

**Joint ECE/Eurostat Work Session on
Statistical Data Confidentiality**

(Skopje, The former Yugoslav Republic of Macedonia,
14-16 March 2001)

Working Paper No. 28
English only

Topic III: Attitudes of respondents towards statistical confidentiality

BUSINESS PERCEPTIONS OF CONFIDENTIALITY

Invited paper

Submitted by the U.S. Internal Revenue Service and the U.S. Bureau of the Census¹

I. BACKGROUND/INTRODUCTION

1. The core mandate of statistical agencies is to collect data on businesses and households. In so doing, each agency enters into an implicit pact with its respondents that the data will not only be used well, but will be protected from unauthorized access and use. This promise of confidentiality is not only a legal and ethical mandate, but also an important contributor to high quality response rates. However, little is known about the substance of this pact – particularly with respect to businesses. This ignorance can have serious potential consequences. If government misconstrues its role or the nature of the pact, and consequently businesses do not trust government to protect their data because they either mistrust or misunderstand the pact, it will be difficult for government not only to maintain high quality and timely response rates but also to frame new ideas on data collection, protection, and access.
2. This study provides one of the first quantitative analyses of business perceptions of the sensitivity of different types of data and their assessment of the protection provided by different agencies – as well as their assessment of the quality of statistical work performed by the agencies. It also examines the knowledge of the business community of the financial and criminal penalties associated with breaches of confidentiality and its willingness to permit data to be shared among different federal and private agencies.
3. This information can be used not only for the maintenance and improvement of current collection systems but also for framing ideas on new data collection – such as data sharing initiatives - and access systems, like secure data analysis sites. In each of these cases, business concerns are likely to differ from those of households. Businesses, unlike households, have multiple requests from different government agencies for information, so data sharing across agencies may be an attractive way of reducing response burden. Similarly, businesses may be more aware of the importance of data quality for their own research needs, and may favor outside researcher access in controlled sites to further this aim.
4. The study is particularly timely given the heightened interest in confidentiality and privacy issues. Media attention has increasingly focused on lapses in security in the private sector – for example, the private marketing of personal dossiers compiled from consumers' electronic sales records as well as the dissemination of medical records. The federal statistical community should do all that is possible to

¹ Prepared by Nick Greenia, Statistics of Income Division, Internal Revenue Service and J. Brandford Jensen, Center for Economic Studies of the U.S. Bureau of the Census. This study would not have been possible without the extraordinary efforts put into the development of the survey by Diane Willimack and Kristin Stettler, U.S. Census Bureau. The authors would also like to thank Dr. Frederick Knickerbocker for facilitating this work. All views expressed represent those of the authors, and not necessarily the institutions they represent.

convey to the respondent community that they have addressed both real and perceived concerns—even when perceived concerns seem at most to be abstract wisps on the far horizon.

5. The paper is organized into four core sections. The first of these describes and discusses the role of business confidentiality in statistical data collection. This is followed by a description of the questionnaire design and the sample frame. The next section provides the quantitative results, and we conclude with some preliminary suggestions for extensions of this research that can ultimately be used to affect data reporting, collection, protection, and access decisions.

II. ROLE OF BUSINESS CONFIDENTIALITY

6. The lack of quantitative research does not mean that no attention has been paid to the confidentiality² of business data provided for federal statistical purposes. In 1992, the U.S. Office of Management and Budget established a working group that not only noted the differences between household and business perceptions of confidentiality³, but also identified several factors that were likely to affect business trust in the protection afforded their data. These factors are more fully developed in a series of papers by Willimack and coauthors. This paper examines a subset of these – namely, the sensitivity of individual items queried; the perceived benefits of the data collection (e.g., survey objectives); the costs of data collection (e.g., survey completion time); and the protection provided respondents.

