter has more than one data record in the same cell, these records are aggregated and suppression is done at the reporter level.

In addition to applying the general rule, several special rules may be applied covering rounded estimates, country and industry aggregates, and "key item" suppression (looking at a set of related items as a group and suppressing all items if the key item is suppressed).

Complementary suppression is done partly by computer and partly by human intervention. The computer programs used include routines that examine different combinations of cells to ensure that suppressions cannot be uncovered through the computation of linear combinations of rows and columns.

Some tables are published on numbers of companies, such as the number of foreign affiliates of U.S. companies in different countries or industries. These number counts are not considered sensitive and are not analyzed for disclosure or suppressed.

Under the International Investment and Trade in Services Survey Act, limited sharing of data with other Federal agencies, and with consultants and contractors of BEA, is permitted, but only for statistical purposes and only to perform specific functions under the Act. Included among these are "Special Sworn Employees", who are allowed on-site access to company-level

microdata for research purposes and who are sworn to uphold the confidentiality of the data on the same basis as regular BEA employees. Certain types of data sharing with other Federal agencies also are authorized by the Foreign Direct Investment and International Financial Data Improvements Act of 1990 and by the Confidential Information Protection and Statistical Efficiency Act of 2002. This data sharing is for statistical purposes only, and any staff of these agencies who must view BEA's unsuppressed data in connection with these activities are required to obtain BEA Special Sworn Employee status.

In another program area, BEA's Regional Economic Measurement Division publishes estimates of local area personal income by major source, based on county-level data on wages and salaries that it obtains from the Federal/state ES-202 Program of the Bureau of Labor Statistics (BLS). BEA is required to follow statistical disclosure limitation rules that satisfy BLS requirements. To prevent either the direct or the indirect disclosure of the confidential information, BEA uses the BLS state and county nondisclosure file to protect the confidential information in the ES-202 data that has been supplied to BEA. The nondisclosure file identifies the sensitive cells that must be protected to avoid release of confidential information.

BEA uses as many BLS nondisclosure cells as possible, but cannot use some of them for various reasons. The most important reasons are that the industry or geographic structure published by BEA does not exactly match the industry or geographic detail provided by BLS and that BEA does not use ES–202 data for the farm sector. For these cases, BEA must select additional cells to prevent the disclosure of confidential information. In order to determine which estimates should be suppressed, the total wages and salaries file and the wages–and–salaries–nondisclosure file are used to prepare a multidimensional matrix. This matrix is tested, and the estimates that should be suppressed are selected. Complementary suppressions, if necessary, are generated by computer and checked to ensure that they are adequate.

A.2.b. Bureau of the Census (BOC)

The Census Bureau conducts its statistical programs under government-wide legislation such as the Privacy Act, the Freedom of Information Act (FOIA), and the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002; and agency-specific legislation such as Title 13, United States Code, of 1954.

Title 13, U.S.C, defines the basis for the Census Bureau standards for confidentiality. Data that identify individuals, businesses, and other organizations must not be shared with anyone unless that person has taken an oath to maintain Census confidentiality and has a business need to know. The Census Bureau protects confidential data through the use of technological safeguards, statistical data protection, and through restricted access. Methods used include encryption software, special dedicated lines, as well as password and firewall techniques.

The Census Bureau has legislative authority to conduct surveys for other agencies under either Title 13 or Title 15 U.S.C. A sponsoring agency with a reimbursable agreement under Title 13 can use samples and sampling frames developed for the various Title 13 surveys and censuses. This would save the sponsor the extra expense that might be incurred if it had to develop its own

sampling frame. However, the data released to an agency that sponsors a reimbursable survey under Title 13 are subject to the confidentiality provisions of any Census Bureau public-use microdata file or tables; for example, the Census Bureau will not release either identifiable microdata or small area data. The situation under Title 15 is quite different. In conducting surveys under Title 15, the Census Bureau may release identifiable information, as well as small area data, to sponsors. However, sources other than surveys and censuses covered by Title 13 must be used to draw the samples. When the sponsoring agency furnishes the frame, the data are collected under Title 15, and the sponsoring agency's confidentiality rules apply.

A Disclosure Review Board (DRB) reviews specifications and proposals relating to each Title 13 data release intended for public use. The DRB ensures adherence to guidelines of the "Census Bureau DRB checklist" and any other criteria previously established by the DRB. It communicates disclosure limitation policy to program managers, Census Bureau officials, data users, prospective sponsors and the general public. The DRB initiates and coordinates research on the disclosure potential in microdata, tabular data, and other statistical outputs; and on the effectiveness of disclosure avoidance techniques as applied to such outputs. Members of the Disclosure Avoidance Research Group in the Statistical Research Division conduct research into the most suitable data protection methods for the materials published.

Some mechanisms exist to provide access to more detailed information on a restricted basis. These include Research Data Centers for approved researchers with Special Sworn Status, as well as remote on-line access in State Data Centers and Census Information Centers via the Advanced Query System for user-defined tables from Census 2000. The latter system allows users to request certain types of tables and then automatically reviews the tables to avoid disclosing confidential information. Users receive only the tables that have passed disclosure review.

Some microdata are accessible to approved researchers at the Census Bureau's Research Data Centers (RDCs). The objective of the Center for Economic Studies (CES) and the RDCs is to increase the utility and quality of Census Bureau data products. Use of microdata can address important policy questions without the need for additional data collections. In addition, it is the best means by which the Census Bureau can check on the quality of the data it collects, edits, and tabulates. These secure research facilities are located at various sites across the country. Access is strictly limited to researchers and staff authorized by the Bureau of the Census. All analysis must be performed within the secure RDC research facility. Ensuring security at RDCs has several aspects: project oversight, a physically secure facility, personnel security, a secure computing environment, an on-site Census employee, and application of disclosure avoidance rules to the analytical results presented to the public.

