Proceedings of a symposium on

PERSONAL INTEGRITY AND THE NEED FOR DATA IN THE SOCIAL SCIENCES

held at Hässelby Slott, Stockholm,
March 15–17, 1976
and sponsored by the Swedish Council
for Social Science Research

Swedish Council for Social Science Research

Proceedings of a symposium on

PERSONAL INTEGRITY AND THE NEED FOR DATA IN THE SOCIAL SCIENCES

held at Hässelby Slott, Stockholm,
March 15–17, 1976
and sponsored by the Swedish Council
for Social Science Research

Editors of Proceedings:
Tore Dalenius, Anders Klevmarken

ISBN 91-970044-6-4

CONTENTS

EDITORIAL PREFACE

The papers which form the present volume are the proceedings of the symposium "Personal Integrity and the Need for Data in the Social Sciences" held on March 15-17, 1976 at Hässelby Slott, Stockholm under the auspices of the Swedish Council for Social Science Research.

The participants in the symposium, as may be seen in part from the roster of people who presented papers and participated in discussions, brought widely varying perspectives to the theme of the symposium: attendants included representatives of various academic disciplines, especially the social sciences, and members of government statistics offices and of the Swedish Data Inspection Board. This variety naturally stimulated much interesting and enlightening discussion which, we believe, both brought about a better understanding of the different viewpoints presented and raised new questions of interest for future research.

The present proceedings reflect - except for the last day's discussion from the floor - the program of the symposium. We are grateful for the cooperation of the authors and discussants in preparing their contributions for the press, and we wish to thank Sten Bergman, New Haven, Connecticut and Anna Landberg, Uppsala for translating several manuscripts into English, Katrina H. Avery, Providence, Rhode Island, for editorial work on some of the manuscripts, and Wera Nyrén and Gudrun Dahlberg, Stockholm for editing and typing the volume for reproduction.

Launching a symposium such as this is a major managerial challenge. We want to express our gratitude to Olivier Guilbaud, Stockholm for his untiring efforts in this, and also for his assistance in producing this volume.

Finally, we want to thank the Swedish Council for Social Science Research for providing the funds for the symposium and for these proceedings.

September 30, 1976

Tore Dalenius
Providence, R.I.

Anders Klevmarken
Göteborg

1. INTRODUCTION

by

Torgny Segerstedt

About fifteen years ago I had a standard question which I asked when I met colleagues from abroad: what do you regard as the most important event in your field of research? The overwhelming majority answered: the new possibility of data analysis with the help of computers.

It is evident that computers and the development of electronics in general have given all scientists and especially social scientists a new opportunity to collect and analyze data. In older days we had to be careful in our data-collecting - we could easily be drowned in the stream of answers to questions and of statistics. But with computers there seem to be no limits. Our optimism, however, turned out to be ill-founded. We soon discovered that our technical and scientific difficulties were replaced by ethical problems. It was asked: How much does a researcher have a right to know about another human being? Is there a limit which must not be overstepped becaused of a human being's right to personal integrity? If we admit that there is a limit, we will probably have to understand that there are areas and relationships in individual and social life which will never be available to sociology and psychology and other social sciences. We may regard that as something good, as a blessing, but we cannot be quite sure that knowledge of those facts and relations might not have been of great importance in solving problems which cause considerable human frustration and suffering.

The problem of the relation between science and personal integrity is not a simple one. That was the reason why the Swedish Council for Social Science Research felt it of great importance to arrange this symposium. We want to thank those who organized it for their work and all participants for their most valuable contributions.

2. Theme No. 1: WHAT REQUIREMENTS FOR PROTECTION OF INDIVIDUAL DATA ARE REASONABLE?

2.1 THE DATA INSPECTION BOARD AND RESEARCH

by

Claes-Göran Källner

The background to this symposium is the criticism of the Data Act that has come from researchers, statisticians and social planners. From that source we have heard that the Data Act restricts free research. It has been said that the Data Act and the Data Inspection Board should not become an instrument of censorship which hinders criticism and independent research. It has also been said that a statistical base must be made available, otherwise reform policies and welfare policies cannot be constructed and this would hurt the weakest individuals in society.

