CONFIDENTIALITY OF SURVEY DATA -- PANEL DISCUSSION

Moderator: Ida C. Merriam

Panel: Robert Boruch, Northwestern University
Tore Dalenius, Brown University
Joseph W. Duncan, Office of Federal Statistical Policy and Standards
Thomas B. Jabine, Energy Information Administration

Ida Merriam:

Privacy and confidentiality of information about individuals and businesses have been subjects of growing concern over the past five or ten years. The ASA has paid considerable attention to the problems involved, particularly as they affect statisticians and the Federal statistical system. In March 1975, the president of the Association, Lester Frankel appointed an Ad Hoc Committee on Privacy and Confidentiality to review and evaluate the statistical implications of current legislative and other proposals. The Committee developed a report - which was published in the American Statistician for May 1977 - that laid out certain principles and discussed their implications. Basically, the report emphasized two factors: the need to find a balance between the individuals' right to privacy and society's need for information-- a theme that runs through all current discussions of the subject--and the fundamental importance of the distinction between data collected for administrative purposes and data collected solely for statistical and research purposes. Most of you are probably familiar with the report of the Ad Hoc Committee. It is worth rereading. The Ad Hoc Committee was followed by a continuing ASA Committee on Privacy and Confidentiality which is attempting to monitor current developments and make sure that the special concerns of the statistical profession are recognized.

This session focuses on survey data and will primarily concern itself with confidentiality rather than with privacy. The two concepts are interrelated, of course, and survey statisticians must pay attention to both. But they are not the same. Privacy relates to the individual's rights to be left alone and not asked to divulge more information about himself than is needed or is justifiable. Confidentiality, on the other hand, is concerned with the restrictions placed on the use and dissemination of information provided on surveys or otherwise. We are going to assume that all surveys any person of this audience would sponsor or support are fully justified as to purpose and general approach, and give our main attention to the problem of assuring the confidentiality of the data collected without hindering analysis and research. First we will review the general issues involved in achieving confidentiality of survey data.

Joe Duncan:

What is the general need for the protection of survey data? Why are we so concerned about it? The need is very simple. It is essential to respondent cooperation for them to know that the data they provide is without any risk to themselves; that it will not in any way affect their own future, including their rights, benefits, or privileges under Federal programs or in any other manner. That sounds like a fairly simple need, yet when the statistical agencies' activities are explored today, we find that in many cases it is very difficult to tell the respondent that the data are absolutely restricted and highly confidential.

In contrast to the respondent's need for protection of the confidentiality of the information provided, there is a statistical and sometimes a broad government need for access to individual responses. First of all, the statisticians would like to review the individual survey instruments to make certain that the data files prepared from them are accurate and that the coding is correct; to check one file against another to make sure that the information appears reasonable; and a variety of other file enhancement activities that are largely statistical in character. These statistical needs have the characteristic that there is a sharing of the data between the original collector and somebody else that is also working with the information. Beyond the statisticians need in government we have seen, especially in the last decade, a growing need by the regulatory agencies, or at least a perceived need, that they too have access to statistical information. Recently, after the energy crunch and latest gas lines, there were Congressional hearings saying "Isn't it terrible that statistical information provided is not available to use in the Congress or to regulatory agencies so we can check up on the oil companies." This is but one example. Plus there is a need generally for avoiding duplication in obtaining information that already exists someplace else.

So this leads to some of the general issues that we will talk about today. First of all, how do you provide clear, legal protection? Second, how do you build the public's confidence so that they will feel comfortable that the government which is regulating them is not using the statistical information gathered from these surveys for individual cases.

Third, how to assure that statistical agencies can share information so that statistics can be evolved with a minimum of burden and a maximum of accuracy.

Ida Merriam:

Can we now look in more detail at the need for legal protection of survey records.

Tom Jabine:

To understand why legal protection is needed for survey records, first we need to ask what is being protected. We are talking of records from statistical surveys, that is, surveys in which the identification of specific individuals is not material to the final results and uses that will be made of the survey data. Second, we are talking only about records that are individually identifiable, either because there are particular identifiers such as names or social security numbers associated with them, or because the detailed content of the record may enable people with outside information to identify particular individuals. If the records are not identifiable, then at least for government-sponsored surveys, the results, including micro-records should be available to anyone with perhaps some exceptions for security reasons. However, there is no clear dividing line between those records that are identifiable and those that are not identifiable. This matter is currently being tested in the courts with respect to particular data sources. We can talk about either public or privately-sponsored surveys. However, I believe that legal protection for survey data, where the government has no role as the sponsor either through contract or grant, is an unlikely prospect in the near future.