For the every-fifth-year economic census and associated surveys, the Census Bureau uses the p% rule to identify sensitive cells in tables but does not publish the value of p. Sensitive cells are suppressed and complementary suppressions are identified using the technique of network flow (which may be viewed as a special case of linear programming) which is computationally very fast, or linear programming which is slower. Network flow is ideal for 2-dimensional tables. It has also been applied to 3D tables although for such tables, linear programming is the preferred

method from a theoretical point of view; i.e. full protection of sensitive cells is guaranteed, obviating the need to run a disclosure audit program to check the extent of protection achieved.

For the 2002 Economic Census, network flow was used for all 2-dimensional tables and the larger 3-dimensional tables. Suppression programs based on linear programming were used for smaller 3-dimensional tables. Certain surveys have 4-dimensional or 5-dimensional data, and linear programming based programs may be used for these tables if runtimes are not excessive. Auditing programs are used when necessary.

For non-census demographic data, the Census Bureau primarily uses a combination of geographic thresholds, population thresholds and coarsening. Microdata cannot show geography below a population of 100,000. For the most detailed microdata, that threshold is raised to 250,000 or higher. Some surveys tabulate only at state, region or Census division. For data products that fall outside the main publications, a threshold may be applied at the cell level or to the population. Multi-dimensional tabular data on specific populations must met a minimum of unweighted cases, usually 50. The cell threshold minimum most frequently used is 3 unweighted individuals from 3 distinct households. Coarsening is used to avoid the application of thresholds. For small populations or rare characteristics noise may be added to identifying variables, data may be swapped, or an imputation applied to the characteristic. Census data, which lacks the component of protection provided by sampling, employs targeted swapping in addition to the combination of table design and thresholds described above.

Most of the Census Bureau's current statistical disclosure limitation practices and research are summarized in three papers Zayatz (2002), Zayatz, Massell, and Steel (1999) Hawala, Zayatz, and Rowland (2004). Other references are found in these three papers.

A.3. Department of Education: National Center for Education Statistics (NCES)

The National Center for Education Statistics (NCES) has strong legislation that requires the agency to protect the confidentiality of its data collections. First under the 1988 Hawkins-Stafford Elementary and Secondary School Improvement Amendments, and then under the 1994 National Education Statistics Act, NCES was required to maintain confidentiality of all individually identifiable data about individuals (e.g., principal, teacher or student data). Although the law did not explicitly protect institutional data, protecting data about individuals within institutions frequently resulted in the protection of data about educational institutions as well. The Education Sciences Reform Act of 2002 explicitly requires NCES to protect the confidentiality of all individually identifiable data about students, their families and their schools. Related to these laws, NCES has a statistical standard on maintaining confidentiality (NCES Statistical Standard 4-2 http://nces.ed.gov/statprog/2002/std4_2.asp). That standard summarizes the relevant laws, identifies employee and contractor responsibilities when handling confidential data, describes alternative methods that may be used to protect NCES data from disclosure, and includes the consent notice to be placed on NCES public use data files. In addition, the NCES Disclosure Review Board (DRB) reviews disclosure analysis plans and proposed public-use data releases to protect the confidentiality of the individual reported values.

Most NCES data collections include some institution data, but additionally include data from any combination of institution heads, teachers, librarians, students or student's parents. It's the individual's data that must be protected. These datasets can be made publicly available through either a public-use file or a data analysis system (DAS) after applying a DRB approved disclosure analysis and resolving any observed disclosure risks. This process is described below.

A public-use file is a file or series of linked files that: 1) contain individuals' responses about themselves, and 2) have gone through a DRB approved disclosure analysis. All direct individually identifiable information (e.g., school name, individual name, addresses) is stripped from the public-use file. Continuous variables are top and bottom coded to protect against identification of outliers. After this has be done, the only way a casual data intruder can identify an individual respondent is by first identifying the sampled institution for the individual.

To prevent identification of the sampled institution, all known publicly available lists of education institutions that contain institutions' names and addresses are gathered. Each list is matched with the sample file using all common variables between the two files. If an institution can be identified to within 2 other institutions, using an appropriate distance measure, then that is a disclosure risk and must be resolved before releasing the data.

If too many disclosure risks are obtained then a common variable(s) may be dropped from the public-use file, or the variable(s) may be coarsened. If there are only a few identified disclosure risks found then the appropriate action is to selectively perturb a set of the common variables until all disclosure risks are resolved. This analysis is repeated sequentially for each list file until it can be repeated for each list file without identifying any disclosure risks.

The matching analysis described above is designed to prevent the casual data snooper from determining survey respondents. It is assumed that if the institution cannot be identified then individuals within that institution also cannot be identified. However, data intruders with detailed knowledge about a sampled institution may be able to identify an institution; thereby, increasing the likelihood of identifying an individual. To reduce the likelihood of correctly doing this, additional disclosure edits are required.

Whenever institution head, teacher, student, or parent data are clustered, a subsampling of respondents is required. Data from respondents selected in this sub-sample, are reviewed using an additional disclosure edit. The edit is either: 1) a blanking and imputing, or data swapping of a sampling of sensitive items collected; or 2) a data swapping of the key identification variable of the respondent or institution. The amount of editing is set at a level high enough to protect the confidentiality of the respondent, while not compromising the analytic usefulness of the data file.

The important aspect of this edit is that all respondents have a chance of selection. Usually respondents at greater risk are given a larger selection probability. Should someone think that they have identified a respondent, they cannot be sure that the data is really for that respondent.