But also other voices can be heard. Thus it is stated in Folkpartiets data policy program, presented in August 1975:

> "Foremost the authorities, but also researchers ought to a higher degree be forced to motivate investigations which require that sensitive personal information be collected. They should be able to show that not only does a specific investigation has a reasonable objective, but also that the objective is actually achieved by the research, that the objective cannot be achieved by other means, and that the investigation does not come into conflict with the individual's right to privacy. An ethical code ought to be established for investigations that imply that sensitive information is to be collected about individuals.
>
> As soon as possible after the processing of statistical data the individual's identifying number ought to be removed. One should not collect a set of sensitive pieces of information about an individual solely on the motivation that it 'could be good to have'. Files on persons must be periodically thinned out".
>
> In a motion by the Moderats in this year's Parliament it reads:
>
> "The researchers have reacted strongly when criticism has been directed against their manner of filing so called soft data. From the researchers' point of view it is put forth that every restriction on the use of the ADB technique in research implies a kind of censorship. 'The freedom in research' must not be restricted. Yet it is apparent that researchers must rethink their position. Citizens can first be expected to allow themselves to become guinea pigs for researchers if not some minimal demands are fulfilled regarding the protection of their privacy."

The problems have become more acute because ADB has created a new situation for research, statistics and social planning. Simultaneously as being able to broaden and deepen the information in records kept for these purposes one has been able to keep the connection to the individuals. It has become practically possible to follow groups of individuals which can be identified decennium after decennium and continously add

new information about the individual. These longitudinal investigations may come to complete to a wider extent the hereto more common cross sectional investigations, i.e. investigations of groups of people selected after the same principals, but which contain different sets of individuals at each investigation period. The longitudinal investigations are used to register every change of an individual with respect to some area of interest.

As is so often the case, here one has to weight different interests against each other.

Not even the most ardent supporter of the independence and freedom of research suggests that the researcher, with their high moral values, has the right to stand above the general rules which govern the interaction of people and certainly the laws which regulate this interaction. The sometimes lofty goals of research does not sanction all methods which can possibly be used. Without putting a cloud over Swedish researches, it should be permitted partly to point at the experiments conducted on people which have taken place occasionally, and partly to remind us that the experimental results which first seemed to be steps forward have turned out to be the opposite: the preparation of DDT resulted in a Nobel Prize - but now it is an illegal environmental poison. Let us ask the following question: can research withdraw from its social responsibility through confining itself to make and account for new discoveries, or ought the researcher also be prepared to take responsibility for the way in which his results would come to be used?

Thus, it must be an obvious starting point that research, as well as other activities in society, should be conducted with considerable respect for the private life and value as a person of the individual. But who should then decide where the boundary is to be drawn?

When it concerns research one can demand that the researchers themselves take the responsibility for setting the boundary: they have the most insight due to their education and wide perspective. Within medicine one has for a long time realized this and has there developed an ethic which ought to be seen as an expression for this responsibility.

In Sweden one can say that this responsibility lies to a great deal on researchers, because of their as of yet so privileged position in comparison with their colleagues abroad.

In the first place, the Swedish public documents rule, i.e. the in principle free access to documents and information possessed by the authorities, brings about unique research opportunities.

Secondly, our rules imply that the authorities can generously give out for objectives of research also such information that is not to be made public.

Thirdly, the existence of an identity number for each individual, which is never to be kept secret, together with the public documents principle, makes possible research projects that could not be conducted without such identity numbers. (In this context, it may be worthwhile to remind ourselves that when in foreign countries it is suggested that research and statistics should be exempted from legislation, it is a question of personally non-identifiable files, i.e. it concerns de-identified information.)

Researchers already have according to the present state of affairs a formal responsibility in this area. It can be worthwhile to remind ourselves of this more formal side of the situation.

A research institute is an official authority. This implies a responsibility to see to that the current freedom of the press rules is observed, that secrecy and the duty of silence are maintained where appropriate. In practice this implies a responsibility to make sure that classified documents and information, that is to say also computer-stored information, will not be handled carelessly or get lost. This requires in part knowledge about the current rules and also in part a certain surveyance and control, tidiness and order. The responsibility lies on the first hand with the institution's director or the corresponding holder of office. They must, for example, take the responsibility for seeing to that their students do not - as is said to have happened - use computer bank data for experimental jokes. From the health sector I've heard it being mentioned that a certain, not so low percentage of all journals are always unaccounted for, which is, of course, natural, but also that a certain percentage of these are said to never be found again.

Naturally, within the area of human research one hasn't overlooked one's moral responsibility. The ethics developed since long ago in the medical field should probably be seen as an expression of this assumption of responsibility.