What are we protecting against? We are protecting against the disclosure of identifiable records for purposes not directly related to the survey. There may be exceptions to this but the possibility of their occurrence should be well understood by the people involved in the process -- the data subjects and the data collectors. Disclosure may be compulsory, as in response to Freedom of Information requests, subpoena or some other legal process. It may be inadvertent through inadequate physical security or through statistical disclosure in the publication of results, or it may be voluntary on the part of the surveyor. We do not necessarily require protection against all types of disclosure. There are some strong arguments that are made for disclosure for statistical purposes not directly related to the initial survey; and possibly for some other purposes such as disclosure to the subjects of the data for their own use or for reasons of health and safety. What is required is control over disclosures and the agreement of the parties involved.

Why do we say that legal protection is needed? First of all, why is any protection needed? That was covered by Joe Duncan; the reasons are equity to the survey respondents and to create a climate in which people can respond to surveys and not feel that the data will be used to affect them as individuals directly. There is a question of equity and there is also a question of not undermining the ability of statistical agencies to collect complete and accurate data to guide public policy. There is some fairly good experimental evidence showing that inadequate protection can hurt the quality of response. The study by Eleanor Singer, with the National Opinion Research Corporation, shows that stronger assurances of confidentiality led to less item nonresponse. Similarly, the study that was carried out by the Committee on National Statistics with the support of the Census Bureau showed that the level of confidentiality promised to respondents had small, but significant effects on response rates. There is a lot of other less direct evidence of the effect of inadequate guarantees of confidentiality. Probably business respondents are more concerned than individuals because they are more aware of what may happen to the information that they provide.

Finally, why do we talk about legal protection? The actions of Federal agencies are guided by statutes, regulations, and policies. The statutes are probably more pertinent than regulations, which have to follow the requirements of the statutes, and the regulations may be more lasting than policies. Ideally, the protection of statistical records should be statutory rather than merely a matter of policy by an agency, which can be fairly easily reversed. Clearly, the present coverage of confidentiality of government surveys by statutes is deficient. A few agencies and some types of surveys have strong protection; however, a price often paid is that the access of others to data from those agencies and surveys is unnecessarily restricted. Other agencies have essentially no protection in their statutes and are vulnerable to Freedom of Information requests and other kinds of compulsory disclosure. The challenge we are faced with is to provide adequate statutory protection without placing unnecessary restrictions on the use of survey records.

Ida Merriam:

Is there any disagreement in this panel on the need for moving beyond what we have now in the way of statutory protection?

Tore Dalenius:

The legal community has, by and large, been most responsive to demands for legal protection of our privacy. It may be argued that it has sometimes reacted too quickly and too strongly. I would welcome that any proposed legislation in this area be subjected to a thorough "requirement analysis" in close cooperation with all parties involved. When legislation is called for, the merits of a proposed law should be assessed with respect to not only the privacy protection it provides but also its impact on the possibilities of making surveys and taking censuses in an efficient manner. To illuminate the importance of this point, I mention the legal restrictions concerning the use of imputation for missing data, which have been introduced in Sweden. Legislation must not be developed in isolation from the needs of the society of information on which to base its program.

Bob Boruch:

One of the things we have tried to do recently is examine the extent to which cooperation in surveys has been influenced notably by stronger assurances of privacy and confidentiality. I agree with Tom Jabine's remark that indeed the Census Bureau's and Singer's studies bear out the need for strong assurance. It is remarkable, however, that so many people in these studies choose to respond without much in the way of assurances. Further it is relatively easy to find smaller experiments and case studies in which people respond irrespective of the assurance that is made, they ignore the assurances made, or they forget the assurances that are made at the beginning of the interview. The point here is that an argument for stronger protective legislation must recognize that there are a subset of people out there who probably respond to a questionnaire written in crayon and on a brown paper bag. These ought to be recognized, just as we must recognize respondents who appreciate assurance.

The second point is related. If we look over the ten or twenty controversial cases in which it has been announced that privacy has been a major factor in the disruption of a survey or a large-scale program evaluation, we see that it is typical that the privacy issue is confounded almost inextricably with a number of other issues. Those other issues include group privacy. For example, during the campus protest years, the Students for a Democratic Society objected to researchers' studies of campus unrest; part of the objection concerned individual privacy while a larger part concerned what is now considered group privacy. Errors in reporting perceptions are chronic both in newspapers and in magazines. There is, for instance, a nice classical case study done in Norway on Project Metropolitan's large-scale developmental survey of adolescents in which some disruption was apparently caused by public confusion over the distinction between "sociological" and "socialist." In any case, the point is that there are other issues that are tangled with privacy matters and separation of the issues is in some cases at least as important as identifying the extent to which privacy plays a role.

My third point is a question. Tom, you alluded to a couple of court cases in which the manner of identifiability of data was an issue. Could you expand on that?