Another way NCES distributes data is through a Disclosure Avoidance System (DAS). A DAS is a table generator program that can generate proportions, means, or correlation coefficients with the corresponding standard errors that have been calculated taking into account the complex

sampling procedures used in the NCES surveys. The DAS is linked to a data file, but all data elements are masked so that the file itself is unreadable to anything or anyone other than the table generator program. The data are also protected through the survey sampling process (i.e., any unit selected is likely to have many other similar units in the universe). However, since there is little control on the type and number of tables generated, further disclosure protections are applied through data perturbation (e.g., data swapping) and data coarsening.

In order for a DAS to be released, the underlying data file must include a series of DRB confidentiality edits: either a blanking and imputing, or data swapping of a sampling of sensitive items collected; or a data swapping of the key identification variable of the respondent or institution.

All NCES tables use either a perturbation technique (i.e. a confidentiality edit approach), or a process of collapsing cells until all cells contain values associated with at least three respondents. The confidentiality edit approach is applied to the restricted-use microdata file. The table can then be prepared with no additional disclosure limitation method applied.

A.4. Department of Energy: Energy Information Administration (EIA)

EIA has established statistical standards (http://www.eia.doe.gov/smg/Standard.pdf) including standards for data protection, accessibility, and nondisclosure. Standard 2002-22, "Nondisclosure of Company Identifiable Data in Aggregate Cells," contains the procedures and policies to ensure that sensitive data cell values are suppressed (i.e., withheld from public release) for the protection of confidential survey data. EIA also requires additional confidentiality training for those who have access to data protected under CIPSEA.

EIA's primary method for ensuring confidentiality protection is the application of the pq rule or a combination rule. Regardless of the parameters chosen, the rule assures that nonzero value data cells must be based on three or more respondents. The combination rule is the pq rule in conjunction with some other subadditive linear suppression rule. The value of the pq sensitivity parameter represents the maximum permissible gain in information when one company uses the published cell total and its own value to create better estimates of its competitors' values. The values of the pq parameter that are selected for specific surveys are not published and are considered confidential. Complementary suppression is applied to other cells to assure that the sensitive value cannot be reconstructed from published data. For information collected under a pledge of confidentiality, EIA does not publicly release names or other identifiers of survey respondents linked to their submitted data.

For many EIA surveys that use the pq rule, complementary suppressions are selected manually. One survey system that publishes complex price and volume tables for crude oil and refined petroleum products uses software to select complementary suppressions. It assures that there are at least two suppressed cells in each dimension, zero value cells are excluded as candidates for suppression, and that the cells selected are those of lesser importance to data users.

Standard 2002-22 also includes separate supplementary materials with guidelines for understanding and implementing the pq rule. Guidelines are included for situations where all

values are negative; some data are imputed; published values are net values (the difference between positive numbers); and the published values are weighted averages (such as volume weighted prices). Much of the same information is provided in Appendix A of this report.

In selected program areas, EIA does not use disclosure limitation methods on statistical data. For certain energy supply data, the number of companies providing information is relatively small and/or the distribution of energy supply companies is highly skewed with a relatively small number of large companies. Statistical data for sub-United States geographical areas (e.g., States, Petroleum Administration for Defense Districts, Refining Districts) typically include some values that are sensitive and would not be published if disclosure limitation methods were applied. If disclosure limitation methods using primary and complementary suppression were applied, the result would be a significant amount of information loss. This loss of information to data users would seriously erode the value of the information for public and private understanding and analysis of energy supply.

In these program areas, EIA uses a Federal Register notice to announce a proposed policy of not using disclosure limitation methods and requests public comments. After considering public comments, EIA decides whether to formalize its policy. If the policy is to not use such methods, EIA explains the policy at the time an information collection undergoes the Office of Management and Budget approval process and when the survey materials are provided to potential respondents at the time information is requested. The explanation states that disclosure limitation procedures are not applied to the statistical data published from that survey's information. The explanation goes on to state that there may be some resulting statistics that are based on data from fewer than three respondents, or that are dominated by data from one or two large respondents. In these cases, it may be possible for a knowledgeable person to estimate the information reported by a specific respondent.

EIA does not have a standard to address tables of frequency data. However, there are only two primary publications of frequency data in EIA tables. Those publications are the Household Characteristics publication of the Residential Energy Consumption Survey (RECS) and the Building Characteristics publication of the Commercial Building Energy Consumption Survey (CBECS). In both publications, cells are suppressed for accuracy reasons, not for disclosure reasons. For the first publication, cell values are suppressed if there are fewer than 10 respondents or the Relative Standard Errors (RSE's) are 50 percent or greater. For the second publication, cell values are suppressed if there are fewer than 20 respondents or the RSE's are 50 percent or greater. No complementary suppression is used.

EIA does not have a standard for statistical disclosure limitation techniques for microdata files. The only microdata files for confidential data released by EIA are for RECS and CBECS. In these files, various standard statistical disclosure limitation procedures are used to protect the confidentiality of data for individual households and buildings. These procedures include: eliminating identifiers, limiting geographic detail, omitting or collapsing data items, top-coding, bottom-coding, interval-coding, rounding, substituting weighted average numbers (blurring), and introducing noise through a data adjustment method which randomly adjusts respondent level data within a controlled maximum percentage level around the actual published estimate. After applying the randomized adjustment method to the data, the mean values for broad population

groups based on the adjusted data are the same as the mean values generated from the unadjusted data.

