STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Working Paper No. 26
English only

Topic III: Attitudes of respondents towards statistical confidentiality

A HOLISTIC APPROACH TO CONFIDENTIALITY ASSURANCE IN STATISTICAL DATA

Invited paper

Submitted by the U.S. Census Bureau[1]

I. THE IMPORTANCE OF PUBLIC PERCEPTION IN DATA RELEASE

1. Increasingly, statistical data are made available to more people who have improved tools to get information relevant to their needs. As a result, governments are making it easy for citizens to know what is happening in society and the economy and the effects that government programs are having on their well being. For example, the U.S. Government Information Integrated Test-bed project (reference) intends to permit queries of statistical databases seamlessly across government agencies despite their use of different hardware, operating systems and database formats. The Census Bureau's existing FERRET system fully integrates geography with the data to permit layering of economic, demographic, and environmental characteristics (see Capps, Green and Wallace, 1999). The Census Bureau's new American FactFinder (AFF) provides users with the ability to tabulate data for user-defined variables across desired geographic boundaries. These new tools greatly enhance the usefulness of data that were collected with public funds for the public good.

2. At the same time, citizens are increasingly concerned about the many infringements on their confidentiality and privacy facilitated by the same technology that is making easy access to statistical information possible. The threats to confidentiality from hackers who break into Web-sites and steal credit card information cause the public to worry that nothing is safe in cyberspace. The public is also worried that information they provide to business or government for one purpose will be used for a very different purpose without their knowledge or consent. Stories they read in the newspaper or see on television about violations involving credit information, medical records, Social Security Numbers, and intrusions into their children's lives, make them less willing to give personal information or to give the correct information. Recent protests over the intrusiveness of questions asked in the 2000 Census (referred to as "Census 2000") in the United States demonstrate these concerns and are remarkable given that essentially the same questions were asked in the 1990 Census with much less public concern.[2]

3. National statistical offices have always been aware that the public's trust is essential to continued success in gaining cooperation in surveys. The old paradigm is to emphasize the legal assurances of confidentiality and statistical use at the point of collection and rely on past disclosure limitation successes as evidenced by the lack of publicized breaches. In today's world, technological advances are providing

[1] Prepared by Gerald W. Gates. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited Census Bureau review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion on work in progress.

[2] Following the intense privacy concerns raised over the Census 2000 long form, the Census Bureau reviewed correspondence and Congressional testimony surrounding the 1990, 1980, and 1970 censuses and, while there was a focused debate in 1970 over privacy concerns, this was not evident in 1980 or 1990.

more data dissemination opportunities and are moving statistical offices in new directions where old approaches may no longer be sufficient.

4. National statistical offices are also beginning to realize that research into public perceptions may help supplement disclosure limitation research. The U.S. Census Bureau, since the mid-1990s, has implemented a program to study perceptions of privacy and confidentiality related to collecting and linking data but primarily focused on the use of administrative records. Results from this research have proven to be useful in measuring attitudes about confidentiality as they may relate to data release decisions. In addition, because of concerns about the exposure of public data on the Internet to many new users, the Census Bureau has also undertaken cognitive research on how best to explain data protection to users. This is intended to make these messages more relevant to non-traditional users and avoid misunderstanding about confidentiality protection. In effect, this new attention to public perception is recognition that communication with respondents cannot end once the information is obtained. The information age necessitates that we listen to and inform respondents throughout the process.

II. HOW PEOPLE PERCEIVE CONFIDENTIALITY AND PRIVACY—SOME RECENT RESEARCH

5. Research done by the Census Bureau following the 1990 census, when compared with 1980 results, showed that the public's concerns about privacy and confidentiality had become increasingly important barriers to cooperation (Fay, Bates, Moore, 1991). In preparing for Census 2000, the U.S. Census Bureau undertook a program of research on public beliefs about confidentiality and privacy as they relate to Census Bureau's efforts to collect sensitive information and link data from various sources for statistical uses. This research was focused primarily on the planned expanded use of administrative records but respondents were also asked their knowledge of confidentiality protections, whether they believed that confidentiality is protected, and concerns they may have about privacy today. Results from the 1995 and 1996 surveys will be compared with results from surveys done just prior to the Census 2000 and after the census was completed.