Tom Jabine:

In regard to your first point, I do not think we should take too much comfort from the fact that there is a large body of the population which seems to respond to almost anything without too many objections. If there were a situation in which it were found that survey data were being used for a non-statistical purpose and this received wide publicity then many members of this group might become very rapidly aware of what was happening. This could be a real disaster for the ability to do a census or anything else that depends on high cooperation rates.

The particular case that I had in mind was the Long case. Susan and Phillip Long are two tax scholars who have been doing battle, so to speak, with the Internal Revenue Service over a number of years for access to various data sets that IRS maintains. Most recently they have requested some microdata files with the specific identifiers removed from the IRS Taxpayer Compliance Measurement Program. This program uses a probability sample of taxpayers and the data base contains the detailed data from their original returns as well as the results of the audits that are performed for this sample of cases. The most recent development I am aware of occurred in the Appeals Court. The District Court said that the IRS was not required to make this information available under the Freedom of Information Act. The Appeals Court reversed this ruling saying that in general they would have to release the records unless they can show that there is some significant risk of individuals being identified by indirect methods through the content of the data files. Now the District Court is being asked to rule on the question of whether the risk of disclosure of individual data is or is not acceptable.

Ida Merriam:

The statutory protection of the statistical system to achieve confidentiality and better statistics is also a protection for the statistician who wishes not to be forced to release statistical data.

Bob Boruch:

I am not sure about the extent of that problem in our own situation. At Northwestern, we have been engaged, for the past four years, in secondary analyses of other people's data, where data were generated originally from large scale social program evaluations. In about half of the cases where we meet researcher resistance to giving up their statistical data, it is not deliberate and malicious. Rather, resistance is due more likely to incompetence or indifference in record keeping. Some people resist for other reasons, such as proprietary interest: they would like to exploit the data well before they give it to someone else. A minority, if any, may want to hide something, but that is very difficult to verify. We have not pursued cases of the sort.

There is however, an important recent case, Lora versus the Board of Education of the City of New York, (74 F.R.D. 565 (E.D.N.Y. 1977) in which the school system decided not to release statistical data. That is, it decided to resist disclosure of statistical data to the Office of Civil Rights on grounds that the individual privacy would be compromised. They attempted to argue that releasing such data affects individual privacy, not in order to impede the OCR's attempt to establish discrimination in the school system. This sort of thing, I suspect, will be more likely; it certainly has a longer history.

Joe Duncan:

I agree that I would like to protect statisticians but there is a risk in tight confidentiality which has to be addressed in the legislation that is drafted. That is you have to protect the public by assuring that the statistician is doing a good job. In other words, there needs to be an opportunity to investigate the records, audit the way in which the statistician made his estimates. So in the sense that you create very strict fortress-like regulations you may eliminate the opportunity for groups like the Congress to audit the statisticians in the government.

Tore Dalenius:

I want very much to join Dr. Duncan on that point. Let me also emphasize that we statisticians have a responsibility to our profession to assess critically allegations about disclosures and when these allegations prove to be false, to counteract. As an example of false allegations, it has been claimed by a prominent writer on privacy issues that the Census Bureau once disclosed sensitive information about identifiable physicians; this example has received wide publicity. It appears, however, that this story has no basis in fact; it is a genuine fabrication.

Ida Merriam:

What is it that statisticians themselves have been doing to make sure that individual information is not used as it should not be?

Tore Dalenius:

The invasion of privacy problem has long been recognized by statisticians and, more generally, by social scientists and the like, in fact long before the public debates in the 1960s. The technological development in the recent past has, of course, enlarged the problems, but it has also provided more powerful tools for coping with these problems.

In the setting of statistical investigations (surveys and censuses), it is clarifying to distinguish three types of concern about invasion of privacy, linked, respectively, to: (1) the substantive focus; (2) the methods used for data collection, processing and storage; and (3) the consequences of releasing the results.

While statisticians pay attention to all three types of concern, they have -- rightly, I think -- paid special attention to the two last-mentioned types. I will give a review of their endeavors; it will by necessity be limited to a few selected example.

With respect to methods for data collection, processing and storage, statisticians do refrain from using methods for data collection (such as "participant observers") which deprive subjects of their control of self-presentation.

When it comes to collection of "sensitive information", statisticians exercise considerable efforts to reduce the possible reluctance of the subjects to providing such information. In a survey of personal income, for example, it is a long-established practice to collect what I will call "interval data" rather than "point data": the subject is asked to state the interval in which his/her income lies. In the last 15 years, statisticians have in addition developed new methods for collecting "sensitive information" in a way that gives the individual subjects protection. Two examples are the techniques for "randomized response" and "combined questions".