Within other areas of human research the ethical questions have, in contrast, not received any attention until the last few years: Among the causes of this is the obvious one that the consequence of lack of ethics within for example the behavioral sciences is hardly concerned with life or death or even physical harm. Here it is question of the citizens' personal integrity and psychic welfare. In this period an explosive development has taken place within the behavioral sciences. According to a report (1974:2) from the Department of psychology at the University of Stockholm concerning ethical problems in psychological and pedagogical research it is noted that the number of citizens partaking in pedagogical, psychological and sociological experiments, field studies and inquiries has grown very rapidly. In the report it says among other things:

> "Also in Sweden the extent in terms of the number of participants in behavioral science research is impressive. The investigations which for the period 1969-1973 are included in the register of reports on research in the behavioral sciences by the Swedish Council for Social Science Research, have engaged between 40,000 and 70,000 persons as experimental subjects. In an endeavor of such magnitude it is apparent that there exist risks for accidents of the form of going over the boundaries of accepted ethics and efforts should be made to prevent this ... .
>
> Another important reason for the increased interest for ethics in research is possibly that there exists a greater awareness of the relationship between research and the wider social context in which that research occurs. The interest from the public for different aspects of research has thereby increased. Finally, research has been faced with new issues and methods which in turn have given rise to new kinds of ethical problems".

The new situation for conducting research, which ADB among other things has helped to create, has been accompanied by various proposals

formulated by the committee for ethical questions of the Swedish Council for Social Science Research concerning ethical principles in research for psychological-pedagogical human research. Their results contain the following rules:

1) The experimental subject, and when the subject is a minor even parents or guardians, ought to be informed in advance of all the moments in the investigation which might possibly affect their willingness to participate.

2) After the information referred to above has been given, formal consent should be received from the participant. The circumstances under which this occurs should be of the kind which permit true freedom of choice. For minors, consent should also be received from the parents or guardians.

3) The participant should be informed ahead of time about his right to terminate his participation.

4) All personal information should be gathered, registered and stored under strict confidentiality. The reporting of results should be done in such a manner that it is impossible to identify the individual participant.

5) It is the obligation of the researchers in as far as is possible to inform the participants of the direct results of the investigation.

The above recommendations are not a consequence of the Data Act. Similar proposals have been presented in other countries which do not have or did not have at the time of the above proposal a Data Act.

The Swedish proposal coincides with the ethical guidelines for research involving human subjects that was published by the American Psychological Association 1973. It also corresponds to the proposals that have been discussed seriously in other countries. For example the British Association for the Advancement of Science in a June 1974 report called "Does Research Threaten Privacy or Does Privacy Threaten Research" proposed ethical rules that are similar to the Swedish proposal. The English report contains among other things recommendations that a certain type of research should be postponed until society is in a position to use the results in an intelligent way. Additional examples can be taken from the debate in Germany. There it has been suggested that the administratively collected data and research data never should be mixed together. Researchers should use unidentifiable data which cannot be misused against an individual person. By exception, administrative data can be used within research only if it is first passed through a rigorous anonymification procedure.

As always it is easy to speak in general terms about a boundary between the freedom of research and an appropriate respect for privacy and the value of a human being. The difficulties become apparent when turning to concrete reality.

I have in the above tried to suggest that it is the researchers themselves who are most able to judge. However, when the question pertains to research done with the help of ADB the state authorities have assigned this problem to the Data Inspection Board. The Board then naturally has to - as would the researchers themselves - find guidance in the legislation at hand.

From researchers it is said that the word research does not even appear in the Data Act or in the directions and guidelines for the Data Inspection Board. It has also been pointed out that the Board never really investigated the special position in which researchers find themselves, their need and their role as both a critical and "innovative" force. It has to be conceded that the Data Act does not give much direction, but at least gives some.

In the first place, the Act gives a guiding main principle concerning the registering of sensitive information, that is to say information about crimes committed, health care received, state of health, sicknesses had as well as political and religious convictions; it is exactly information of this kind which to a large degree is of interest for research. Such information may be registered with ADB

a) by authorities who, according to the law, are directed to register such information; this is usually not the case for researchers. In counterexample we have the National Central Bureau of Statistics,

b) in other cases only if there exist exceptional and particular reasons for such registration. (This is a common legal loophole.)

As a help, in the motivating preface to the Act it expressively says that "exceptions ought to be made for statistical analysis and scientific investigations, under the precondition that the individual subject's name is not made public". In addition it reads: "Even in certain other cases an exception should be made if there exists a strong public or similar interest that the registration takes place and if there are guarantees that the information is not used or distributed in such a manner that unacceptable breaches of privacy occur."

Furthermore the Data Inspection Board is explicitly directed in the Data Act to consider the disposition of those who are subject to registration.

From the above the Board in applying its authority has come to the following conclusions. Registration for the purposes of research is in practise by no means generally exempted from the ruling principles of the Act. Every proposal for registration for the purposes of research must be reviewed separately.