A. Sensitivity

7. Knowing what types of data businesses consider to be most sensitive might be used by statistical agencies to accord different levels of protection and permit broader access to, and analysis of, subsets of data. It is well known that different types of household data have different levels of sensitivity – item response levels on income measures vary substantially from those on age and number of children. Although no hard evidence exists, the sensitivity of business data is likely to be different for a number of reasons: the existence of publicly available information; the structure of business entities and the existence of competitors.

8. Businesses, unlike households, are often routinely required to provide information for administrative or regulatory purposes which quickly becomes publicly available (see Willimack et al, 1999). For example, publicly traded companies are required annually to provide to the SEC extensive financial information, much of which also ends up on commercially available datasets such as Compustat. Further, all employers sponsoring an employee benefit plan (retirement, health, etc.) are required to file an annual Form 5500 series return so that IRS, DOL, PBGC, and SSA can administer their respective provisions of the 1974 Employee Retirement Income Security Act. Private corporations, such as ABI/INFORM, use information in the yellow pages to create business lists with name, address, employment, payroll, and industry information on the universe of businesses. In sum, virtually all of the data provided for these purposes are publicly available, raising the question of **whether** confidential datasets containing subsets of the provided information really need to—or even can—protect all of their information equally. It may well be that the overriding issue to businesses in such cases is to avoid further reporting burden, rather than with how to obtain maximal “confidentiality” protections for data already in the public domain.

² The protection of confidentiality in this chapter is defined as the restriction of access to information about the individual party/entity once it has been provided—for statistical or administrative purposes—to a second party charged with the collection responsibility. The confidentiality protection responsibility is traditionally viewed as residing with the collecting party, even when the law permits third and fourth parties to access the data. Indeed, the consequences of any breach of confidentiality would almost always be borne by the collecting party in the form of reduced response rates and less precise responses as the cost exacted for such violations

³ These are similar to the differences between collecting household and business data generally (see, eg, Box and Chiannapa, 1995).

9. Other items, not in the public domain, may not be sensitive simply because of the structure of business entities. For example, the taxpayer identification number, which is quite sensitive for individuals (the Social Security Number), may not be sensitive for firms, since the Employer Identification Number (assigned by the IRS) often appears on publicly available datasets and also may be changed several times over the business' lifetime. This difference makes businesses more elusive to track and monitor over time than individuals and hence does not enable instant access to complete lifetime data.

10. The inherently competitive nature of business is also something to be considered in analyzing the sensitivity of data. Many businesses may consider some information, such as name and mailing address, less sensitive than individuals and households, but other items necessary for profitable strategic planning, such as sales at the establishment level or trade secrets, very sensitive indeed. Similarly, the time sensitivity may well be different: in the rapidly changing business world, data more than one or five years old may need much less protection than current data.

B. Benefits

11. One concern raised by the working group, and supported by the research of Willimack et al (2000), is that the direct benefits to businesses of the data that are collected from them by government statistical agencies may not always be readily apparent to them, and indeed, the data themselves may permit strong competition, even when privately collected. In fact, the major producers of statistical data on businesses, Census, BLS and BEA have as their primary mandate to produce data on the economy for government policymakers (e.g., the Congress and the Administration), not for businesses. Although the Census Bureau notes that economic census data can be used by businesses to, among other things, study their industry, gauge competition, calculate market share and study business markets, this purpose can often be satisfied by using more detailed microdata from the private sector. A brief perusal of private company websites such as Dunn and Bradstreet, American Business Information, and the Donnelly Information Files provides convincing evidence that businesses can analyze quite detailed and quite current firm level information on competitors⁴ rather than relying on aggregate, federal government data.

C. Costs

12. The burden imposed on businesses of filling out surveys and censuses is a clear concern of OMB, but the full cost estimate is not well known. The Census Bureau has estimated that the cost to business of filling out surveys in non economic census years is about 2 million hours; economic census years add an extra 5 million hour burden. There is some evidence that businesses are unhappy about filling out the same information for different agencies (Nichols et al., 1999) due to the redundant burden and because they do not understand the need. Voluntary mail survey response rates on businesses range from xx to yy%

13. Our own anecdotal experience supports this assertion. We had originally intended to explore the topic of confidentiality with the CIO's of the 200 largest US corporations, but the initial set of 15 phone calls revealed not one company willing to participate in a voluntary survey. In fact, almost all of these calls were unable to make it past the initial company screener.

