The "Checklist on Disclosure Potential of Proposed Data Releases"

Virginia A. de Wolf, OMB; Alvin Zarate, NCHS; Laura Zayatz, Bureau of the Census

Federal statistical agencies and their contractors often collect data under a pledge of confidentiality. Before disseminating results as either public-use microdata files or tables, agencies should apply statistical methods to limit disclosure of the data. A discussion of such methods can be found in the Federal Committee on Statistical Methodology's (FCSM) 1994 report, *Report on Statistical Disclosure Limitation Methodology*, Statistical Policy Working Paper (SPWP) # 22. One recommendation in SPWP # 22 was that agencies should centralize their method for reviewing disclosure limitation procedures and products. The "Checklist on Disclosure Potential of Proposed Data Releases" (referred to as Checklist) is one tool that can assist in such reviews.

This paper contains four sections. The first describes the history of the Checklist. The second contains a list of suggested uses. The third describes the three main components of the Checklist. The last section has a short summary and description of future plans.

■ Background

Federal statistical agencies and their contractors often collect data under a pledge of confidentiality. They collect data from a wide variety of sources—persons, businesses, or other types of organizations. Before disseminating results as either public-use microdata files[1] or tables, agencies apply statistical methods to protect the confidential information that they collect. A review and evaluation of the statistical disclosure limitation techniques used by Federal statistical agencies can be found in FCSM's SPWP # 22. In addition, SPWP # 22 contains a set of 12 recommendations to improve disclosure limitation practices.

In discussing the recommendation that agencies should centralize their review of disclosure-limited data products, 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 paper, the term Disclosure Review Board refers to formal agency disclosure review even though such a review might be handled by one person in the agency. The "Checklist on Disclosure Potential of Proposed Data Releases" (called Checklist) is one tool that can assist in such reviews.

The Checklist consists of a series of questions designed to assist an agency's Disclosure Review Board in determining the suitability for release of microdata files and tabular data collected from individuals and organizations under an assurance of confidentiality. It contains three main sections. The first section pertains to microdata files that contain information from individuals or establishments, while the next two sections refer to tabular data from individuals and establishments, respectively. In creating this Checklist, the Interagency Confidentiality and Data Access Group (ICDAG) has liberally borrowed descriptions and definitions from SPWP # 22.

This Checklist is based on one that is used at the U.S. Bureau of the Census. In 1996, it was modified and adopted for use by the National Center for Health Statistics. Staff at the Bureau of Labor Statistics expanded the section on tabular products from establishments and organizations. ICDAG members, who represent about 20 different Federal statistical agencies (for a description of ICDAG, see de Wolf, 1997), became very interested in broadening the applicability of the Checklist. Along the way, many members contributed to its development and refinement.

■ The Checklist

The Checklist begins with a cover sheet that asks for basic information about the proposed data release. It has three main sections:

Microdata

Most microdata files contain demographic information. *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 if located within a single State, but more respondents—especially those with rare and visible reported characteristics—could be identified if located 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 to be released). This auxiliary information is often difficult to assess for its disclosure risk. "Coarsening" a data set by dropping survey variables, collapsing response categories for other variables, and/or introduction of statistical perturbation called "noise" in 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 as such entities are often selected from very skewed populations. 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 geographic information as large as a Census region.

Tabular Data from Persons or Households

This section pertains to tables based on data collected from persons or households (referred to as demographic data) 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.

The second section 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 cover especially sensitive topics.

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

The Checklist was developed with the following considerations:

1. It should be completed by a person who has appropriate statistical knowledge and is famil-

iar 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 is intended to accomplish. (Those who need a "primer" on statistical disclosure limitation methods should see Chapter 2 of SPWP # 22; or for information on the use of noise for tabular data, Evans, Zayatz, and Slanta, 1996. Two additional useful resources are the 1996 Eurostat publication and the book by Willenborg and de Waal, 1995.)

