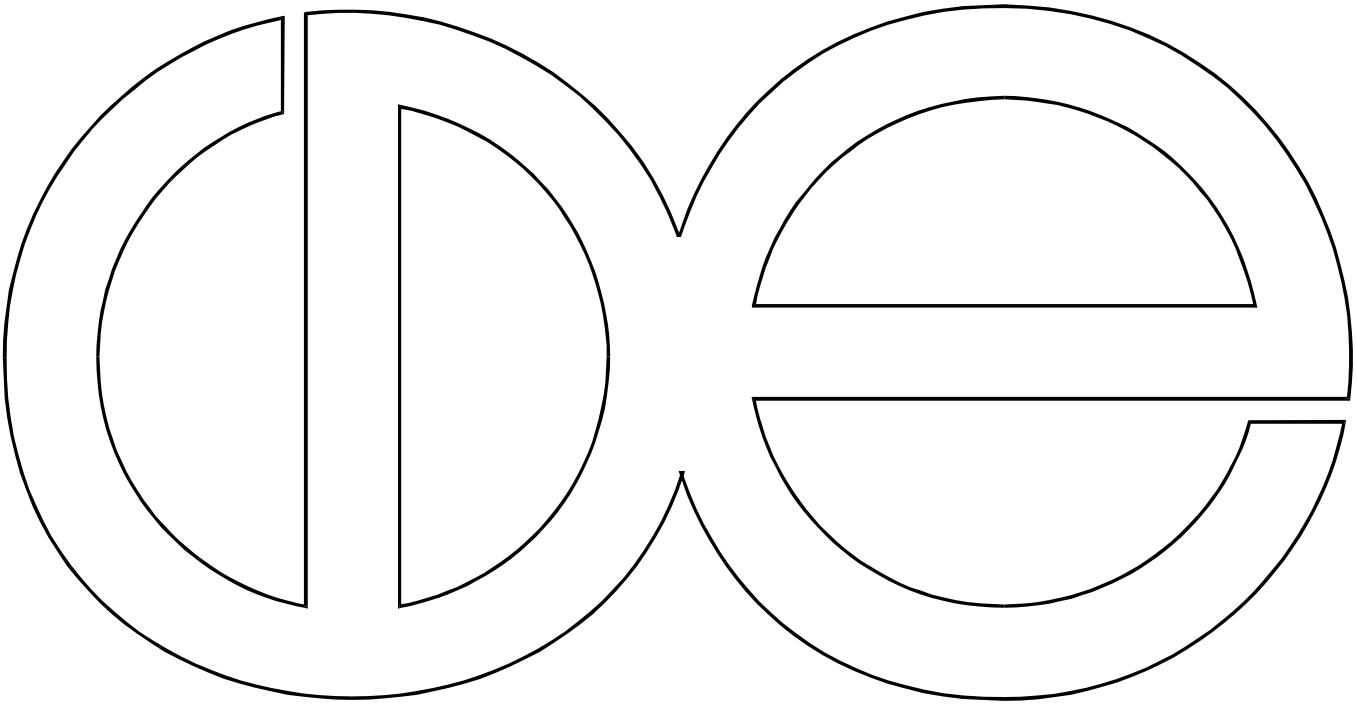
Center for Demography and Ecology

University of Wisconsin-Madison

Proactive Social Science Defends Confidentiality and Data Access

Martin H. David

CDE Working Paper No. 2001-03



PROACTIVE SOCIAL SCIENCE DEFENDS CONFIDENTIALITY AND DATA ACCESS

Martin H. David
Joint Program in Survey Methodology
University of Maryland – College Park

Correspondence: 425 8th St N.W. #1144, Washington DC 20004
E-mail: Mdavid@survey.umd.edu

Presented at Population Association of America
March 30, 2001
Revised 30 April 2001

Acknowledgments

The Science, Technology, and Law Program of the National Research Council held a workshop on the impacts of the Shelby amendment on the conduct of science on 12 March 2001. A report on that meeting is in preparation.

The American Statistical Association Committee on Privacy and Confidentiality was responsible for developing an organizational response to OMB rules for the Shelby FOIA process in 1999 and again in 2000. I had the privilege of taking the lead in drafting the response. Members of the committee and the president of ASA engaged in thoughtful and constructive development of the draft that made me proud to be a member of ASA.