D. Knowledge of protection provided respondents

14. Although statistical institutes make much of the legal protections that are afforded by statute, it is not known whether businesses know about or value these protections. This lack of knowledge may be due to the decentralized nature of the US statistical system, since different data collection instruments utilize both different privacy and confidentiality protection statements/pledges⁵.

⁴ See, for example, <http://www.mscnet.com/prodserv/nationaldatabases/index.htm>

⁵ See Annex 1 for the privacy and confidentiality pledges published by the federal statistical agencies, Census and BLS, as well as the federal income tax agency, IRS. Penalties for confidentiality violations (also known as unauthorized disclosures) are provided for the same three agencies in Annex 2.

15. This situation raises several questions. Do respondents understand what is meant by confidentiality protection? Do they realize and understand the variation which exists across different collection agencies? An additional implicit question is how and even whether the system should address the authorized secondary disclosures which occur among the collection agencies⁶. How much “comprehensive” information should be provided to respondents to bolster their informed consent status, without adversely affecting response rates at the same time? Should the same data collected from different sources (administrative or survey) be treated differently from a confidentiality standpoint?

16. At least part of any current answer would seem to depend upon the absence of any data about relevant perceptions of the respondent community itself, which certainly seems a reasonable point of departure in any discussions on this subject. However, analysis by Singer et al. (1997) suggests that the volume of information can itself adversely influence response rates. Obviously, such a result would seem to be in the ultimate interest of neither respondents nor policy makers, so some optimal combination of information and brevity (the right amount of the right sort of information) would appear to be the objective.

III. QUESTIONNAIRE DESIGN AND DATA

17. Although this brief discussion raises many questions, it is clear that there is no quantitative base on which to answer them. We developed a questionnaire to begin to answer a subset of these questions, while recognizing that this represents only a first step towards the development of a much broader research agenda. We administered the survey through the mail using a commercially available business database (Dunn and Bradstreet) to designate the respondent population.

A. Questionnaire design

18. The questionnaire itself was designed to inform two discrete components of the confidentiality knowledge base:

- i) What kinds of data/information do businesses consider sensitive – and for how long are they perceived to be sensitive?
- ii) What are businesses’ perceptions of collection agencies’ ability to collect and protect data efficiently? In particular, do businesses believe there are differences in the quality of data collection across both private and public agencies, and are they aware of (and do they have confidence in the efficacy of) penalties for disclosure violations?

19. These are complex concepts – particularly given that most businesses may not think much about confidentiality issues (Willamack et al., 2000). Because education/information is in itself known to influence/change perceptions, we decided to define confidentiality only minimally in this first pass at data collection⁷.

20. The first set of questions deals directly with the question of whether businesses believe that different types of data have different levels of sensitivity – despite the fact that statistical agencies treat all these data with virtually the same level of protection. For example, we ask whether the business considers its primary identifiers--name, address, and phone number--to be sensitive, as well as whether employment, payroll, sales, profits or tax liabilities are sensitive items. We expect, *a priori*, the former to be not as sensitive, since they are typically available in the phone book and are even advertised to

⁶ For example, universe extracts of records with limited item content are provided annually by IRS to Census primarily to reduce both respondent burden and collection costs, but with the secondary usage of providing an additional input source against which to check and verify some survey and even census information. While such redisclosures are statutorily authorized (USC Title 26 section 6103 j1A, Title 13, etc.) such arrangements are [not] mentioned [at all?] only briefly in these actual collection instruments used by these agencies

⁷ Nevertheless, the questionnaire does use the item responses themselves to begin to establish differences in perceived definitions. For example, item 7 seeks answers to core questions which frame business belief systems regarding confidentiality, including whether statistical collection agencies— release identifiable data to anyone, keep collected business data confidential, release any collected data outside government, and share collected data with other agencies.

accomplish basic profit objectives, but the latter set of questions to be increasingly sensitive. We also ask whether similar data are more sensitive at the establishment or company level.