By the same token, statisticians carefully protect the data collected in the course of processing and storage. Thus, elaborate schemes are used to control access to the facilities where the processing takes place, and likewise to the computers themselves on which the processing takes place.

If the data are not destroyed after they have been processed, schemes for data protection are used. Two examples are "file splitting" and "encryption".

With respect to the consequences of releasing the results, methods nowadays known as methods for "disclosure control" have long been used in official statistics. Clearly the development of new media for release of statistics has enhanced the problems. The computer, for example, has made it economically feasible to release "micro-statistics". But at the same time, the computer is a key instrument in implementing schemes for disclosure control; using such methods as "cell suppression" or "privacy transformation" is an integral part of the operations which yield the results.

In summary, statisticians can rightly be proud of their contributions to endeavors to strike a rational balance between society's need to know and citizens' right to privacy.

Bob Boruch:

It is puzzling that legal scholars who are also interested in statistics, have not examined randomized response and similar strategies for protecting data to judge the extent to which they are good, bad, or indifferent relative to legal standards. Based on some of the work that a colleague and I have done, it does seem clear that there is a precedent for the contention that, in some courts, probalistic evidence about the individual of the sort produced by randomized responses would not be admissible as evidence. In other courts, it might be admitted but there is usually a good deal of controversy over it and so would be of limited probative value. In any case, if one defines invasion of privacy or identifiability in deterministic, rather than probalistic terms, it is clear that the method does not let you identify anybody or attach meaning to the response.

Tom Jabine:

The question of how safe the data are from statistical disclosure was addressed by a subcommittee of the Federal Committee on Statistical Methodology, which is chaired by Maria Gonzalez. The Subcommittee, in working over about two years to produce a report (number 2 in the OFSPS Statistical Policy Working Paper series), found some instances where they felt that statistical disclosure had occurred. However, in spite of zealous efforts, including a specific request that was published in the Statistical Reporter, the Subcommittee was not able to come up with examples where individuals felt that they had been harmed as the result of any statistical disclosure. I think it is important to recognize that if we use a broad definition of statistical disclosure as the Subcommittee did, some disclosure will take place in virtually any kind of release of aggregated data if you know which individuals are members of the population on which the data are based.

I would like to get back to one thing that Tore said about not using participant observers. I heard of a case recently which would probably come under the heading of non-participant observers. The Environmental Protection Agency is now doing some surveys in which they have observers at retail gas stations. The observers know which pumps are being used - leaded or unleaded; they record the license numbers of the automobiles and then by checking with the Motor Vehicle Bureaus they can find out what kind of cars these are and which kind of gasoline they should be using. How do you feel about that kind of study?

Tore Dalenius:

I will give a specific example. A social scientist undertook a study of homosexuality. As a means of getting a sample of homosexual men, he observed secretly people in the men's room at a bar; the behavior there of some men indicated that they were homosexual. When they left by car, the social scientist recorded the license plate of their cars and was thus able to get the names and addresses of the drivers. Authorities on ethics in research with human beings would no doubt consider the procedures described unethical; I fully agree.

Joe Duncan:

I think the principle I started with in the beginning that in a statistical survey the data of the respondent are confidential and do not directly infringe on his privilege, has been violated; I doubt that he has been looking at it as a statistical survey as much as punishment or model building activity.

Tom Jabine:

As I understand it the EPA survey was purely statistical. They are not taking any action against the violators.

Tore Dalenius:

The problem would be that even if they do something which is not a statiscial survey, the public may conceive of it as a statistical survey and generate reactions against the survey. That is a risk that is being taken.

Ida Merriam:

It would be worth reviewing legislation which provides protection against release of data files relating to specific research areas.

Bob Boruch:

There are indeed a small group of laws that provide a researcher with some protection against appropriation of information on identifiable research subjects. As such, they facilitate the researchers' adherence to ethical standards, and to the extent that research participants find legal protection necessary, desirable, or attractive, then the laws may also help induce people to cooperate in research. Each of the statutes focuses on a different area of research, for example, mental health, criminal justice, and others. Each contains slightly different provisions affecting quality and level of protection.

The standards used to gauge quality of these laws generally include matters like: whether immunity is automatic rather than must be authorized in each case or for each project; whether immunity refers to administrative, or judicial, or legislative agencies seeking to appropriate the data or the individual records; whether all information or just identifiers is covered by the statute, and whether provisions for secondary analysis are included in the statute. My description here is brief, adapted from a more thorough treatment in Boruch and Cecil (Assuring Confidentiality in Social Research, University of Pennsylvania Press, 1979).