A difficult question has sometimes therefore been whether or not the purpose of a proposed scientific investigation and its scientific value should at all be judged. The following example should be mentioned.

A psychiatric clinic wished to establish a register on 143 patients. The register was to contain all kinds of sensitive information, for which registration by ADB legally requires exceptional or particular reasons. The scientific council of the Swedish Board of Health and Welfare claimed that the investigation had three weak points: it was retrospective, it contained only a small and select group of patients, and the investigational method was strongly influenced by the doctor in charge of all the investigations during the period of data collection. Furthermore, the results of the investigation were said to come to be of only limited value; the possibilities for drawing general conclusions from the results were judged to be small. In spite of this, the scientific council and the Swedish Board of Health and Welfare approved of the investigation.

The Data Inspection Board permitted ADB registration in spite of the apparent limited scientific value of the investigation. This demonstrates that the Board is extremely careful with such judgements: What the results would have been if the Swedish Board of Health and Welfare and the scientific council had ruled against the investigation is difficult to say. A fact is, nevertheless, that the Data Inspection Board, which by now has ruled on quite a few cases concerning the establishment of register for research purposes, has so far never ruled against a single one.

The Data Inspection Board has, however, in granting permission, given directions concerning the protection of privacy. According to the Data Act the Board is required to give such directions when granting permission. A point of departure for this has been, as already mentioned, that the Data Act requires the Board to take into consideration the attitude of the subject. In particular this requires that the subject is informed that he will be registered and in what way.

Occasionally one can notice an attitude of those who are to organize a registration which is to the effect "if we don't go into the details of how the registration is to proceed we won't worry people unnecessarily". Such an attitude cannot be accepted. Beyond that it goes against the spirit of the Act, it implicitly declares the subjects incompetent who otherwise are acknowledged such rights of citizenship that have information and personal judgement as their natural prerequisites.

Against this background it is interesting to note that the previously mentioned ethical rules contain the right to information; according to certain sources the National Board of Education is said to be going to apply these rules within pedagogical research. It should also be mentioned that the National Board of Health and Welfare in its statement of approval of the above mentioned issue is granting that the persons concerned be informed before the investigation about its purpose, together with the extent and content of the files to be created.

Information to those who are to be registered and the consideration of their attitude obviously depends on the existing rules concerning coercion and consent within this area.

The main principle is that citizen can be required to give information without the rule of law. This is true, for example, where people are requested to fill in questionnaires or answer a set of questions.

Sometimes there is an implicit form of coercion in question answering even though the right to refuse to answer is formally present: In order to receive social help from public funds one is required to give information about one's situation; in order to get adequate health care one has to answer the doctor's questions - and besides he is obligated to register the answers in a journal, for health purposes.

In order to illuminate the difficulties that the Data Inspection Board may come across here and the solutions they have so far tried, one must first keep in mind that the ADB-register, which is used in research, can be of three different types as regards source material:

1) They may consist only of answers on polls, obtained through interviews or from forms which the subject has filled in himself.

2) They may consist of answers on polls together with information gathered from other already existing registers.

3) They may consist only of information gathered from other already existing registers.

In parenthesis one ought to mention concerning the gathering of data from already existing registers, that the Data Inspection Board in principle has nothing against that information which has been collected for administrative purposes, e.g. the rulings on petitions concerning individual persons, and which exists in the administrative registers, is used for statistical purposes, research, and planning. In contrast, it is a guiding principle that information gathered for statistical analysis or the like, cannot be used in determining a decision which concerns an individual.

For polls the Data Inspection Board has consistently required that those who answer the questionnaire be informed about

- the extent and purpose of the register,
- that the answers are to be used for ADB,
- that the participation is voluntary and that there is a right to refuse to answer certain questions.

Apparently this is consistent with the ethical rules.

From what has been said it follows that those who plan to arrange a poll must in advance have knowledge about the rules for the gathering of this information. The Data Act - or more correctly its motivation - considers this in the following way: The Minister of Justice has said with respect to this: "Belonging to the concept of a person register ought to be included among other things the gathering of personal data if the purpose is that they should be included in a person register."

Since the creation of a person register cannot be done without permission, it is illegal to start conducting a poll until one has received permission.

However, we are confronted with the transitional case where polls have been conducted before the enactment of the Data Act and where information hasn't been given to the participants to the extent that is now required. In such cases the Data Inspection Board hasn't requested information to be specially delivered in order to give permission to the continued establishment of the register. When in the future the person in charge of the registration applies for permission in order to extent a register with information concerning e.g. results from polls, still the demand on information will be asserted as far as this issue is concerned; the person in charge of the registration is still bound to come into contact with the subjects of registration and can thus deliver the information without too much effort. In other words, the principle of voluntary participation can thus be maintained retrospectively in these transitional cases. Consequently, as soon as the practical circumstances and costs so allow, the subjects of registration should obtain information about the moments of the investigation that are likely to influence their willingness to participate in it.