21. Building on this foundations, we then attempt to probe whether businesses feel that there is a time dimension to the sensitivity of their data (again – statistical agencies typically treat business data as sensitive regardless of the age of the data). We differentiate again between types of data and length of time (1, 5, 10 and 30 years).
22. The next set of questions asks whether businesses are more or less concerned about the datacollecting agent/recipient – whether they distinguish among federal regulatory or statistical agencies, not-for-profit researchers, for profit researchers, other businesses or the general public.
23. We then attempt to capture how businesses feel about the performance of federal statistical agencies (and will later correlate this with their other responses). In particular, we ask whether the federal government is better than the private sector at collecting information, providing information, and protecting information. This is followed by a more detailed set of questions to find out how businesses perceive the protection provided by federal statistical agencies – whether their data are kept confidential, whether their data are disclosed to, or shared with other agencies. We also test our idea of absolute respondent cynicism regarding government trust by asking whether the respondent believes that any federal agency, including the IRS, can access data provided by businesses any time it wants. The analytical section will examine the interactions across the responses to these different items.
24. An important - hitherto unknown – element is whether businesses are aware of the legal and financial penalties imposed by federal agencies for individuals who divulge confidential information without authorization. Question 8 addresses this contingency, and we follow it with a question which asks whether these penalties should be increased, or more stringently enforced, and if either, whether the business would be more inclined to provide confidential data as a result .
25. The last set of questions generally investigates whether the respondent might view responsible datasharing as a solution to some of their concerns on burden and asks for their input into other concerns about the use of their data by federal statistical agencies.

B. Data base

26. The sample frame for the survey was derived from Dunn and Bradstreet’s commercial database which represents 11,300,000 business establishments in the U.S.⁸. This database has several advantages. It provided us with information on the name and title of up to four levels of management (e.g. the Owner, Chief Executive Officer, Chief Information Officer and Chief Financial Officer), information on industry, employment size and single/multi unit status, as well as being quite current (updated on a daily basis). Dunn and Bradstreet implement a number of checks to improve the quality of the data (see Annex 2 for a more detailed discussion). The most obvious disadvantage is that the sample frame is not necessarily representative of businesses in the U.S.

D&B FREQUENCIES BY INDUSTRY DIVISION	
Industry	Number of firms
Agriculture, Mining, Contract Construction	454
Manufacturing	991
Transportation, Public Utilities, Communications	270
Wholesale and Retail Trade	1,028
Finance Insurance, and Real Estate	454
Services	927
Public Administration	875
Total	5,000

⁸ This is derived from the Dunn and Bradstreet report, which is at variance with the size of the Census Bureau’s Business Register.

27. Two types of pilots were conducted. The first was a mailout of 25 questionnaires to a random subset of the population. The second was a set of cognitive interviews conducted by Kristin Stettler of the U.S. Census Bureau on a subset of 8 out of 25 respondents who agreed to be interviewed about their understanding of the questions.

28. Standard Dillman methodology was employed for a paper mailing by First Class Postal service, with the first wave of questionnaires (Annex I) transmitted with a cover letter on 28 November 2000. A follow-up reminder postcard was sent to non-respondents on 8 December, and a second mailing of the questionnaire and cover letter was transmitted on 28 December. The survey was sent to a random sample of 5,000 businesses, 1250 from each of four employment size classes: 0-49, 50-249, 250-499 and 500 or more. Of the 5,000 total businesses, 2,530 were multi-unit businesses with headquarters locations, and 2,470 were single unit companies with only one business location. Some 400 surveys were returned as undeliverable [address might have been correct but business might have been defunct] , and some 400 completed surveys were received as of 11 January 2001.