Proactive Social Science Defends Confidentiality and Data Access

1 A generation ago

1.1 Confidentiality and access to data

In the late 1950's and early 1960's survey research boomed under the leadership of academic research centers that innovated in demographic studies (Freedman's *Growth of American Families*, Miller and Campbell's studies of the electoral process, NORC's *General Social Survey*, and many others). A norm of data-sharing was born for these major studies, and is currently alive and well. The data are documented, archived, and widely-used in secondary analyses. The Bureau of the Census broke new ground by creating a public use file of the *1960 Census of Population* that again fostered extensive secondary analysis of individuals and households (Converse 1987).

Two principles characterized shared use of survey data. The shared data were anonymized. Use of the data was to create scientific knowledge. Leaders in the survey community foresaw no negative consequences of widespread data-sharing for the process of gathering responses. To the contrary data-sharing was a means of producing scientific findings more rapidly and validating the expense of large probability samples to the funding agencies (many private, some Federal government).

1.2 Data banks, privacy, and legislative initiative

A storm of anxiety over Orwellian power attributed to data banks and electronic computation broke over the U. S. in the early 1970's. Fear that administrative records were widely and inappropriately shared within government lead to two major pieces of legislation: The *Privacy Act of 1974* and The *Tax Reform Act of 1976.* The former affirmed individuals' constitutional rights to privacy and implemented fair information practices in Federal record systems. The *Tax Act* provided parallel, but not identical, protections for tax returns.[1]

That legislation has framed discussion of confidentiality and access to data for research for a generation (Duncan-Citro 1993). Evidence in the 1980's that individuals could be re-identified from anonymized public use data led to technical procedures for limiting the probability of disclosure. But the need to study identifiable data led to academic research through other modes. Access to identifiable data was sparingly granted through special sworn employee status. In other instances researchers were licensed to use confidential data. Licenses required users to observe restrictions that assured scientific use and the maintenance of confidentiality promised to the data suppliers.

2 Framing confidentiality and data access since 1995

In the last six years, new storms have buffeted the American population. Most of the population now knows that it faces substantial risks of loss of privacy that are not wanted and are, in part, illegal. Telemarketing invades the home. Viruses, cookies, and bugs afflict our personal

computers. Our medical information has been accessed by persons with no need.

At the same time, the U. S. citizen is reduced to a number, the Social Security Number (SSN). No transaction with a financial institution can be completed without the SSN. Many people understand that 1099 reports increase the efficacy of income taxation. Our national resistance to registration of the population has been totally overcome: IRS mandates assignment of SSN to infants.

These storms confound the traditional mechanisms for protecting the confidentiality of respondents to surveys and maintaining widespread sharing of survey data for scientific research. No single piece of recent legislation does more to change the course of the last forty years of scientific analysis of survey data than the Shelby amendment of 1998.

3 The Shelby amendment, 1998-

3.1 Pertinent history

The Omnibus Appropriations Act (Public Law 105-277), signed into law towards the end of 1998, contained a two-line amendment introduced by Senator Richard Shelby (Alabama). The amendment instructs OMB to amend Section __.36, of Circular A-110.[2] The substance of the amendment reads:

> "[OMB Circular A-110 is] to require Federal awarding agencies to ensure that all data produced under an award will be made available to the public through the procedures established under the Freedom of Information Act." (FOIA)

Literal reading of these lines suggests that any data collected by any non-profit organization under a Federal grant can be commandeered by any private person (including corporate persons through a FOIA request. (The Federal award may be a trivial part of the financing for the research, but it contaminates the entire investigation stemming from related data collection with the burden of FOIA release.)

The Office of Management and Budget drafted rules, effective November 1999, to interpret the new law. OMB restricted the target of the amendment to *research data* that support *published* findings and are *used by Federal agencies to develop administrative rules*. The rule declares:

> "(i) *Research data* is defined as the recorded factual information commonly accepted in the scientific community as necessary to validate researching findings. ...
>
> (ii) *Published* is defined as either when: (A) research findings are published in a peer-reviewed scientific or technical journal; or (B) a Federal agency publicly and officially cites to research findings in support of a regulation. ...
>
> (iii) *Used by the Federal government in developing a regulation* is defined as when an agency publically and officially cites to research findings in support of a regulation for which

notice and comment is required under 5 USC 553."

Immediate, and highly critical comment on the rule, from Shelby and others, suggested that the OMB had overreached in narrowing the scope of the law and interpreting the relationship of data to administrative rule-making.