2. Responses to questions in the Checklist are not intended to supply all the information that might be 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 other potential users.

Of course, there are two sides to the application of statistical disclosure limitation procedures: limiting disclosure risk and releasing data that are analytically useful. Products that meet the criteria for public release will often not meet all user requirements. When there are several methods of meeting requirements for public release, the alternative that provides data with the greatest analytic value is clearly preferred.

There are instances when the tabular/microdata products are not adequate for certain types of research uses. To accommodate such users, some agencies have developed restricted access procedures that place conditions on who can use the data, for what purposes, at what locations, etc. (see Jabine, 1993, for an excellent description of a broad array of restricted access procedures used by U.S. statistical agencies). One example of a restricted access procedure is the Census Bureau's Research Data Centers that enable researchers to have access to establishment microdata under carefully controlled conditions at one of three data centers (the Bureau plans to add other remote sites in the near future). A similar research data center was started at the National Center for Health Statistics in 1998. Another example would be the licensing of researchers to use, at their university or research center, datasets that contain more detailed information than the "standard" public use microdata file. This approach is used by the National Center for Education Statistics.

■ Uses for the Checklist

In addition to helping an agency's Disclosure Review Board to 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; and
- ❒ 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.

Note that the Checklist reflects the current standards of the Census Bureau and the National Center for Health Statistics for the release of data for public use.

■ Summary and Future Plans

Users should complete the cover sheet and answer all questions for the applicable section(s). (Obviously, if it is distributed as a paper document, those who need more space for an answer would attach a continuation sheet and identify the number of the question.)

The Checklist is not a "fixed" document, and agencies are encouraged to modify it to suit their particular needs. With appropriate modifications, the Checklist can be adapted by Federal agencies, as well as other

organizations, and used to review material of varying levels of confidentiality. ICDAG encourages Federal agencies to tailor the Checklist, as needed, for their own uses.

The Checklist is a "work in progress" that will be changed, refined, and modified as new approaches and techniques are developed. However, there is a need to have it in a "semiformal" stage so that it can be more widely disseminated. Therefore, the Checklist will be reviewed by ICDAG's parent committee—FCSM—during the latter part of 1998. Once it is approved for broader dissemination by FCSM, ICDAG will make the Checklist available on the FCSM web site in several alternative formats (including PDF).

Comments, suggestions, and questions about the Checklist can be sent to any of the authors:

- Virginia de Wolf: (V) 202-395-7314; (F) 202-395-7245; vdewolf@omb.eop.gov;
- Alvin Zarate: (V) 301-436-6044; (F) 301-436-3503; aoz1@cdc.gov; and
- Laura Zayatz: (V) 301-457-4955; (F) 301-457-2299; laura.zayatz@ccmail.census.gov.

Those interested in obtaining a copy of the Checklist should include a mailing address.

■ Authors' Note

The views expressed in this paper are those of the authors and do not necessarily represent the views of their respective agencies (Office of Management and Budget, National Center for Health Statistics, and the Bureau of the Census).

■ References

de Wolf, V.A. (1997), The "Interagency Confidentiality and Data Access Group," *American Statistical Association, 1997 Proceedings of the Government Statistics Section and Social Statistics Section*, pp. 323-328.

Eurostat (1996), *Manual on Disclosure Control Methods* (Catalogue #: CA-94-96-283-EN-C), Luxembourg: Eurostat.

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.

Jabine, T. B. (1993), "Procedures for Restricted Access," *Journal of Official Statistics*, 9(2), pp. 537-589.

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.

Marsh, C.; Dale, A.; and Skinner, C. (September 1991), "Safe Data versus Safe Settings: Access to Customised Results from the British Census," *Proceedings of the 48th Meeting of the International Statistical Institute*, pp. 63-91.

Willenborg, L. and de Waal, T. (1995), *Statistical Disclosure Control in Practice* (Lecture Notes in Statistics 111), NY: Springer-Verlag, Inc.

■ Footnote

[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.