A.5. Department of Health and Human Services

A.5.a. Agency for Healthcare Research & Quality (AHRQ)

The disclosure limitation procedures used by AHRQ are similar to those of NCHS. The Medical Expenditure Panel Survey (MEPS) conducted by AHRQ utilizes the National Health Interview Survey as its sampling frame. Therefore, the disclosure limitation procedures used by AHRQ for MEPS public use data files follow the procedures used by NCHS for the MEPS. All public use data file releases are required to be reviewed and approved by the NCHS Disclosure Review Board before they are released. AHRQ also reviews and cross clears release of public use files from the NHIS.

AHRQ has established an on-site data center within the Center for Financing, Access, and Cost Trends (CFACT) to facilitate researcher access to selected non-public use MEPS data.

The CFACT Data Center is a physical space at AHRQ located in Rockville, Maryland where researchers, with approved projects are allowed access to data files not available for public dissemination. These data are classified as "restricted" and contain information that are not released to the public. These data sets may contain geographic variables at a lower level than released for public use, more detailed condition information, or may consist of unedited data base segments not yet prepared for public release. These restricted data sets do not contain information that directly identifies a respondent (name, social security number, street address).

Researchers are allowed access only to the information required to complete their project. No researcher can remove any materials from Data Center until the materials have been reviewed by specific CFACT staff for disclosure avoidance. Only summary output (tables, equations) may be removed from the Data Center. No microdata files are permitted to be removed from the Data Center.

All materials to be removed from the data center are subject to disclosure review. CFACT staff is responsible for insuring the confidentiality of data being used in the data center. In the case of onsite users, CFACT staff reviews output or tables prior to the material leaving the Data Center. In the case of researchers using the Data Center remotely, CFACT staff will conduct a disclosure review of material before forwarding output to the researcher. The development of formal criteria for review of tabular materials is an ongoing process.

For users, the Manager of the CFACT Data Center is the point of contact for arbitration of confidentiality review. Every attempt will be made to work with the researcher to develop specifications for tabulations that will "pass" a confidentiality review. Projects with continuing confidentiality issues will be discussed with CFACT senior staff before a final decision is rendered.

Any output that could potentially identify respondents or small geographic areas, either directly or inferentially cannot be removed from the data center. Tables with geographic areas as one of the tabs (except for those identified on public use files) cannot be removed, nor can tables containing cells with less than 100 observations. Data Center Users are never given access to files with direct identifiers such as name or address. Users may be given access to files with dummy codes for places. However, since data center users have no need to discern the identity of the places, they will not be given the key that would allow the association of a place name with the code. Upon request the entire file can be pre-coded into categories (i.e. residing in a state with high/middle/low Medicaid generosity). Models using geographic area as the dependent variable cannot be removed from the Data Center. The identity of sampling units, which could assist in the identity of the data subject, cannot be removed. In general, any direct or inferential identities not revealed on public use data files cannot be removed from the Data Center.

A.5.b. National Center for Health Statistics (NCHS)

NCHS is the principal federal agency that releases health statistics. It is part of the Department of Health and Human Services Centers for Disease Control and Prevention (CDC). CDC's NCHS statistical disclosure limitation techniques are presented in the NCHS Staff Manual on Confidentiality (September, 2004), Section 9 "Avoiding Inadvertent Disclosures Through Release of Microdata " and Section 10 "Avoiding Inadvertent Disclosures in Tabular Data". No magnitude data figures should be based on fewer than five cases and an (n, k) rule is used. Commenting on an earlier edition of the NCHS Manual, Jabine (1993b) states that "the guidelines allow analysts to take into account the sensitivity and the external availability of the data to be published, as well as the effects of nonresponse and response errors and small sampling fractions in making it more difficult to identify individuals." In almost all survey reports, no low level geographic data are shown, substantially reducing the chance of inadvertent disclosure. The NCHS staff manual states that for tables of frequency data a) "in no table should all cases of any line or column be found in a single cell"; and b) "in no case should the total figure for a line or column of a cross-tabulation be less than 5". One acceptable way to solve the problem (for either tables of frequency data or tables of magnitude data) is to combine rows or columns, or to use cell suppression (plus complementary suppression). Other approaches are in development.

It is NCHS policy to make microdata files available to the scientific community so that additional analyses can be made for the country's benefit. Such files are reviewed for approval by the NCHS Disclosure Review Board following guidance and principles contained in the Staff Manual and the NCHS Checklist for the Release of Micro Data Files. These guidelines require that detailed information that could be used to identify individuals (for example, date of birth) should not be included in microdata files. The identities of geographic places and characteristics of areas with less than 100,000 people are never to be identified and it may be necessary to set this minimum at a higher number if research or other considerations so indicate. Information on the drawing of the sample that could identify data subjects should not be included.

All new microdata sets must be reviewed for confidentiality issues and approved for release by the NCHS Confidentiality Officer who consults with the NCHS Disclosure Review Board in making agency decisions.

Upon successful application to the NCHS Research Data Center, researchers may be provided access to special files that do not permit the identification of individual respondents. This may take place on site at NCHS offices or remotely over secure electronic lines. While information concerning named geographic entities cannot be accessed, data ordered by such units can be analyzed at a level not possible with public use data.

Prospective researchers must submit a research proposal that is reviewed and approved by a committee whose judgment is based upon the availability of RDC resources, consistent with the mission of NCHS, general scientific soundness, and the feasibility of the project. Although researchers sign confidentiality agreements, strict confidentiality protocols require that researchers with approved projects complete their work using the facilities located within the RDC. Researchers can supply their own data to be merged with NCHS data sets. Completed by the RDC staff, the merged files are only available to the originating researcher unless written permission is given to allow access to others. Further details on NCHS' Research Data Center are available at http://www.cdc.gov/nchs/r&d/rdc.htm.

Areas under current investigation include software for balancing data quality and statistical disclosure limitation (SDL) in tabular data and enhanced procedures for SDL and disclosure risk assessment in microdata.