An irritant that may have provoked Shelby's amendment was the *Six Cities* study. From 1973 to the present Professors Ben Ferris and Douglas Dockery at the Harvard School of Public Health have collected epidemiological and health data from individuals in six cities with large differences in ambient air quality. Their careful longitudinal measurement of health risks, disease, and mortality for their sample population provided the basis for inferring that differences in the ambient particulates of less than 10 millimeters in size and ozone concentrations across the six cities contributed to a decline of 2-3 years in life expectancy, conditional on age and other health risk factors.

3.2 EPA rule-making and *Six Cities*

Six Cities research was relevant to EPA's mission. EPA financed a part of the *Six Cities* study with grants. Shelby's amendment would apply to *Six Cities* under the OMB rules, except that the rule pertains prospectively after November 1999. Retroactive application to earlier grant awards would upset a timing principle that is deeply embedded in legislative rule-making.

The *Six Cities* published findings were cited in support of a 1996 EPA rule that sets an 8-hour ozone ambient air quality standard and an ambient air quality standard for fine particulate matter. During the public comment period on the EPA rule numerous affected parties requested access to the *Six Cities* research data.[3] EPA demurred saying that it did not hold the data and could not compel the investigators to release the data.

The EPA rules were hotly contested prior to 1996. Large costs were attributed to complying with the new EPA rules. The rules dramatically increased scope of non-attainment areas. And the rules apply to sources of pollution distant from non-attainment areas. Each of these aspects of the rule created discomfort among industry and energy users. Electric power generating facilities, automobile and truck manufacturers, the trucking industry, and construction were particularly aroused and sought to vacate the regulation.

By 1998 the EPA rule was under review in U.S. Supreme Court. Early in 2001, the Court determined that the 8-hour ozone standard was promulgated without exceeding the authority of the EPA. We now begin a period of transition to the new standards that was averted by the intervening litigation.

3.3 Restricted access to *Six Cities* data

Meanwhile, what became of *Six Cities*? Between 1996 and 2000 the *Six Cities* Harvard group arranged for the Health Research Institute (HRI) to audit its methodology, survey materials, and analyses. The HRI employees became "special sworn employees" of the Harvard project, so that

the audit could be carried on without violating confidentiality pledges. Confidentiality pledges had been made to three groups:

- Individuals who contributed their responses to surveys and their bodies for diagnostic testing of lung and related functions;
- National Center for Health Statistics whose *National Death Index* was used to establish mortality in the study population; and
- Several states who hold death certificate information (cause of death) as confidential.

The audit, the choice of the auditor, and the procedure for the audit constitute a most unusual scientific undertaking. It was expensive. Several advisory groups provided direction and vetted reports. Parties litigating or opposed to the EPA rule probably found the mechanism for restricted access to data unnecessarily awkward and uncontrollable. Whatever view one has of openness for rule-making, it seems unlikely that every request for data used in rule-making can be handled with this model. I hypothesize that disaffection with the HRI mode of operation stimulated Senator Shelby to introduce his amendment.

3.4 Where are we now?

OMB rules enforce the use of FOIA to obtain research data *regardless of pledges of confidentiality to respondents*. Data access is the privilege of anyone willing to pay the costs of reproduction of data, applicable metadata, and related efforts of scientists that are required to make the data usable by secondary analysts. The institutions receiving grant awards are beholden to provide data access for each project award that they receive; principal investigators are legally required to comply with A110.

The Shelby clause does not restrict the recipient of data to research use. FOIA applications can be made by journalists or marketing companies who hope to find a gold mine in lists of individuals that emanate from samples of specialized populations. Lawyers representing parties in any adversarial proceeding where the data might be relevant may use the data for discovery as well as for assisting or harming particular individuals (including corporate persons).

The U. S. Chamber of Commerce has declared that it will litigate the circumscribed force of the OMB rule. Richard A. Merrill, Professor of Law, University of Virginia, a member of the Science, Law, and Technology Program of the National Academies, commented on this declaration in two ways (March 12, 2001):

> • Some researchers foresee the value of their findings to Federal rule-making. The Six *Cities project* understood that their findings would be contested in an adversarial atmosphere. Many principal investigators will not anticipate the link of their findings to rule-making.

