CONTROLLED ACCESS AS A MEANS OF BALANCING RESPONDENT PRIVACY AND ANALYTICAL UTILITY IN SURVEY RESEARCH

BY

Phillip A. Windell
Bonneville Power Administration
U.S. Department of Energy

This paper discusses issues related to the privacy and confidentiality of survey data and describes procedures for protecting the privacy and confidentiality of survey respondents. Of primary concern are the impacts of these procedures on the research value of the data. A method for striking a balance between these two competing interests is presented together with a case study in its implementation at the Bonneville Power Administration.

THE CHALLENGE TO INDIVIDUAL PRIVACY

Sensitivity to issues involving individual privacy lies at the root of our nation's political and legal systems. The advent of very large, extremely fast electronic data processing machines presents a challenge of unprecedented magnitude, because these machines have made it possible for governments to assemble and readily access vast quantities of information concerning individuals. Lest the likelihood of such occurances be too readily dismissed, it should be recalled that during World War II, the Department of War and the Department of State inquired about access to individually identifiable records for the U.S. Bureau of the Census in an effort to identify Americans of Japanese ancestory living on the west coast.(1) More recently, in the state of New Jersey, law enforcement officials requested individually identifiable information concerning participants in the New Jersey Negative Income Tax Experiment.(2) These are but two of many incidents that could be cited as evidence of the need for limitations on and close scrutiny of governmental uses of information pertaining to individuals.

The U.S. government is not the only beneficiary and potential abuser of data pertaining to individuals. Advertising and door-to-door sales companies, debt collection agencies, and even electric utility companies regularly use government collected data which derive from individuals, could benefit greatly from access to government data which permit identification of individual sources, and therefore represent a potential threat to individual privacy.

The Federal Government has sought to limit the increasing threat to individual privacy through appropriate legislation, including the Privacy Act of 1974 (Pub. L. 93-579). In response to this and other legislation, federal agencies have regularly attempted to avoid invasions of individual privacy from data collections through data reporting techniques that make it impossible to identify individual respondents. For example, the U.S. Bureau of the Census does not report data for individual respondents. Furthermore, where the characteristics reported for geographic clusters would enable identification of individuals, the data for that cluster are suppressed. And, indeed, the U.S. Bureau of the

Census has been exemplary in maintaining the highest ethical standards in the conduct of surveys and in protecting the privacy of survey respondents.

The Energy Information Administration, U.S. Department of Energy (EIA/DOE) relies on a different technique to protect the privacy of respondents in their energy consumption sample surveys. Among other objectives, these annual surveys are designed to produce a data base for use by analysts in accounting for variations and changes in energy consumption among individual residential units.(2) As a result, data for individual respondents (residential units) are available on computer tapes. In an effort to protect the privacy of the respondents, the EIA/DOE suppresses certain information (for example, geographic location other than Census Region and climate zone). In addition to the interview responses, the data from the EIA/DOE surveys also include actual billing histories for the primary fuels used by the unit (electricity, natural gas, and fuel oil), including delivery or billing period dates and amounts of fuels consumed. Since the fuel suppliers could use this information to identify their own customers, the EIA/DOE "masks" this data by systematically altering the dates and fuel amounts. Thus, the billing dates are randomly altered by a factor of $\pm$ 3 days and the fuel amounts are randomly altered by a factor of $\pm$ 4 percent.

THE IMPACTS OF PRIVACY PROTECTION ON RESEARCH UTILITY

When properly implemented, both data suppression and data masking techniques are effective methods of protecting the privacy of individuals. However, the application of these techniques to a data set can also have significant impacts on the utility of the data for analytical purposes. For example, decennial census data obviously cannot be used to analyze the relationships between different characteristics of individuals or households. The smallest unit of analysis is the block (in SMSA areas), and even here, the amount of data reported is limited and suppression often has significant impacts on the results. In the case of the EIA/DOE's energy consumption surveys, the suppression of geographic identifiers prohibits state level analyses, among other desirable topics of investigation. In addition, however, the masking techniques used by the EIA/DOE (i.e., alteration of the billing history information) may have even more serious effects on the analytical utility of the data. It is likely that these techniques have little or no effect on the estimates of total values. However, the effects on subsample totals and on the results of analytical explanations (for example, multiple regression) are unknown and this author is not aware of any published attempts to estimate the possible effects. As a result, the analytical utility of the data may be severely limited.

The point of this discussion is that there is a strong inverse correlation between the traditional procedures for protecting individual privacy and the utility of the data for analytical purposes. As is too frequently the case, in order to protect ourselves against unethical usages of data, we have restricted and in some cases prohibited legitimate and profitable usages of the data. The task, and the point of this paper, is to design procedures which strike a balance between these two competing interests.