A.6. Department of Justice: Bureau of Justice Statistics (BJS)

The same requirements under Title 13 of the U.S.C. that cover the Census Bureau are followed by BJS for those data collected for BJS by the Census Bureau. For tabular data, cells with fewer than 10 observations are not displayed in published tables. Published tables may further limit identifiability by presenting quantifiable classification variables (such as age and years of education) in aggregated ranges. Cell and marginal entries may also be restricted to rates, percentages, and weighted counts. Standards for microdata protection are incorporated in BJS enabling legislation. Individual identifiers are routinely stripped from all microdata files before they are released for public use.

A.7. Department of Labor: Bureau of Labor Statistics (BLS)

Commissioner's Order 3-04, "The Confidential Nature of BLS Records," dated October 4, 2004, contains the BLS' policy on the confidential data it collects. One of the requirements is that:

> "Publications shall be prepared in such a way that they will not reveal the identity of any specific respondent and, to the knowledge of the preparer, will not allow information concerning the respondent to be reasonably inferred by either direct or indirect means."

A subsequent provision allows for exceptions under conditions of informed consent and requires prior authorization of the Commissioner before such an informed consent provision is used.

The statistical methods used to limit disclosure vary by program. For tables, the most commonly used procedure has two steps--the threshold rule, followed by a concentration rule. BLS programs use the p percent rule or the (n, k) rule to assess concentration depending upon program. The value of the parameters used for thresholds and various concentration rules used by BLS is not released to the public. Current practice at BLS is to replace use of the (n, k) concentration rule by the p percent rule.

For example, the Quarterly Census of Employment and Wages (QCEW), a census of monthly employment and quarterly wage information from Unemployment Insurance filings, uses a threshold rule and the p percent rule for calendar year (CY) 2002 data and beyond. Prior to CY 2002, QCEW used a threshold rule and a concentration rule of (n, k). In a few cases, a two-step rule is used--an (n, k) rule for a single establishment is followed by an (n, k) rule for two establishments. The Survey of Occupational Injuries and Illnesses is using a threshold rule and the p percent rule for the CY 2003 data replacing the threshold rule used in conjunction with a concentration rule of (n, k).

The National Compensation Survey uses an approach that combines two threshold rules and an (n, k) rule. The threshold rules require that each estimate be comprised of establishments from at least m companies (unweighted) and that there are at least t distinct occupational selections (unweighted). It also uses an (n, k) concentration rule, which requires that the weighted employment among all establishments contributing to the estimate that are part of n companies cannot exceed k percent of the weighted employment of all establishments contributing to the estimate.

The Consumer Price Index Program uses a combination of a threshold rule and a minimum number of quotes from distinct sample units. The Producer Price Index uses a threshold rule on units and quotes in conjunction with the (n, k) rule.

BLS releases very few public-use microdata files. Most of these microdata files contain data collected by the Bureau of the Census under an interagency agreement and Census' Title 13 authority. For these surveys (Current Population Survey, Consumer Expenditure Survey, and four of the five surveys in the family of National Longitudinal Surveys) the Bureau of the Census determines the statistical disclosure limitation procedures that are used. Disclosure limitation methods used for the public-use microdata files containing data from the National Longitudinal Survey of Youth, collected under contract by Ohio State University and the National Opinion Research Center at the University of Chicago, are similar to those used by the Bureau of the Census.

The Bureau of Labor Statistics (BLS) has opportunities available on a limited basis for researchers from colleges and universities, government, and eligible nonprofit organizations to obtain access to confidential BLS data files for exclusively statistical purposes. These data files are derived from BLS surveys and administrative databases for which no public-use version is available. These confidential BLS data are available for research that is exclusively statistical,

with appropriate controls to protect the data from unauthorized disclosure. BLS confidential data files are available for use only at the BLS National Office in Washington, D.C., on statistical research projects approved by the BLS. Researchers granted access to the confidential data sign agreements stating that they are responsible for adhering to the confidentiality policies of the BLS.

The BLS considers applications for research proposals four times a year. Research proposals should be between 5 and 10 pages and should contain detailed information about the research project, including a literature review and an indication of how the proposed research contributes to the literature, the hypotheses to be tested, the data set and variables to be used in the analysis, the empirical methods to be used, and the specific data outputs that will result from the project.

A.8. Department of the Transportation: Bureau of Transportation Statistics (BTS)

The Bureau of Transportation Statistics (BTS) collects transportation-related data. BTS' confidentiality statutes and a set of comprehensive confidentiality procedures protect these data. The BTS *Confidentiality Procedures Manual* documents the confidentiality procedures for the agency.

BTS' confidentiality officer (CO) is responsible for the day-to-day operations of the confidentiality program. The CO also chairs the BTS' disclosure review board (DRB), which is responsible for reviewing microdata, tabular data and other information products for disclosure risks prior to public release. BTS staff and contractors are required to have annual confidentiality training, and to sign non-disclosure agreements when they enter or leave service with BTS.

BTS confidentiality program objectives guide the data review process for whether disclosure limitation methods should be applied. These objectives seek to:

- Protect confidential data while increasing access to data,
- Apply statistical disclosure limitation (SDL) methods on a case-by-case basis, and
- Take into account data user opinions on applications of SDL methods.

For most microdata and tabular data products, BTS program managers are required to complete a checklist identifying potential disclosure risks and outline any steps taken to mitigate such risk. The BTS' DRB reviews the data product and checklist and makes a final determination on disclosure risk. The DRB can recommend application of SDL methods prior to public dissemination.