Secondly, Merrill asks the research community to understand the point of view that energizes the proponents of a broad application of the Shelby rule. He observed:

> • In the 1950's few regulations impeded the conduct of business that might engender health,

> safety, and environmental risks for third parties. Torts constituted the remedy for those damaged, and every business had access to tort-based judicial determination of how precedent, common law, and particular facts might affect their financial liabilities. OSHA, EPA, and other agencies have created a web of regulatory obligations since the 1950's. Compliance with regulations is rarely adjudicated. Instead, businesses are forced to deal with an administrative hierarchy to obtain relief from what appears to be unreasonable intrusion in business operations.

These comments suggest two consequences, the first pointed out by Merrill in his remarks. Investigators who do not anticipate FOIA requests may find themselves blind-sided by legal expense, disruptive interventions in their research activities, and uncomfortable allegations about the propriety of their behavior. On a larger scale academic institutions need to act to minimize the consequences of litigation that uses FOIA as javelin to pierce established standards of conduct for the collection of survey data.

I do not doubt that FOIA will be used to pry data from investigators. And will future investigators have the power to negotiate restricted access (that the *Six Cities* study achieved)? Power to negotiate restricted access will require each research institution to endorse the use of confidentiality pledges in social surveys.

Mindful of these observations, the social science community must assume that the force of FOIA requests will impact their research. How are we to respond? What kind of precautions can be undertaken to protect the confidential relationships that we have with respondents and the assurances that we give agencies who release data for our use on a restricted basis?

First, we need to be clear that conflict created by Shelby's rule is inherent in two widely supported ethical imperatives:

> • Government should be open. The basis for its law and rules should be clear to the entire polity.
>
> • The contract between data givers and the research community must be confidential to assure that voluntary response will be complete and truthful.

Where these two principles apply at the same time, dispute will be generated by individuals who weigh the imperative to openness more heavily than respondent privacy. Judicial adjudication is inevitable.

4 The social science response

4.1 What does science require?

First, both ethical imperatives relate to social science. Social scientists can research attitudes towards openness in rule-making and confidentiality of responses to surveys. Furthermore, social science research has the most to lose if survey methodology is impaired. Other fields that

might defend promises of confidentiality to respondents are not required to deliver data through FOIA. Most importantly, commercial testing, including clinical trials that pledge confidentiality, is excluded from the Shelby rule. (Clinical trials undertaken in non-profit research organizations and financed by grants from Federal agencies will be subject to FOIA requests.) In addition, the OMB rule excludes physical data from the FOIA requirement. Specimens and proprietary secrets need not be revealed.

Second, since the conduct of social science research is impaired by Shelby, the social science community will need to justify the use of confidentiality pledges, and the conduct of survey research to obtain a different mechanism for revealing data than FOIA. As we have seen in the case of *Six Cities,* restricted access to data can satisfy public need to verify scientific findings without changing the relationship between data givers and scientific data collectors. Each of us will need to contribute to and support justification of confidentiality pledges in survey research. Further, we will need to take special care to maintain an audit trail of the handling of all data collected under confidentiality pledges, as well as a complete record of judgements that went into estimation from the data.[4]

Few in the general public have an understanding of scientific enterprise, particularly when that enterprise entails collection of data through surveys. The public is asked on a regular basis to supply information (useful to this cause, or that product registration, or the needs of some membership association). The public can not easily discriminate data collections that propagate investigations of deep scientific value from the chaff of information gathering.[5]

Reputations of institutions and individuals in the community serve as proxies for the value of data collection to scientific research. That is why many data collectors obtain the support of community leaders for their data collection projects and appeal to public loyalty to some locally respected institution (e.g., of higher education) in requesting cooperation with the data collection.

This appears to be a tenuous fabric to protect the vitality of survey data collection. How can we strengthen the fabric? What will weaken it? The remaining remarks focus on these questions.

4.2 Professional conduct and governance

We can be sure that the institutions we serve are weakened by every instance of scientific misconduct that is alleged. With Shelby, scientific misconduct cuts academic scientific credibility in two ways: We can be berated for failing to uphold pledges of confidentiality. We can be excoriated for failing to yield data collected with confidentiality protection to those requesting FOIA access.

Our response must be to burnish the reputation of scientific independence and reliability of our investigations through continuing extension activity with a large number of publics.

Academic scientists need to publicize standards that the survey community has adopted:

- The public should learn about the Association of Public Opinion Research standards for

survey research.