IV. SURVEY RESULTS [THE DATA] [NOTE: This section is incomplete as the survey is still in progress.] Number of responses/non-responses/non-deliverables for each mailout, and in total. Individual Item responses (How code multi or non-responses for one item?)

29. All Classified by:
- i) Industry (what level of precision?)
 - ii) Employment size (the 4 classes only or others, too?)
 - iii) Single vs. Multi unit?
 - iv) Geo (how broadly?)
 - v) Some financial info size
 - vi) Age?

V. CONSEQUENCES FOR THE FUTURE [FUTURE PLANS] [NOTE: To be completed after survey results analysed.]

29. Discussions for the development of a list frame for controlled access, but minimal restrictions on bona fide researchers. Assuming Name, Address, Telephone Number? Employment size? Industry Code? Considered non-sensitive, TIN?

30. Development of a continuing dialogue
Respondents <-> Data Collectors <-> Researchers/Analysts <-> Policy/Decision Makers

- a) Perception studies to monitor response climate;
- routinize so that surveys/censuses other data gathering constructs (including admin data uses, time series linkages, etc.) not undertaken without firm perceptions underpinnings, foundation;
 - use perception study results to justify/explain & to demonstrate good faith efforts on part of stat community;
 - New collection systems: when possible, avoid statutory changes due to unwanted consequences of legislation.

b). The implications for Administrative Data Use;

Perceptions<->Response Rates<->New Collection Systems <-> New Storage/Access Systems

c) The implications for data provision;

d) The implications for accountability.

References

- Biemer, P.B. and Fecso, R.S. (1995), "Evaluating and Controlling Measurement Error in Business Surveys." Chapter 15 in B. Cox, et al. (eds.), *Business Survey Methods*. New York: Wiley.
- Christianson, A. and Tortora, R. (1995), "Issues in Surveying Businesses: An International Survey." Chapter 14 in B. Cox, et al. (eds.), *Business Survey Methods*. New York: Wiley.
- Cox, B.G. and Chinnappa, B. N. (1995). "Unique Features of Business Surveys." Ch. 1 in B. Cox et al. *Business Survey Methods*. New York, Wiley.
- Dillman, D. A. 1978. *Mail and telephone surveys: The total design method*. New York: John Wiley & Sons.
- Edwards, W.S. and Cantor, D. (1991), "Toward a Response Model in Establishment Surveys." Chapter 12 in Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A. and Sudman, S. (eds.), *Measurement Errors in Surveys*. New York: Wiley.
- Mackie, C and N. Bradburns (eds) 2000 *Improving Access to and Confidentiality of Research Data: Report of a Workshop* National Research Council.
- Nichols, E., Willimack, D. K. and Sudman, S. 1999 "Balancing Confidentiality and Burden Concerns in Censuses and Surveys of Large Businesses". Paper presented to the Washington Statistical Society, U.S. Bureau of the Census, Washington, D.C., September.
- Singer, Eleanor, S. Presser and J. Vanhoewyk 1999 "Public Attitudes Toward Data Sharing By Federal Agencies", in *Record Linkage Techniques - 1997 Proceedings of an International Workshop and Exposition* National Academy Press.
- Wallman, K and J. Coffey 1999 "Sharing Statistical Information for Statistical Purposes" in *Record Linkage Techniques - 1997 Proceedings of an International Workshop and Exposition* National Academy Press.
- Willimack, Diane K., Elizabeth Nichols and Seymour Sudman (forthcoming) "Understanding Unit and Item Nonresponse In Business Surveys," in *Survey Nonresponse*, Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J.A. Little (editors), New York: John Wiley and Sons.
- Willimack, Diane K, Seymour Sudman, Elizabeth Nichols and Thomas Mesenbourg 1999 "Cognitive Research on Large Company Reporting Practices: Preliminary Findings and implication for Data Collectors and Users" mimeo, the Census Bureau.