BTS uses various microdata SDL methods based on the disclosure review findings and the unique characteristics of the data files. Some SDL procedures used include data suppression and modification. Data modification includes recoding continuous variables into categorical variables, collapsing categories, top and bottom coding, introduction of noise, and data swapping. BTS program managers must also identify any external data that could be matched to BTS datasets and take steps to minimize the ability to match.

The DRB conducts disclosure review of tabular data products when they are developed from microdata files that are not released to the public. BTS also uses tabular data SDL methods based on the disclosure review findings and on the characteristics of the tables.

A.9. Department of the Treasury: Internal Revenue Service, Statistics of Income Division (IRS, SOI)

The Statistics of Income (SOI) function within the larger organization Research, Analysis, and Statistics (RAS) is to establish and implement IRS guidance rules for the public release of tax data in tables and public-use microdata files. This role is primarily necessitated by sections 6108(c) and 6103j(4) of the Internal Revenue Code (IRC), which require that the data in statistical publications produced by IRS and authorized recipient agencies be anonymous.

The administrative rules are found in Chapter VI of the SOI Division Operating Manual (January, 1985), and require that at or above the state level each cell in a publicly released tabulation be based on at least three observations. Below the state level the requirement is at least ten observations. Data cells not meeting these thresholds are suppressed or combined with other cells. Combined or deleted data are included in the corresponding column totals. These rules also apply for secondary disclosure in which taxpayer identities might be revealed by subtraction of associated cells within a table or between tables, and even indirectly through similar data in other publications.

SOI documents disclosure procedures in its own publications. For example, disclosure limitations are discussed in "SOI Sampling Methodology and Data Limitations" in the Appendix to the quarterly SOI Bulletins and online at http://www.irs.gov/taxstats.

SOI produces one annual public-use microdata file, known as the SOI "tax model", containing a sample of data based on the Form 1040 series of individual tax returns. The disclosure protection procedures applied to this file include: (1) subsampling certainty records at a 33% rate; (2) removing certain records having extreme values; (3) suppressing certain fields from all records and geographical fields from high income records; (4) top coding and modifying some fields; (5) blurring some fields of high income records by locally averaging across records; and (6) rounding amount fields to four significant digits. To help ensure that taxpayer privacy is protected in the SOI tax model file, SOI has periodically contracted with experts who employ so-called "professional intruder" techniques to both verify that confidentiality is protected and to inform techniques to be applied to future releases of the SOI tax model file. For additional details on the disclosure avoidance techniques used to produce SOI public-use files see: Sailer, P., Weber, M. and Wong, W., (2001);

In addition to its own role in producing tax statistics, SOI is also responsible for coordinating the provision of tax data for statistical purposes to authorized recipients under section 6103j of the IRC. This function includes ensuring that authorized recipients of tax data also follow the rules of 3/10 described above or an equivalent methodology approved by SOI, as stipulated in the IRS Publication 1075, *Tax Information Security Guidelines for Federal, State, and Local Agencies (June 2000)*. Because of the considerable onus this requirement can entail for both SOI and agencies using alternative disclosure protection methodologies, recent efforts have begun to

establish inter-agency agreements with experienced users, such as the US Census Bureau, in which responsibility for alternative tabular protection methodologies is accepted by the recipient agency. The IRS-Census agreement for this purpose was effective June 2, 2003. Because the challenges of protecting public-use microdata files are considered unique and such data are deemed more sensitive to disclosure risk, public-use microdata files are excluded. That is, under these agreements, IRS approval would still be needed before an outside agency could release a public-use microdata file based on tax data.

Currently, the IRS Office of Research within RAS is working with Census to ensure that all data in a proposed Census public-use file based on tax data [earnings] linked to Census' Survey of Income and Program Participation (SIPP) will be anonymous. The proposed SIPP/earnings public-use file methodology is exploring using "synthetic data" to produce public-use files tailored for particular users, as opposed to a "one size fits all" approach.

A.10. National Science Foundation (NSF)

The National Science Foundation (NSF), Division of Science Resources Statistics (SRS), balances the requirement to guard the confidentiality of its respondents against the desire of the research community to access data collected using taxpayer dollars. NSF applies either the (n, k) dominance rule or p-percent rule, or sometimes both rules in conjunction with each other depending upon the survey. When it is possible to create a microdata file that is useful to a broad group of researchers while protecting respondent confidentiality, SRS releases public use data files consistent with these dual objectives. When releasing public-use microdata files, individual identifiers are removed from all records and other high risk variables that contain distinguishing characteristics are modified to prevent identification of survey respondents and their responses. Top-codes and bottom-codes are employed for numeric fields to avoid showing extreme field values on a data record. Values beyond the top-code or bottom-code are replaced either by the average of the values in excess of the respective top-code or bottom-code or through the application of various imputation methodologies.

When the researcher demonstrates that available SRS public use data files do not meet research needs and in keeping with SRS's mission to help provide the statistical information about the US science and engineering enterprise, it is sometimes possible to accommodate the request by providing access to restricted data files. One method for access is a recently created on-site secure analysis area for visiting researchers. Another method of access is off-site licensing.

Under the Office of the Director, SRS, the Chief Statistician coordinates a restricted-use data-licensing program. To acquire restricted-use files, the researcher and the researcher's institution indicates their knowledge of confidentiality issues and willingness to ensure protection of the data by completing a formal legal contract, the license agreement, that details the use of the data, promises to prevent disclosure of confidential data, agrees to a prepublication review by SRS, and stipulates the return of the data to SRS upon expiration of the license. Research conducted by licensees often is found in scientific journals as well as highly cited in policy forums.

A.11. Social Security Administration (SSA)

The Office of Research, Evaluation, and Statistics (ORES), the statistical office of the Social Security Administration, reviews and establishes methodology and procedures for protecting the confidentiality of data. For the release of statistical tables, ORES uses a strategy combining both suppression and rounding to prevent the release of identifiable information.