First among these is the Public Health Services Act, which entitles the Secretary of HEW to authorize persons engaged in research on mental health, including research on the use and effect of alcohol and psycho-active drugs, to protect the privacy of individuals who are subjects of the research. The protection is sustained by legally mandating the researcher to withhold from persons not connected with the research the names or other identifying characteristics of such individuals. A researcher so authorized cannot be compelled in any Federal, State or local civil, criminal, administrative, legislative or other proceedings to identify the individuals. The Act is of considerable importance to mental health researchers who find the identity of respondents can be protected legally; moreover, the references to "other identifying characteristics" assures that the researcher can attempt to prevent deductive disclosure if that was a problem. The protection must be authorized formally by the Secretary. Consequently, researchers who initiate politically controversial types of research, such as the effects of marijuana on sexual behavior for example, are at risk of not obtaining the grant of immunity or having it rescinded. There is no provision in this Act for disclosure of records for secondary analysis.

The second important statute is the Crime Control Act of 1973 which directs attention to criminological research. It is a bit more complete in coverage than the Public Health Services Act, in that it specifies that information contained in research records or copies of them must not be disclosed by the researcher to any person other than the person from whom it was collected. The use of identified information in judical and administrative proceedings against the individual research participant, is expressly prohibited. Unlike the Public Health Services Act, this law provides immunity to all information not just to identifiers. It does not, however, prevent legislative committees from appropriating records as evidence.

The Controlled Substances Act, the third stereotypical law in this class, authorizes the Attorney General to permit persons engaged in research on controlled substances to withhold identification of research subjects from being disclosed in legislative, administrative, criminal and other proceedings. It is an important statute to professionals such as statisticians, sociologists, and so on who are engaged in such research. Like the Public Health Services Act, the grant of immunity depends on authorization by a Federal agency executive and as such it is subject to the same problem--pressure to refuse or grant the demand on political rather than scientific grounds. A related law, the Drug Abuse Office and Treatment Act protects records of patients maintained in connection with drug abuse prevention programs assisted by the Federal government and the implementing regulations include coverage of records maintained for research purposes. The law covers material other than identification of the subject, and moreover, it is automatic rather than dependent on the executive authorization. The immunity is limited in that a court may in fact subpoena the identification and identified records.

There are other acts, such as the Health Services Research and Medical Libraries Act of 1974, other special legislation, which was enacted for the special studies for runaway youth, venereal disease, and the like, which also protect data of certain kinds. The general character of these laws is such that the level of protection is mixed and interpretation sometimes gets very difficult; each one has some disadvantages that another one may not have. Partly in the interest of clarifying the matter and generating broader, more coherent and better coverage for the researcher and the respondent, President Carter introduced the Privacy of Research Records Act in April 1979 which is among the pending legislation which Joe Duncan and Tom Jabine will discuss.

Joe Duncan:

As Bob just mentioned, in April there was a bill introduced that was called the Privacy of Research Records Act; but there has been no action on this bill thus far; there have been no hearings and there are no scheduled hearings to date. The committee is tied up with legislation affecting Medical Records, which is, of course, related to medical research.

Let me read the key section of that proposed Act that sets forth the principles; it says: "Research records collected or maintained for a research purpose or collected with the assurance it will be only used for research purpose by an agency or by a contractor or grantee of an agency, shall be confidential, shall not be used in whole or in part in individually identifiable form to make any decisions or to take any action directly affecting the individual to whom the records pertain." That is a very clear statement of objectives and goals which directly relate to my opening remarks. The bill itself is then filled with a series of exceptions as to when that principle will not hold. I will just touch on this very lightly so you can see the complexity of issues. This bill provides that they can be disclosed with the prior written consent of the respondent. It can also be disclosed in the case of medical emergencies; for example, disclosure is authorized to deal with the problem when something is found out during medical investigations that needs to be dealt with immediately and that was not anticipated in the medical case; this is a very specialized case.

The bill does specify a very limited set of situations in which the judicial proceedings can obtain the record; the access is quite restrictive and, in my judgment, quite satisfactory. They provide an exception that deals with the audit of the research itself; the case I mentioned a little earlier in our discussion. This provision makes it possible for an outsider to examine the research itself to verify that the records were properly maintained, properly interpreted and so forth. This way we could verify the research itself. Finally, there are exceptions for related research purposes, an exception which is also rather carefully controlled.

This legislation, like the next piece of legislation we will talk about--the statistical bill--has been drafted very carefully. The lawyers have spent a lot of time worrying about obscure details and implications. However, as the bill moves through the legislative process it is exposed to the risk that something will be added, dropped, or replaced with something else. So that even if it may be cleverly drafted at the outset, such bills merit very careful attention throughout the entire legislative process.