6. Singer and Van Hoewyk have begun analyzing the findings from the 1999 and 2000 Survey of Privacy Attitudes (SPA) under a contract with the Census Bureau. Their early results have shown that people are becoming more aware of the confidentiality protections applicable to Census Bureau data, especially between 1999 and 2000, and probably because of the intense advertising campaign surrounding the census. They also found that the public's concern about confidentiality is significantly greater than it was as recently at 1996. When asked what confidentiality means to them, respondents to these surveys most often said that it means that information would not be sold, shared, or forwarded. Other responses indicated that information will not be released—it would remain private, confidential, or protected. With regard to questions pertaining to privacy and confidence in government officials, the surveys showed that people's attitudes about privacy as well as their trust in government "to do what is right" have changed little since 1995 and in fact may be improving slightly (see Singer and Van Hoewyk, forthcoming). These findings are important as they show that a steadily increasing public concern for privacy in the United States may be leveling off, perhaps due to greater government and media attention to the problem.

7. In Census 2000, surveys were also undertaken by a private organization, InterSurvey, comparing attitudes in 1990 and 2000 and changes occurring during Census 2000. The data suggest there may have been increases over the decade in the perception that people's census answers could be used against them (Martin, forthcoming). Most significantly, the measures of privacy[3] and confidentiality lead one to conclude that distinctions made by the public between privacy and confidentiality in 1990 no longer hold and that a threat to privacy is seen as a threat to confidentiality and vice versa.[4] According to Martin, "a

[3] In this context, the notion of privacy is related to the legitimacy of the request for, or release of, information in the first place. For example, if one believes that the census is an invasion of privacy they likewise contest the legitimacy of the effort to obtain the information either because it is not seen as worthwhile or because the requester is not trustworthy.

[4] The apparent contradiction between the findings of Martin that privacy and confidentiality are linked and Singer and Van Hoewyk that privacy concerns have not increased while confidentiality concerns have increased can be explained by the fact that the SPA surveys measure confidentiality only in terms of the Census Bureau whereas privacy is measured in much more general terms.

more diffuse, general, and perhaps extreme reaction to the census (and perhaps, to other perceived privacy threats) would seem to be a possible consequence of this apparent change in attitude." These results are provocative and cause us to reexamine our whole communication strategy that has been primarily focused on confidentiality.

8. To fill a gap in knowledge about the opinions of businesses about confidentiality of establishment data, the Urban Institute, on behalf of the Census Bureau, is planning a new survey of businesses for early 2001. The survey asks business respondents about their knowledge of confidentiality laws, opinions about what information is most sensitive and whether sharing of business data with other statistical offices and organizations is of concern to them. The results of this survey will be published in a book on confidentiality and disclosure research also commissioned by the Census Bureau.

III. THE ROLE OF PERCEPTIONS ON DATA RELEASE DECISIONS (THE NON-TECHNICAL COMPONENT)

9. Data collectors make decisions to release data to the public based on the likelihood that someone can be identified because they are unique in the population or because information about them is readily available to third parties (for example, from public lists). To limit reidentification, statistical disclosure limitation (SDL) techniques are applied to the data to frustrate intruders who wish to target specific individuals as well as those who are interested in targeting any individual just to prove it can be done. Once the "acceptable" level of protection is applied, the data collector confidently makes the release in accordance with its general mission to provide data to inform public policy. Individual respondents are often unaware of this process and probably not interested as long as they felt they could trust the original confidential pledge and there was no evidence that breaches had occurred.

10. This traditional paradigm is being tested in today's electronic age as respondents may be more aware of the uses of their data and are able to easily access and manipulate the results to which they have contributed. It is further complicated by the growing concern of the public that they are losing control over their personal information and that they cannot trust those in possession of their data.[5] Disclosure Review Boards have always been confronted with addressing the technical issues in deciding an acceptable level of risk for data they release. Today, this job is made harder by the development of publicly available databases and new methodologies that counter efforts to protect identities. A less understood but nagging uneasiness is also surfacing around how these technical approaches are understood and accepted by those unsophisticated users who find the data on the Internet and perceive that they are something they are not (e.g., not purely for statistical use and not disguised well enough).