THE PRIVACY ACT OF 1974

The Privacy Act of 1974 (Pub. L. 93-579) is most frequently cited as justification for the suppression and masking of data. Sponsoring governmental agencies either simply assume that the Act prohibits the publication of data in which

individual respondents might be identified or, if they understand the Act, fear that the procedures required by the Act in order to provide access to identifiable records will adversely impact the resulting data.

The first instance is a clear and simple misinterpretation of the Act, for it does not prohibit the publication of individually identifiable data.(4) What the Act requires is "informed consent." Respondents must be informed of the authority for and the purposes of the collection, what uses will be made of the data and by whom, and the effects on the respondent, if any, for not participating, prior to giving their voluntary consent to participate in the data collection.(5) Since the Act has been in place for nearly ten years, it is likely that few experienced sponsoring agencies continue to suffer under this misinterpretation of the Act.

A more frequent reason for engaging in data suppression and masking is likely the sponsoring agency's hesitancy to inform respondents that certain users may be able to identify them. The agency fears, first, that response rates will be adversely impacted; second, that in an effort to avoid refusals, interviewers may avoid clearly informing the respondents; and third, to avoid the second problem, the sponsoring agency must require the respondent to read and sign a consent form, which, in turn, will have further adverse effects on the response rates.

In certain instances, the sponsoring agency's interest in complete confidentiality for its respondents is entirely justified. For example, in surveys of individuals who engage in illegal activities, or of individuals who have been the victims of personal crimes such as rape or family abuse, or of individuals whose behavior might be considered unethical or reprehensible by others (for example, extramarital sexual relationships), complete respondent confidentiality is required in order to obtain accurate information, to achieve respectable response rates, and, in some cases, to protect the life of the respondent.

In data collections involving less sensitive topics, however, such guarantees of complete confidentiality are neither necessary (in terms of insuring high response rates and a high degree of response validity), nor advantageous (in terms of producing data of maximum analytical utility). Concerning the effect on response rates, a major study by the Census Bureau and the National Academy of Sciences in 1976 included an experimental design in which respondents were assigned to one of five treatments in which the nature of the statement concerning confidentiality was systematically varied.(6) Although the study concluded that there was a statistically significant difference in the refusal rates between respondents presented with offers of complete confidentiality, on the one hand, and those presented with statements that answers might be publicly available, the difference was only 1 percentage point (1.8% versus 2.8%).(7) In addition, a higher percentage of refusals occured prior to the reading of the confidentiality promise (2.9%). As Turner concludes,

> Obviously, there are other factors besides confidentiality conditions that made someone refuse a Census Bureau survey, and our evidence does not support the notion that confidentiality concern is the principal motivator.(8)

As we emphasized previously, the effect of confidentiality conditions on response rates is likely to vary depending on the sensitivity of the survey subject matter. Unfortunately, there are no well-defined studies which precisely document these effects.

The Census Bureau commissioned another experiment in an effort to determine whether varying levels of confidentiality have any impacts on the validity of responses.(9) Conducted by Response Analysis Corporation in November, 1976, the study involved a matched sample of 500 households in Taylor, Michigan. The results of this small experiment suggest that the varying conditions of confidentiality have no significant impacts on the validity of responses to nonsensitive questions, including income.(10)

In general, then, there does not seem to be any advantage in offering guarantees of complete confidentiality to respondents in nonsensitive surveys. The validity of the data is no greater than it would be if the confidentiality guarantees were less stringent. However, the application of procedures to insure complete confidentiality can severely diminish the analytical value of the data. Thus, in many circumstances, unconditional guarantees of complete confidentiality are distinctly disadvantageous.

CONTROLLED ACCESS

In the two cases discussed previously, the procedures for protecting individual privacy were implemented unconditionally. That is, everyone except the sponsoring agency was provided with the same suppressed or masked copy of the data. As a result, even users who can guarantee restricted access and who use the data for statistical purposes only, are prohibited access beyond the single published level. The result can only be a restriction of unknown extent on our ability to understand social processes. This seems not only wasteful, but tragic, especially in view of the crises which all societies currently are facing.

In place of the procedures described thus far, we are proposing here a technique which we shall call "controlled access." Strictly speaking, this technique does not "replace" suppression and masking. Rather, it differentiates between users on the basis of some well-defined criteria, and grants them varying levels of access to the data. Thus, certain users may be provided unrestricted access to the data, others may be granted partial access, while still others are permitted access only to completely confidential versions of the data.

Clearly, the user screening criteria, the procedures for applying the criteria, and the procedures for enforcing the contingent user restrictions are key elements in a controlled access environment. Among the user screening criteria, it is likely that the sponsoring agency will want to include the nature of the user's analytical objectives, the ability of the user to control access to the data, the user's potential for invading the privacy of respondents, and the potential harm that would result to the respondent from such invasions. For example, a sponsoring agency might restrict full access to users interested only in statisical analyses, who present little potential threat to the respondents, and who agree in writing not to contact any of the respondents. Further, the sponsoring agency might provide such access only on the agency's own premises.