A second piece of legislation is one that we have been talking about for a very long time. It is called the Confidentiality of Federal Statistics Records. It has many of the same points that are in the Privacy of Research Records Bill. I happen to feel that overall the drafters of the statistical bill did a little better job of drafting. The bill on Federal statistics has not been introduced; it has been subject to agency comment and debate. My own personal judgment is that this is an excellent bill and I hope it sees the light of day.

What the proposed statistical bill does is, first of all, to establish a statutory basis for the traditional promise of confidentiality which has long been given to respondents of statistical collection. It is interesting to me that many of the sophisticated participants in the statistical system, that is the outside providers like corporations, have assumed that when they give price data to the Bureau of Labor Statistics they are protected. It is true that they are protected because BLS has an unblemished record of not releasing individual records. But I fear that under a legal test there might be some trouble because there is no specific piece of legislation that specifically guarantees the confidentiality of those records. Tradition and practice have provided a sense of confidentiality; what this piece of legislation does is put it into statutory form so that it is not subject to somebody's whim or policy change in the future.

The second thing this bill does, which I think is equally important and is of particularly great interest to the statistician, is it provides for a limited sharing of individually identifiable records for statistical purposes within the statistical system. Once the statutory protection has been given to an agency, these statistical records cannot be used in an individually identifiable form. They can, however, be traded among covered statistical agencies for purposes of enhancing the data file or verifying individual records so that statistical information will be improved.

The protected statistical centers that are named in the legislation at the moment include the Bureau of the Census, the Bureau of Economic Analysis, the Bureau of Labor Statistics, the National Center for Health Statistics, the Economics, Statistics and Cooperatives Service of the USDA, the Office of Research and Statistics in the Social Security Administration of HEW, and the Division of Science Resources Studies of the National Science Foundation. This would be the initial set. There is a provision that additional agencies would become certified if they will follow certain practices. For example, right now the legislation under which the Energy Information Administration operates has a mixing of regulatory and statistical purposes. That would not be compatible with this bill, but the energy bill could be amended. An enclave could be created within the Department of Energy that would then participate in this whole system. It also provides for the establishment of protected statistical files, that is, statistical files not held by these major agencies could be certified so that they would be confidential and protected as well by action of the Central Statistical Office.

There are a series of sanctions in the system. There are penalties if people violate this statute and there is a provision for controlling the exchanges among statistical agencies so that it is not a willy-nilly thing without need and without justification.

If both of these bills were enacted, in the forms in which they presently exist, many of the problems of the day-to-day use of statistical and research data would be, in fact, resolved. There is still a lot of effort that has to take place. For example, a major job of public education is required. Many people have a distrust of government. They believe that any data that they give may be used against them. It is going to take a long time before this feeling is offset. These bills do provide a rather clear basis to set aside statistical data collection and research data collection as highly protected, highly confidential pieces of information about individuals.

Ida Merriam:

What about the capability of differentiating between statistical files and other files with statistical information?

Tom Jabine:

The two bills that have been described take slightly different approaches to this. In the Privacy of Research Records Bill, there is a definition of research records. Then, it is up to each of the agencies, in advance of creating new records, to determine whether or not these are going to be research records and to treat them accordingly. In the case of the Confidentiality of Federal Statistical Records Bill, this is done primarily, although not entirely, by designating separate agencies or parts of agencies that are clearly identifiable as protected statistical centers. One of the principles is that any records that those agencies maintain, except for internal administrative records, must be treated as statistical records under this legislation. So those are two different approaches; but I would like to say that there was full coordination between the people who worked on these two bills, and they are intended to complement each other. The statistical records bill does certain things that the research records bill does not. First of all, it covers individual and business records, whereas the research records bill covers only records about individuals. Second, it sets up separate units, the protected centers, that can maintain only statistical records. It is a complex bill, but what it does was summed up very well by Dorothy Rice as discussant in a session this morning, namely, that the things that would be accomplished by this bill in terms of protection and sharing of information within the enclaves are things that would happen fairly naturally if this country had a centralized statistical system instead of a decentralized one. This bill offers a way of getting some of the benefits of centralization without losing the significant advantages that we get from decentralization, with the statistical units being close to their primary users.

Bob Boruch:

One peculiar aspect about the coverage of these laws has to do with the separation of educational statistics from all other statistics; by educational statistics here I mean also evaluation of Title I programs, surveys, and the like. I notice that no one in the President's Reorganization Project is from NCES. Yet, a lot of relatively innocuous educational statistics are collected, they sometimes generate privacy concerns, and they certainly are extensive enough to justify a place in Federal statistical systems theory.

Joe Duncan:

I will state one problem for you. The National Center for Education Statistics collects records from public schools, elementary and secondary schools. That information under this bill would then become protected and individual records could not be revealed. Part of NCES's role is to publish information about individual institutions. This presents a conflict since the intent of this proposed bill would be to protect the information of individually identified institutions.