Statistical tables for Social Security beneficiaries and benefits consist of frequency counts for beneficiaries and summary benefit amounts. Detailed beneficiary information is suppressed when the marginal total is less than a cut-off value and only the marginal value is shown. For the rows in which only the marginal counts are shown, dollar amounts are suppressed when the number of cases contributing to the total is less than a cutoff. Detailed frequency counts are suppressed when all details for a marginal total are in a single category. When suppressions are introduced to prevent disclosure in an individual cell, complementary suppressions are employed to prevent the inference of a suppressed value. Controlled rounding is also used as a disclosure avoidance method in statistical tables for frequency counts.

Publications that include earnings and employment information conform to IRS rules when presenting tables (See section A.9 of this chapter). In particular, table cells with fewer than 3 persons at the state level and 10 persons at the county level are suppressed and the corresponding summary income is also not shown. Whenever data cells are suppressed, complementary suppressions are introduced to prevent inferring a suppressed value. All dollar amounts are shown in thousands of dollars. Earnings and employment statistics are derived from a sample of IRS records rather than a 100-percent file of earnings and employment information.

When releasing public-use microdata files, individual identifiers are removed from all records and other distinguishing characteristics are modified to prevent identification of persons to whom a record pertains. Records are sequenced in random order to avoid revealing information due to the ordering of records on the file. Top-codes and bottom-codes are employed for numeric fields to avoid showing extreme field values on a data record. Values beyond the top-code or bottom-code are replaced by the average of the values in excess of the respective top-code or bottom-code. Top-code and bottom-code values are derived at the national level and the replacement values are derived and applied at the state level when appropriate. Values shown for some categorical fields are combined into broader groupings than those present on the internal file and dollar amounts are rounded. Top-code and bottom-code values, replacement values, and related information are provided to users as part of file documentation.

A Disclosure Review Board (DRB) reviews proposed public-use microdata files prior to their release. The DRB consists of staff from ORES who are familiar with the underlying data files, their uses, and confidentiality requirements. In addition, confidentiality specialists from other federal agencies may serve on the DRB to provide further perspective and additional confidentiality expertise. Staff who are responsible for file creation complete the *Checklist on Disclosure Potential of Proposed Data Releases,* prepared by the Interagency Confidentiality and Data Access Committee, and the Checklist is included in the DRB review.

B. Summary

Most of the 14 agencies covered in this chapter have standards, guidelines, or formal review mechanisms that are designed to ensure that adequate disclosure analyses are performed and appropriate statistical disclosure limitation techniques are applied prior to release of tabulations and microdata. The agency standards and guidelines exhibit a wide range of specificity: Some contain only one or two simple rules while others are much more detailed. Some agencies publish the parameter values they use, while others feel withholding the values provides additional protection to the data. Obviously, there is great diversity in policies, procedures, and practices among Federal agencies to appropriately protect the wide variations in the content and format of information released.

B.1. Magnitude and Frequency Data

Most standards or guidelines provide for minimum cell sizes and some type of concentration rule. Some agencies (for example, ERS, NASS, and NCHS) publish the values of the parameters they use in (n, k) concentration rules, whereas others, such as Census and BLS, do not. Minimum cell sizes of 3 are routinely used, because each member of a cell of size 2 could derive a specific value for the other member. Some agencies cited accuracy standards as guidelines for releasing certain tabular data. **Accuracy standards** refer to specific rules that an agency applies to the data that relate to some measure of data quality such as a threshold level for relative standard error or coefficient of variation estimates.

Most of the agencies that published their parameter values for concentration rules used a single set, with n = 1. Values of k ranged from 0.5 to 0.8. The most elaborate rule included in standards or guidelines were EIA's pq rule and BEA's and Census Bureau's related p-percent rules. All these rules have the property of subadditivity. The p percent and pq rule give the disclosure analyst flexibility to specify how much gain in information about its competitors by an individual company is acceptable.

One possible method for dealing with data cells that are dominated by one or two large respondents is to ask those respondents for permission to publish the cells, even though the cell would be suppressed or masked under the agency's normal statistical disclosure limitation procedures. Agencies including NASS, EIA, the Census Bureau, and some of the state agencies that cooperate with BLS in its Federal-state statistical programs, use this type of procedure for some surveys to allow publication of those sensitive cell values. Another disclosure limitation method used by two agencies is to apply noise to the underlying micro data before aggregating the reported values.

B.2. Microdata

The agencies that release public use microdata files have established statistical disclosure limitation procedures for releasing microdata. Some agencies noted that the disclosure limitation procedures for surveys they sponsored were set by the Census Bureau's Disclosure Review Board, because the surveys had been conducted for them under the Census Bureau's authority (Title 13). Major releasers of public-use microdata--Census, NCHS and NCES--have all

established formal procedures through Disclosure Review Boards for review and approval of new microdata sets. As Jabine (1993b) wrote, "In general these procedures do not rely on parameter-driven rules like those used for tabulations. Instead, they require judgments by reviewers that take into account factors such as: the availability of external files with comparable data, the resources that might be needed by an `attacker' to identify individual units, the sensitivity of individual data items, the expected number of unique records in the file, the proportion of the study population included in the sample, the expected amount of error in the data, and the age of the data."

Geography is an important factor. Census and NCHS specify that no geographic codes for areas with a sampling frame of less than 100,000 persons can be included in public-use data sets. If a file contains large numbers of variables, a higher cutoff may be used. The inclusion of local area characteristics, such as the mean income, population density and percent minority population of a census tract, is also limited by this requirement because if enough variables of this type are included, the local area can be uniquely identified. An interesting example of this latter problem was provided by EIA's Residential Energy Consumption Surveys, where the local weather information included in the microdata sets had to be masked to prevent disclosure of the geographic location of households included in the survey.