Amoung the procedures required by the Privacy Act prior to the establishment of a new system of records is the the designation of a records system manager. Depending on the anticipated demand for the data, the frequency with which the sponsoring agency collects data, the sensitivity of the data, and the potential harmful effects resulting from abuse, the agency may wish to leave all judgments in the hands of the system manager or, on the other hand, may wish to establish an elaborate mechanism of review committees and appeal processes.

With regard to enforcement procedures, users who are provided access to other than fully confidential versions of the data should be required to sign contractual agreements. The agreements should clearly specify the data to be provided to the users as well as the applicable restrictions on the distribution and permissible uses of the data.

AN EXAMPLE

The Bonneville Power Administration (BPA) is a power marketing agency within the U.S. Department of Energy. To assist in resource acquisition and transmission contruction planning and decision-making, BPA develops forecasts of energy demand. The forecasts are produced by relatively sophisitcated, data intensive computer simulation models. To support these models, BPA conducted an "energy consumption" survey in 1979. Personal interviews were conducted with approximately 4,000 residents of the Pacific Northwest region (Washington, Oregon, Idaho, and Montana), and fuel billing histories were obtained for those respondents who signed waiver forms. In designing the survey, no provisions were made either for maintaining the list of respondent names and addresses or for providing the raw data to the participating electric utilities and natural gas companies. As a result, the electric utilities and natural gas companies were unable to obtain copies of the raw data which were of analytical utility to themselves. Since the utilities had voluntarily invested some of their own resources in the survey (the utilities selected samples of their own customers and provided the fuel billing histories for their customers), utility analysts and managers were less than pleased with the result.

At about the same time, BPA joined the EIA/DOE in an experimental survey of commercial buildings in the Pacific Northwest. The original purpose of the survey was to test the feasibility of using utility billing records as a sampling frame as compared with more traditional areal sampling techniques. In return for using three Pacific Northwest areas as the test sites, BPA contributed sufficient funds to insure the completion of the fieldwork and processing of the data. Unfortunately, the terms for delivery of the data were not entirely clarified prior to the initiation of the survey so that BPA and the participating electric utilities were seriously disappointed when they discovered that the final data were fully suppressed and masked using the usual EIA/DOE procedures.

As a result of these experiences, when BPA began preparations for a second Pacific Northwest Residential Energy Survey, several electric utilities clearly and firmly articulated their desires for guaranteed access to data of maximum analytical utility to themselves. Indeed, unless BPA would satisfy these desires, several utilities made it clear that they would refuse to participate in the survey. Thus, the machine-readable copy of the data to be made available to the participating electric utilities should contain a code identifying the serving utility, together with the respondent ZIP code, all interview responses and complete, unaltered billing histories for each survey respondent.

Given this information, an interested electric utility could identify its own customers by matching the billing history information from the survey data with their own master records. With the exception of certain potentially idiosyncratic cases, however, the respondents would not be identifiable to any other agency or organization. Thus, from the standpoint of potential invasions of respondent privacy, the user audience can be easily and clearly divided into two groups: the electric utilities and natural gas companies, on the one hand; and

all other users, on the other hand.(11) Conveniently, several other criteria divide the user audience into the same two groups. For example, apart from BPA and the respondents themselves, only the electric utilities and natural gas companies have invested resources in the survey--the electric utilities assisted in the selection of the customer samples and the electric utilities and natural gas companies will be asked to provide billing histories for their customers.

Second, the billing data which the electric utilities and natural gas companies collect and maintain for all their customers is itself proprietary. Thus, the electric utilities and natural gas companies are experienced at, and have procedures in place for restricting access to certain data sets. Third, the electric utilities and natural gas companies have legitimate analytical interests in the data--to support their own planning and decision processes.

On the other hand, there is a potential for the electric and natural gas companies to abuse the privacy of the survey respondents based on the data from the survey. For example, the survey inquires about the presence of various conservation measures in the dwelling unit. The utilities could use this information to target conservation promotion campaigns. In an effort to prevent such abuses, BPA has developed an agreement which each requesting utility is required to sign prior to receipt of the data. By signing the agreement, the utility agrees to restrict access to the data to its own employees whose official duties require access; to refrain from contacting the respondents as a result of their participation in the survey; and to refrain from discriminating against the respondents. The agreement was reviewed by BPA's General Counsel and by analysts and attorneys of several local utilities prior to final implementation.