Tom Jabine:

Your explanation is correct. The National Center for Education Statistics was consulted in the process of determining which agencies should be named as protected centers. It was clear that it was not appropriate for those reasons. However, there does remain the possibility that files from some of their surveys which deal with individuals rather than with institutions could attain the status of protected files.

Joe Duncan:

There is the possibility that one should create a subunit that is the protected statistical center and have another unit, for example, that handles the other types of information. The other possibility for this bill is to create protected files which are protected under the legislation.

One of the interesting things that the bill tries to do is to say that there is a lot of administrative information which the government collects which has a value for statistical analysis and for statistical purposes. This, in effect creates a one way flow process where data collected, let us say for regulatory purposes, flows into the statistical agency and is then enhanced within the statistical agency's data base. That is good for the statistical agency. Further, it is protected once it is inside the statistical agency. If somebody wants to access the regulatory information they have to go back to the original part of the department that collected the information. They cannot have access to the enhanced information which may include information beyond what was submitted from the regulatory agency.

Bob Boruch:

I would like to ask a question about both of the acts. One of the interesting features of the Medical Records Act, another bill submitted by Carter, which does not stand out in the research act or the statistics act, concerns those people who fraudulently represent themselves as physicians seeking medical records. Under the Medical Records Act, such people could be prosecuted for their deception. In our case, counterfeit researchers in market research are not unknown; Baxter, for instance, has written a couple of nice articles about salesman claiming to be pollsters. Is there any provision for impeding the counterfeit researcher in the new proposed laws?

Tom Jabine:

No, I do not know of any such provision. Both of these bills deal only with records that are maintained by or under the sponsorship of the Federal agencies and all the penalties relate to improper disclosure by the holders of the records.

QUESTIONS FROM THE AUDIENCE

Question--Aryness Wickens:

I would like to ask an operational question of Joe Duncan. You have been discussing this confidentiality situation for two years and a very considerable number of points have been raised. What is the present status of the legislation that you have been discussing today. Second, would you talk a little to us about the legislation required to assure that the statistical enclaves were adequate for their purposes? Where do we stand in all of this? What apparatus would make this legislation, which we have been discussing for the past several years, work?

Joe Duncan:

When this session of the ASA was planned, it was anticipated that the President's Reorganization Project study entitled the Statistical System Project would have been complete. There would be proposals both on the Hill or at least for the public and one thing we would do at this meeting of the ASA is discuss and debate the details of what has been proposed. In the case of the legislation, the draft legislation has gone through a process of interagency clearance and review which means there is a concensus that this is an appropriate position for the Administration to take. However, the final step in that process once it has cleared the agencies is to clear the Office of Management and Budget and the White House and become part of the President's legislative package. This has not yet happened, although we anticipate that, if there is an action on the Statistical System Project overall, the legislation will be part of that action. In fact, the legislation was basically available at the same time that the Research Records Act was available in April, and it was withheld partly because it was tied together with the Reorganization Project.

Now on statistical reorganization itself, the hope is that within a very few weeks, there will be a decision in the White House about the future organization of Federal statistical activities. The Statistical System Project basically focuses on one question: "How can we have stronger planning and coordination of the decentralized system?" It is not a study of whether the Census Bureau and BLS should be combined. The fundamental question of how should the statistical system be strengthened for planning and coordination has not been controversial. The issue has been where should the resources for planning and coordination come from and where should they be located organizationally?

When the mechanism for stronger planning and coordination is in place, there will be no problem with the operational dimension of this bill because it says that the Chief Statistician will make the determinations necessary as to what agencies are enclaves and what exchanges occur among agencies. In the absence of a strengthening of planning and coordination functions, it would fall back on my Office, and I would have to say that we have relatively limited resources to deal with the problem presently. It is not a major task in the limited sense that it is proposed here. At the start, it should be fairly easy to administer. It is the expansion of the protection to other data files and the establishment of new protected centers that would require more resources.

Question--Michael Lamphier:

This discussion has concerned itself principally on the matter of individual records--protection of individual personal confidentiality. Professor Boruch noted in passing that there has been some resistance to response by those who want to maintain group privacy. Is there any development in terms of the protection of any collective rights? Has this issue arisen for any official consideration in deliberations on the protection of statistical files where there might be the possibility of exposure of vulnerable minorities whether we are talking about ethnic minorities or those of socio-economic status, or whatever?