11. Discussions of the Census Bureau's Disclosure Review Board over the last few years, have raised issues about various perceived threats to confidentiality that are partly due to unknown but potentially unacceptable levels of risk and partly due to how respondents might react. The following summarizes these concerns:

12. *Finding oneself on the file*— All SDL protections for microdata depend on the fact that the intruder does not know that a particular target is on the file. Sampling provides the uncertainty needed to make these data releases possible. The survey respondents necessarily know that they are on the file and can easily identify themselves or other household members even with the standard SDL procedures applied to the data. Whereas in 1990, DRBs did not view this as a serious concern, on the Internet, it is not hard to imagine a respondent looking at a file containing his own records or those of his spouse or children and assuming that others could also find them. How do we explain that only those who know someone is in the survey can find him?

13. *Single cells in tables where data have been swapped*—With the use of data swapping or switching techniques in the census, it is likely that some cells containing one observation will appear in the tables. Unless one is aware that these cells may have been swapped and that the apparent single

[5] Public opinion surveys conducted for Equifax by Louis Harris, as well as surveys conducted for the Census Bureau by Westat have shown a growing concern by the public about uses of their personal information without their knowledge or consent.

person is actually someone else, one may assume that that data reveal confidential information. How likely is this to be of concern? Can this procedure be explained adequately to all data users so that there is no misunderstanding?

14. *Multiple race reporting means up to 63 categories*—In Census 2000, for the first time the U.S. Census Bureau permitted people to record multiple races (five categories plus other). This means that 63 categories will be reported in the census whereas only five were reported in 1990. Some of these combinations will contain very few cases in some or most geographic areas. Swapping and rounding procedures are applied to protect confidentiality but will the public perceive that the protections are sufficient? This is of special concern as there has been some evidence of criticism by leaders of racial groups of individuals who report multiple races.

15. *Increase of potential risk (many more users on the Internet)*—With greater use of the Internet to provide access to Census Bureau data, the pool of likely users has grown tremendously. Likewise, the skills of users and tools available to them raise the risks that confidentiality protections will be broken. Some Disclosure Review Board members are concerned that the possible threats to the data may have grown to unacceptable levels but cannot be sure. If data collectors are worried about the risks, cannot we expect that respondents would also be concerned? How can we reassure them and ourselves that the risk is minimal (if in fact it is)?

16. *"Semi-statistical" uses (targeting groups)*—Statistical data products are designed to eliminate the possibility that individuals can be identified and targeted but, by their nature, are used to examine characteristics of groups of individuals. Some uses may appear to be non-statistical in that companies or government officials may assign characteristics to people on their own files based on geographic summaries provided by a statistical office (for example, average income assigned to targeted groups living is a certain ZIP code in the United States). Sometimes the uses are misrepresented in advertisements and appear to imply that the individual records were linked to the other party's database. How do we set the record straight without sounding defensive and raising additional concerns?

17. *Existence of Public Use Microdata Samples (PUMS) (appears too risky?)*—It is unclear what respondents know about the existence of public use microdata products. In the past there has been a limited number of users of these products because of the computer equipment needed and the skills required to process them. Traditional academic and government users are aware that these files undergo a process to prevent identifying individuals. Uninformed users who come upon PUMs products on the Internet may find it hard to understand how these files could be safe since they show very detailed responses for specific (but anonymous) individuals. If respondents happen upon PUMs, will they understand or will they worry that their confidentiality is at risk?

18. *Increased public concerns about other data collectors*—We continue to hear stories of how information brokers have violated their trust with the public by not informing them how their information will be used or not getting consent for new uses. The public is becoming concerned about their privacy and is demanding government protection. Will these concerns spill over in some way to government surveys? Will misunderstandings about the nature of our data releases cause the public to associate statistical offices with information brokers?

19. These concerns are not easily quantifiable but they do play a role in data release decisions. Each involves a possible misunderstanding and suggests the need for better communication about the nature of data products and statistical disclosure limitation procedures. It is not realistic to think we can eliminate misunderstandings or that everyone will agree with what we say; nevertheless, our research will help determine if these concerns are real and the best way to alleviate them.