All other interested analysts will have access to a version of the data in which elements by which the user could identify individual respondents will be suppressed or otherwise masked. That is, respondent ZIP codes will be removed and elements such as respondent race, household income, and dwelling unit size will be examined to determine whether they enable the identification of individual respondents. If so, certain categories of these variables will be collapsed or, if necessary, the elements will be removed from the data set. Since the electric utilities and natural gas companies are the only users capable of identifying individual respondents through the billing history data, it will not be necessary to alter this data in order to protect the privacy of survey respondents.

As required by the Privacy Act, the respondents will be fully informed of the authority for and objectives of the survey, who will have access to the data, and the purposes for which the data will be used and that each respondent's serving electric utility and, where applicable, serving natural gas company may be able to identify them. This information will be presented verbally and in writing at the outset of the interview. In addition, near the end of the interview, each respondent will be asked to sign a form authorizing the respondent's serving electric utility and, where applicable, natural gas company, to release the respondent's billing history to the fieldwork contractor and, ultimately, to BPA. At this time, the respondents are once again informed, both verbally and in writing, that the information may be provided to their electric utility or natural gas company and that these companies may be able to identify them. Thus, the signed authorization form serves the dual purpose of authorizing release of the respondent's billing history information and documenting the respondent's informed consent to participate in the survey.

The fieldwork for the second Pacific Northwest Residential Energy Survey was scheduled to begin May 23, 1983. Thus, what effects, if any, the proposed confidentiality statements will have on response rates is yet to be determined. Needless to say, the BPA staff will monitor the response rates closely, and there are plans to conduct an analysis of the response rates as soon as possible following the completion of the fieldwork.

With regard to the BPA-utility agreements, generic copies have been distributed to all the participating electric utilities. To date, ten of the 57 participating utilities have expressed an interest in obtaining the data and a willingness to sign the agreement.(12) Several other utilities reviewed a previous draft of the agreement and, after submitting comments, expressed a willingness to sign.

SUMMARY AND CONCLUSIONS

The right to individual privacy is one of the tenets of the American political and legal system. The Federal Government has sought to protect the individual right to privacy through legislation like the Privacy Act of 1974. In response to this and other legislation, federal agencies which collect and publish data have sought to avoid privacy invasion through data suppression and masking techniques. The unconditional use of such techniques has the unfortunate consequence of impairing further analysis of the data.

This paper has offered an alternative procedure which seeks to balance the interest in protecting the privacy of the individual respondents with the interest in maximizing the analytical utility of the data. The procedure involves the provision of differential access to the data based on some well defined criteria. Thus, users with legitimate interests in conducting statistical analyses, who present little threat to the privacy of the respondents, who contractually agree not to violate the privacy of the respondents, and who either can demonstrate an ability to limit access or agree to use the data on the sponsoring agency's premises, may be granted acccess to the complete set of data. Users who do not satisfy all of these criteria may be granted access only to partially or fully suppressed and/or masked versions of the data.

For purposes of illustration, the controlled access system developed by the BPA for its second PNWRES was presented. This survey is just now going into the field, so what effects, if any, the confidentiality statements have on the response rates is not yet determined. To date, none of the utilities expressing interest in obtaining the data have expressed any hesitation to signing the agreement. It will likely be several years before we know whether any of the utilities have violated the agreements, or whether there are any other problems with enforcing the terms of the agreements.

NOTES AND REFERENCES

(1) See A.G. Turner, "What Subjects of Research Believe About Confidentiality," in J.E. Sieber (ed.), The Ethics of Social Research: Surveys and Experiments, New York: Springer-Verlag, 1982, pg. 152.

(2) See D.T. Campbell and J.S. Cecil, "A Proposed System of Regulation for the Protection of Participants in Low-Risk Areas of Applied Social Research,"

in J.E. Sieber (ed.), The Ethics of Social Research: Fieldwork, Regulation and Publication, New York: Springer-Verlag, 1982, pg. 110.

(3) The EIA/DOE also conducts surveys of commercial energy consumption. For purposes of illustration, we focus here only on the residential surveys. The techniques used to protect respondent confidentiality in the commercial surveys are basically similar to those used in the residential survey data bases.

(4) See 5 U.S.C. Sec. 552a and American Statistical Association, "Report of Ad Hoc Committee on Privacy and Confidentiality," The American Statistician, vol. 31, no. 2, pgs. 59-78.

(5) See 5 U.S.C. Sec. 552a,c.

(6) Reported in A.G. Turner, op. cit., pgs. 154ff.

(7) Op. cit., pg. 159.

(8) Ibid.

(9) Op. cit., pgs. 160ff.

(10) Op. cit., pgs. 160-161.

(11) Since BPA markets and transmits electric power only, our attention thus far has focused solely on the electric utilities. However, the fuel supplier survey portions of the end-use surveys include natural gas billing history data and the natural gas companies frequently are interested in obtaining copies of the final data.

(12) It should be noted that many of the utilities do not have the machinery or personnel to conduct statistical analyses, and are therefore not interested in obtaining a copy of the final data.