Joe Duncan:

I will give you a tangential reply. In the discussion of this proposed legislation, the answer is "no" directly, because the purpose of this legislation is to make it feasible for the individual records to be published very broadly in aggregates and all aggregates are subclasses of one sort or another. People in a subclass might suffer because a piece of legislation affects the subclass. These bills do not deal with that aspect, but there is still the philosophical question: What should you do about subclasses? The related development is that as a result of growing intentions of various minorities populations, which are defined quite broadly since women

are a minority under current definitions (even though they are a majority to statisticians), the pressure has been quite the reverse. This is, there is pressure for more data about minorities, defined various ways, so that newly defined minorities can get their rights along with the other minorities. What we have seen to date is concern by the minorities that they are going to have their rights violated in some way; and, therefore, they want to make sure that there is more information. That puts a tremendous demand on the statistical system, because the more subclasses that you want to try to define statistically, the larger your sample, the greater the tabulations costs and so forth.

Tom Jabine:

This is an historical comment on the questions. The predecessor of the present ASA Committee on Privacy and Confidentiality, the Ad Hoc Committee, did discuss this issue at length. I think the final conclusion was that we would not make any recommendations in that area. I think our feeling was that depending on what particular classification of individuals you were using, we are all members of minorities of one kind or another; if you try to place restrictions on data that affect particular population subgroups, you just have an intractable problem.

Bob Boruch:

I agree that the problem is terribly complicated and one ought to examine it. It does seem to be an old problem in the sense that early epidemiological surveys in New York during the late 1700's and early 1800's encountered a great deal of resistance from, for example, the Irish immigrants because they figured the government was out to get them or that government would change immigration laws in a way which was not in their interest. Certainly the problem affects other cultures. Britain, for example, with its population of Caribbeans, Indians, and the like, has given some time to exploring the notion of group privacy, a social ethic, if you will. It is a pervasive problem, but perhaps it ought not be prematurely incorporated in such things as these laws. We know too little about the matter to do so, and the laws are complicated as it is. It does seem to me there is a real political imbalance here: all the sophistication is on the side of data gatherers rather than on the side of the individual who provide the information. This possible inequity is a chronic and pervasive problem, it does seem to be a legitimate intellectual problem, it does deserve serious intellectual attention even if it cannot be tackled under the rubric of what we are talking about here.

Question:

I would like to ask a hypothetical question. A certain Federal agency lets a contract with a private survey company and in order to check up on the company's work asks it to supply a list of the respondents in the survey; the Federal agency proposes to call the respondent and ask them if they have been contacted. What is the privacy and confidentiality aspects of this? Is it legitimate or are there problems with it?

Tom Jabine:

This sort of thing does happen, not only in government surveys but in private surveys for market research and other purposes. Very often when the surveyor explains to the respondent what is going to happen, the surveyor will say my supervisor may call you in a day or two just to find out if I have been here. I think that is in keeping with the notification principles that are generally accepted as desirable in surveys, i.e., that you tell people what will be the consequences of agreeing to take part in the survey.

Bob Boruch:

There is some evidence which may be helpful here. It concerns the U.S. General Accounting Office's interest in reinterviewing participants in the Experimental Housing Allowance Program (EHAP). As you may know, EHAP was run by Abt Associates and Rand, under contract with the Department of Housing and Urban Development. These agencies resisted GAO's request for a list of participants and for data, under the argument that participants were assured of confidentiality. GAO persisted under the argument that GAO is mandated to oversee quality of such research under current law. The resolution was to have the original investigators elicit from participants' written consent to be reinterviewed by GAO. For the three experimental sites, 65-85% of participants gave permission and were reinterviewed. This cooperation rate evidently did not satisfy GAO; frankly, I'm surprised that cooperation rate was so high. Other tactics for reconciling conflicting interests in this context are reported in a monograph produced by the Social Science Research Council's Committee on Program Evaluation, and in the monograph by Boruch and Cecil cited earlier.

Joe Duncan:

Let me just make two quick comments. First of all, the legislation I described does have a provision in it explicitly for performing an audit of individual records. Let me give you another brainteaser. In the collection of price data, one of the issues is how do you develop a sampling frame. The BLS has developed relationships with several associations representing different industries, hospitals, universities, and so forth, to help collect data. One of the industries became unhappy with the numbers in the Consumer Price Index affecting their industry. So under the Freedom of Information Act they filed for a list of the sample that were used. (Incidently, it was drawn from a complete industry list they had provided.) My Office took the position that they could not have it, that this was statistical information that was confidential. The reasoning was really quite simple: if we made a practice of giving out sample list on every survey, we would certainly increase reporting and harassment and people would stop filling out forms for anyone. But in the individual case,

this is not terribly unlike the audit which we are speaking of because in effect this association wanted to audit their own members. It becomes a little of a brainteaser in that it demonstrates the difficulties one gets into when you push these things to the limits.

Ida Merriam:

One point that might be worth repeating is that the research bill applies only to individuals, whereas the statistics bill covers also corporations and other legal entities.