Top-coding is commonly used to prevent disclosure of individuals or other units with extreme values in a distribution. Dollar cutoffs are established for items like income and assets and exact values are not given for units exceeding these cutoffs. Blurring, swapping, blank and impute, noise introduction, recoding, threshold rules, and rounding are other methods commonly used to prevent disclosure.

Summary of Agency Practices

Agency	Magnitude Data	Frequency Data	Microdata	Waivers	Restricted Access Allowed for Researchers
ERS	(n, k), (1,.6) 3+	Threshold Rule 3+	No	Yes	Yes
NASS	(n, k), p-percent Parameters Confidential	1+ Not Sensitive for Est. Surveys	No	Yes	Yes
BEA	p-percent c=1	1+ Not Sensitive for Est. Surveys	No	No	Yes
CENSUS	p-percent Parameters Confidential Noise addition	Data Swapping, Access Query System rules, Threshold Rule	Yes -- Disclosure Review Board	Yes	Yes
NCES	Data Swapping Data Coarsening Accuracy	Data Swapping Data Coarsening Accuracy	Yes – Disclosure Review	No	Yes

Agency	Magnitude Data	Frequency Data	Microdata	Waivers	Restricted Access Allowed for Researchers
	Standards/Threshold Rule 3+	Standards/Threshold Rule 3+	Board		
EIA	(n, k), pq, Parameters Confidential	Threshold Rule Accuracy Standards	Yes – Office Review	Yes	No
NCHS	(n, k), (1,.6)	Threshold Rule 4+	Yes – Disclosure Review Board	No	Yes
AHRQ	N/A	Threshold Rule 4+	Yes – Disclosure Review Board	Yes – Disclosure Review Board	Yes
SSA	Threshold Rule 3+	Threshold Rule, 5+ Marginals, 3+ cells	Yes - Agency Review	No	No
BJS	N/A	Threshold Rule 10+, Accuracy Standards	Yes - Legislatively Controlled Agency Review	No	No
BLS	(n, k), p% rule, Parameters vary by survey and data element	Minimum Number varies by survey	BOC Collects Title 13	Yes	Yes
IRS	Threshold Rule 3+	Threshold Rule 3+	Yes - Legislatively Controlled	No	No
BTS	Varies by data	Threshold Rule 3+	Yes – Disclosure Review Board	No	No
NSF	(n, k) and/or p as appropriate	Varies by risk	Yes – Meet or exceed Census public use products which are merged	Yes	Yes

Notes: Details of specific methodologies being used are shown in this table and discussed in the text to the extent they were included in the individual agencies' responses. Rules shown in the various table cells (p-percent, (n, k), for example) are explained in the text.

The following page contains a brief explanation of the key terms used in the table.

The Threshold Rule: With the threshold rule, a cell in a table of frequencies is defined to be **sensitive** if the number of respondents is less than some specified number. Some agencies require at least 5 respondents in a cell, others require 3. Sometimes, the threshold rule is applied to the universe of a table. For example, a minimum size may be needed to publish values in all cells of a table. An agency may restructure tables and combine categories or use cell suppression, random rounding, or controlled rounding. The "+" notation (3+ for example) means at least that many non-zero observations must be present for the cell to be published. (See Section II.C.3)

Data Swapping is the procedure that was used by the U.S. Census Bureau to provide protection in data tables prepared from the 2000 Census. The technique applies statistical disclosure avoidance to the microdata records before they are used to prepare tables. The adjusted microdata files are not released, they are used only to prepare tables. For both the 100 percent data file and the sample, a small sample of households were selected and matched with households in other geographic regions that had identical characteristics on a set of selected key variables. Most variables in the matched records were interchanged. This technique is called swapping. The key variables used for matching were selected to assure that Census aggregates mandated by law would be unchanged by applying this procedure. NCES recommends using data swapping and coarsening for all internal and external microdata records. If these techniques are not used, NCES prohibits the publication of any cells with fewer than three cases and prohibits the use of cell suppression. Tabulations must be reconfigured until there are no remaining cells with fewer than 3 cases

The p-Percent Rule: Approximate disclosure of magnitude data occurs if the user can estimate the reported value of some respondent too accurately. Such disclosure occurs, and the table cell is declared sensitive, if upper or lower estimates for the respondent's value are closer to the reported value than a pre-specified percentage, p. This method assumes that before data are published a user can estimate the true value to within plus or minus 100%. This rule is referred to as the "p-percent estimation equivocation level" in Statistical Policy Working Paper 2, but it is more generally referred to as the **p-percent rule**. (See Section IV.B.1.a)

The pq Rule: The pq rule is similar to the p% rule, but assumes that before data are published the general public can estimate a company's data to within q% (where q<100). Hence, an agency can specify how much prior knowledge there is by assigning a value q which represents how accurately respondents can estimate another respondent's value before any data are published ($p < q < 100$). (See Section IV.B.1.b)

The (n, k) Rule: The **(n, k) rule**, or dominance rule was described as follows in Statistical Policy Working Paper 2. "Regardless of the number of respondents in a cell, if a small number (n or fewer) of these respondents contribute a large percentage (k percent or more) of the total cell value, then the so-called **n respondent, k percent rule** of cell dominance defines this cell as

sensitive." Many people consider this to be an intuitively appealing rule, because, for example, if a cell is dominated by one respondent then the published total alone is a natural upper estimate for the largest respondent's value. (See Section IV.B.1.c)