IV. REACHING AN UNDERSTANDING WITH RESPONDENTS

20. When collecting personal information for statistical purposes, the data collector is required by law and/or ethical principles to disclose information related to who can see the identifiable information (confidentiality) and why the information is needed and how it will be used (privacy). Respondents are assured that no information that can be associated with them personally will be revealed in any of the data

releases made by the data collector. They are further assured that the information will be used only for statistical purposes—not to make decisions about them personally based on how they responded. National statistical offices have been made aware of the growing concern for confidentiality and privacy in the information age but the messages about confidentiality and statistical use are generally the same as they were 20 years ago—before personal computers and the Internet.

21. The messages used to establish the bond between the data collector and the respondent and lay out the conditions of access and use are conveyed at the time of collection. Once the data are collected and verified there is little, if no, contact between the collector and provider about how the data are protected and used (Gates and Groves, 1992). The initial agreement to participate and any future agreements depend entirely on trust. Voluntary cooperation (the cornerstone of survey research) works only as long as the respondent continues to believe that the data collector is reputable and there is no evidence to the contrary. In today's electronically connected, privacy-sensitive society there are frequent reminders that trust can be misplaced. In the statistical arena, there are also opportunities to test this assumption on the backend when the results are published. The failure of data collectors to maintain communication between collection and dissemination may contribute to a situation where simple misunderstandings can damage the trust built up over the years.

22. Of course, it is not easy to maintain contact with respondents and even harder to convey descriptions of technical procedures that do not alarm people unnecessarily. I have even argued in the past that technical procedures contributing to the intended statistical use are less relevant to respondents and that, so long as the statistical office is confident the respondent is informed about permitted uses and legal protections, descriptions of technical procedures are unnecessary (Butz and Gates, 1986). In the year 2001, we can no longer avoid making some statement about technical procedures. The goal is to make these messages clear, understandable, and relevant. There is also an issue of when and how to deliver the message to increase the chance that we will succeed in this goal. Previously, we looked at the point of collection as the only opportunity to deliver messages to respondents. Primarily, this is done as part of the introductory letter sent to the respondent prior to the interview. But, messages on SDL procedures are less relevant at that stage of the process. With the current technology, we now have an easy means of communicating this information to data users, including the respondent, as they view the survey results on the Web.

23. With the Census 2000 results planned for release on the Web through the Census Bureau's new Internet Data Dissemination tool, the American FactFinder (AFF), the Census Bureau decided it needed to aggressively deal with possible misperceptions about the confidentiality and quality of the data. Confidentiality messages were prepared to describe in general the confidentiality law and penalties for violations. Additional messages were developed to describe the technical steps taken to ensure confidentiality. To determine how these messages are received and interpreted, we recruited ten individuals to participate in interview sessions to test the meaning of specific terms used in two versions of the messages (one general and one more detailed). The goal was to come up with the optimal set of messages that conveys meaningful information that AFF users find relevant (Mayer, 1999).

24. The research showed that an intermediate-length message was preferred to the two we developed. In addition, the results suggested that this new message contain hyperlinks to additional information about complex confidentiality terms or procedures. In that way users who are knowledgeable or most concerned can dig deeper to get the answers they want. Also, we found we needed to revise and eliminate some references that were not meaningful or could be misinterpreted. For example, concepts that were particularly difficult and needed to be better developed included: public use microdata, magnitude data, primary suppression, complementary suppression, tract level and above, visibility within the population, and sparseness of cells in the table. One interesting finding was that some participants were concerned that suppression may be seen as a malicious act to conceal data rather than a benevolent means of ensuring confidentiality.

25. In March 2001, the explanation of confidentiality was revised on the basis of cognitive testing. The new confidentiality message and related links are as follows:

Confidentiality

The Census Bureau has modified some data on this site to protect confidentiality. Title 13 United States Code, Section 9, prohibits the Census Bureau from publishing results in which an individual's or business' data can be identified.