• The public should have the opportunity to review the excellent pamphlets prepared by the American Statistical Association on surveys and on confidentiality.

• The public should receive a continuing stream of translations of the value of scientific findings to their lives.

I believe that principal investigators who neglect these forms of outreach imperil the rest of the profession.

Secondly, academic researchers need to make the administrators of their research aware of the need to defend confidentiality of research data in court. Faculty governance in research institutions must convince its administrative hierarchy that more damage to scientific investigation will follow from complying with Shelby than from compelling its research scientists to make FOIA disclosures. We have already seen, in a case that was cited by the AAU in its comments on Shelby, that an investigator whose institution did not support confidentiality pledges resigned when commanded to disclose data collected under a pledge of confidentiality. It will be tragic to sacrifice other careers in this way.

Think of all the data that demographers collect under pledges of confidentiality. Will surveys of contraceptive use, of sexual behavior, of cohabitation, and out-of-wedlock births elicit response, let alone truthful response, without a pledge of confidentiality? I can vouch from experience with surveys of income and wealth, that response rates will decline if confidentiality is not offered.

I conclude that survey takers, and academics accustomed to secondary analysis of surveys for their creative research findings, will need to mount a strong campaign to support continued use of confidentiality pledges.

4.3 Educating the next generation

The last, and most important, defense against the application of FOIA to data collected under confidentiality pledges lies in education. *USA Today* has demonstrated that Americans are statistical "junkies". We need to train everyone who takes a single course in social science that a major source of the statistical information that we have about our society comes from surveys taken with pledges of confidentiality to the data suppliers. What is the practice of the Federal government in this area? What do the major not-for-profit survey-taking organizations do in their data collection? If we degrade the quality of surveys, because confidentiality is not offered to the respondent, in what ways is the statistical picture of us distorted? What information will we be able to collect from illegal immigrants? From persons using drugs?

5 Outcomes

The prognosis I give you is not pretty. Either survey research will be degraded to accommodate Shelby or social scientists must mount political action to maintain the quality of their research methods. These alternatives are not overdrawn. It is certain that some researcher's life work will

be crucified by FOIA in the next decade, unless social scientists choose to oppose the unbalanced and unnecessarily destructive means to openness in rule-making that is now law.

References

Converse, Jean M. *Survey research in the United States: roots and emergence*. Berkeley CA, University of California Press, 1987.

Duncan, George and Constance Citro, editors. *Private lives and public policies: confidentiality and accessibility of government statistics*. Washington, D.C., National Academy Press, 1993

Fienberg, Stephen E. and Margaret E. Martin. *Sharing research data*. Washington, D.C., National Academy Press, 1985.

Endnotes

1. The Executive Office of the President obtained copies of individual income tax returns and used information on those returns to harass political opponents in the 1970's. Congress deemed it appropriate to create special legislation governing access to tax returns IRS. Access to returns outside the IRS was denied except for the purpose of estimating revenues and related policy initiatives, where returns could be released to the Joint Committee on Taxation staff and staff of the Office of Tax Analysis, US Treasury.

2. OMB Circular A-110 is titled “Uniform Administrative Requirements for Grants and Other Agreements with Institutions of Higher Education, Hospitals, and Other Nonprofit Organizations”. OMB circulars are standing rules of two or more years duration providing information or instructions to Federal agencies, in this case the appropriate terms for administering grants to non-profit organizations.

3. Persons requesting data presumably wished to attack the statistical model used by *Six Cities* or believed they could find questionable procedure and judgement used in the data collection process.

4. Some pledges of confidentiality can be protected by destruction of identifying information after individual response has been reconciled with the sample selected for study. However, panel data require maintaining identifying information. Contact addresses and phone numbers, as well as links that relate individuals to changes in name and establish alternate contacts through close relatives or friends are essential for following the sample members. When the location of data collection is not concealed, as was the case for *Six Cities*, the risk of reidentifying some individuals in anonymized data is substantial.

5. Steven Fienberg comments that this lack of understanding emanates from public confusion of privacy with confidentiality.

Center for Demography and Ecology
University of Wisconsin
1180 Observatory Drive Rm. 4412
Madison, WI 53706-1393
U.S.A.
608/262-2182
FAX 608/262-8400
comments to: dkuo@ssc.wisc.edu
requests to: cdepubs@ssc.wisc.edu