The Census Bureau's internal Disclosure Review Board monitors the disclosure review process and sets the confidentiality rules for all data releases. A checklist approach is used to ensure that all potential risks are considered and addressed. A list of possible concerns is created and the Disclosure Review Board makes sure that the appropriate steps are taken to ensure the confidentiality of the data.

For more information on how the Census Bureau protects the confidentiality of data, please explore the following links.

Disclosure limitation procedures
Suppression
Data Swapping
Protection of Microdata

Questions about confidentiality may be addressed to webmaster@census.gov Attention

_____

26. The reader is encouraged to visit the AFF site at factfinder.census.gov to view the placement of these messages and the descriptions of the linked terms.

27. One aspect that has not been tested is whether privacy messages are needed at the dissemination stage. As mentioned previously, people are more clearly associating privacy with confidentiality and are less likely to believe the confidentiality pledge if the survey is perceived to be an infringement on privacy. One aspect of privacy is the feeling that the request for information is legitimate. Even thought it may be seen as intrusive, if the use is considered legitimate and of value, the individual may agree to participate if confidentiality is assured (Groves, Cialdini, Couper, 1992). Given this interaction between privacy and confidentiality, we need to determine if the messages we use to achieve an understanding with the respondent on confidentiality should also provide reassurance that the data we are releasing support legitimate and worthwhile interests of the respondent (e.g., better targeted government programs, new neighborhood stores).

V. A FRAMEWORK FOR INTEGRATING PUBLIC OPINION AND COGNITIVE RESEARCH WITH SDL RESEARCH IN DATA RELEASE STRATEGIES

28. In order to address confidentiality/privacy in data release decisions in the 21st century we need to look beyond the technical solutions offered by SDL research. We need to factor in the human side by looking at what the public understands about our processes and what bothers them about how their personal information is being used. An approach that incorporates technical research and behavioral research stands a better chance of avoiding major problems down the road when and if something does go wrong.

29. The suggested process for integrating perception research into the data release decision starts by acknowledging the agreement reached with the respondent when the data were collected. The limits imposed by the messages related to confidentiality and use determine, in part, what data can be released. If the messages are vague and imprecise the respondent may not be reassured that his confidentiality is protected but if messages are very specific the resulting uses may be overly limiting. In addition, research has shown that over-emphasizing the confidentiality and privacy protections to respondents can have the unintended consequence of heightening the respondent's concerns (Singer, Hippler, Schwartz, 1992). The appropriate balance requires knowledge of the specific legal protections and the intended uses as well as research on the cognitive meaning of the words used to convey the protections and uses. The U.S. Census Bureau has recently developed and tested a standardized confidentiality message as part of a

larger effort to improve its survey introductory letters and to address some inconsistencies in the various confidentiality messages that have been in use.

30. Public opinion research has showed us that there are other areas that require some attention. Given that privacy and confidentiality appear to be more closely linked in the public's mind, we need to consider messages that reassure respondents about the legitimacy of the statistical office and the collection effort. Trust in government as an institution that serves an important purpose also plays a role in establishing legitimacy. If the respondent sees the statistical office as just another arm of a government that is not relevant, efforts will be needed to show relevance in terms that the respondent can particularly relate. Also, since we know that technology is a particular concern we will need to emphasize the positive aspects of technology, such as to reduce burden, and the safeguards in place to protect confidentiality. The U.S. Census Bureau has not yet evaluated how best to convey these concepts but I see this as a logical next step.

31. Once we have settled on the "right" messages at the point of collection, we need to turn our attention to the messages that will explain the data we release. (I have previously mentioned the research surrounding the messages used in the AFF.) These are not the same messages as those in the "respondent letter." The data release messages are designed to assure the respondent that the data have been protected and that only the uses that were agreed upon are possible. The messages need to address various possible concerns or perceptions (as I described above) and should convey complicated SDL procedures in a simple and concise manner. They may also help reinforce the legitimacy of the data collection as a way of addressing the privacy concerns. Again, results of public opinion surveys and cognitive interviews will serve to determine what we need to say and how we need to say it.

32. After the data are released, it is recommended that we employ a quality assurance program consisting of user surveys designed to see how well we have done in conveying our message. If there is still confusion or misunderstanding, it is easy enough to modify the message. We also need to continually monitor public opinion to identify changes in privacy attitudes that may necessitate changing our messages.

33. Some may also argue that, in addition to disclosure limitation research and cognitive/opinion research, we need to explore legal/legislative options to impose penalties on those who break the confidentiality protections we apply to the data. Such an approach may allow statistical offices to release more data to the public by reducing the risk that intruders will break the protections. The theory behind this assumption is that if the penalties are steep, existing or reduced SDL protections may be sufficient even with new tools to reidentify individuals. This is a third area that is outside the scope of this paper but needs to be explored as we attempt to shore up all threats to data dissemination.

VI. AN AGENDA FOR CONTINUING RESEARCH

34. What I have attempted to outline here is a rationale for an integrated approach to data release that requires a commitment to research on SDL techniques and research on public perceptions. The results will provide a holistic approach to reassure national statistical offices about their decisions to release data and to reassure respondents that their interests have been looked after. Although opinions will vary across countries and cultures, the methodologies employed will be similar word-wide. National statistical offices, working together, can fund and support continued research that will provide guidance to all.

35. Over the past two years, a proposal has been submitted to Eurostat under the Fifth Framework to undertake research on perceptions of confidentiality. The proposal came out of an international workshop held by the U.S. Census Bureau in May 1999. The proposed research would consist of surveys of the public, businesses, and public opinion leaders. The results would help us understand better how individuals, establishments and those who advocate for the rights of individuals and businesses perceive the threats to confidentiality and privacy from statistical data releases. With this research, the global statistical community would be able to develop an integrated approach to addressing this important component of data dissemination. Unfortunately, this research has not yet been funded but I hold out hope that it will be approved in subsequent rounds of proposals.

36. I truly believe that the statistical community will soon recognize the need for perception research and will collaborate in measuring opinions about privacy and confidentiality. Surveys of public opinion will help us identify and respond to the important concerns both at the time of data collection and dissemination. This research must be ongoing since opinion is continually changing as recent findings have showed. Once we know what to address, we then need to know how to address it. Cognitive research on message content is a critical follow-on to the survey work. As they have done in SDL research, national statistical offices can accomplish much more by integrating their research efforts on perception. It is important that this research begin without delay.

References

Butz, William and Gerald Gates (1986), "Consent, Matching, and Release for Publicly Collected Data," 1986 Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria Va., pp. 16-18.

Capps, Cavan, Ann Green, Mark Wallace (1999), "The Vision of Integrated Access to Statistics: the Data Web." Of Significance: A Topical Journal of the Association of Public Data Users 1:2, p. 42-47.

Fay, R.E., Bates, N., and Moore, J. (1991), "Lower Mail Response in the 1990 Census: A Preliminary Interpretation," Proceedings of the 1991 Annual Research Conference, Census Bureau, Washington DC, pp 3-32.

Gates, Gerald and Robert Groves (1992), "Data Protection: Do Respondents, Data Collectors and Users Agree on its Meaning?" Proceedings of the International Seminar on Statistical Confidentiality, Dublin, Ireland, September 8-10, 1992, pp. 49-59.

Groves, Robert and Robert Cialdini and Mick Couper (1992), "Understanding the Decision to Participate in Surveys," Public Opinion Quarterly, Vol. 56:475-495.

Martin, Elizabeth (2001), "Public Opinion Changes During Two Censuses," paper presented at the Federal Committee on Statistical Methodology's Statistical Policy Conference in Bethesda, Maryland, November 2000, forthcoming.

Mayer, Thomas (1999), "Cognitive/Usability Testing of Confidentiality Statements for American Fact Finder," Statistical Research Division, Bureau of the Census, Washington DC, December 21, 1999.

Singer, Eleanor, Hippler, H., and N. Schwartz, (1993) "The Impact of Privacy and Confidentiality Concerns on Survey Participation." Public Opinion Quarterly, Vol. 4:256-268

Singer, Eleanor and John Van Hoewyk, "Trends in Attidudes toward Privacy and Confidentiality, 1995-2000," forthcoming.