In the second case, when an investigation is based on both answers from polls and information from already existing registers, the Data Inspection Board maintains similar principles: the participant persons

should accordingly also be informed of the fact that information is being obtained from already existing registers and which these are. The giving of this information and the receiving of consent to participation do not bring about any extra work worth mentioning since the person in charge of the registration is in contact with the participants on the occasions used for interviews and the filling in of forms.

There are more complications when dealing with investigations merely based on information obtained from already existing registers. The advantage with this type of investigations is precisely that the subjects of registration don't have to be "troubled" and that the cost of the investigation is kept low. We haven't yet established any common-law suggesting how to deal with uses like this. One possibility is to let the Data Inspection Board presume in the ruling of permission that the persons concerned are prepared to participate provided it's a question of a cross-sectional investigation, and maybe also a following-up within a certain time limit of for example 5-years. As an alternative such a procedure could be complemented by informing the subjects of registration that the investigation is being conducted, and that they according to the tenth paragraph in the Data Act have the right to be given access to the register. In longitudinal investigations the ethical rules are in accordance with the completely voluntary participation of the persons concerned. Considering the ethical rules in research this voluntariness should also imply that the persons in question be informed of the long prospective period of investigation and of their participation through prohibiting the continued use of information about themselves.

To sum up:

1) The Data Inspection Board - and its committee of laymen - has contrary to what has been alleged, indeed "considered the special situation of researchers". But of course we have done so within the framework of the general rules prevalent in a society and within the framework of legislation set up by the state authorities to be applied by us. And we have been forced to do balancing of issues that we sometimes thought should properly have been done by the researchers themselves.

2) We have not prohibited any research records.

3) We have tried to maintain the requirement that information should be given those people being registered, as an act of defending their value as human beings and as a basis for the voluntariness to participate which in this field is a self-evident right of the citizen.

4) For some registers we have set up time limits. This isn't an expression of a general prejudice against so called longitudinal investigations but offers the possibility to make a new review after a couple of years.

5) We have tried to give directions about the deidentification of registers used for research, which doesn't necessarily hinder longitudinal investigations; we have actually judged deidentification to be of great importance for the confidence of the public in those who request information, among whom are the researchers.

To end up, it can be mentioned that the parliamentary committee who according to the decision of the parliament is to scrutinize the Data Act definitely will be given the task of dealing with the issue of the research records.

2.2 CONTRIBUTION TO THEME NO. 1

by

Kerstin Anér

Let me start this talk on a grass roots note. I think you should listen for a moment to what your subjects, or your objects or whatever you call them, actually think of you while you are studying them. Three stories about what the citizens of this country write to their M.P.s:

A student I know was requested by the Sociological Institute in Stockholm to answer a lot of questions of a rather intimate nature, in order to help with a certain piece of research: how did you like your parents when you were at school? did you drink and how much? did you use drugs? did you sell drugs to your friends? etc.

He did not answer. This meant he got several letters of reminder, finally in a rather menacing tone. At the fifth letter he answered at last, telling them: "There is a running number on this questionnaire you have sent me, and since I do not know at what stage in the research you will take away this running number, or whether you will take it away at all, I do not intend to answer. I cannot be certain that my answer cannot be identified as being given by me."

Then the research leader wrote back and said: "By this time, about 20 or 30 % of the people questioned have refused to answer, so now we have decided to scrap that running number from the beginning." The student then finally sent in his answer.

But both he and I wonder very much what happened to the running numbers of the people who did answer from the beginning. We also think it rather unfair to treat these two groups differently. We also wonder why they put in the running number at all, since the whole thing worked just as well without it.

I also cannot help remembering what the Swedish Association of Sociologists has written recently to the Data Inspection Board. They consider it quite unnecessary, the say, to make any suppositions about how the persons registered understand the purpose of the register. I quote: "An interview or questionnaire is generally answered after a brief explanation that this is for purposes of research. That the questions are answered is in itself a proof that the person registered accepts the purpose." I cannot quite see that this brief optimistic description really matches reality.

My second story:

A Swedish citizen recently boarded the ferry from Malmö to Copenhagen. Before he could do so, however, he was approached by a line of girls who said they came from the Central Bureau of Statistics (SCB), and asked him a lot of questions about his travel patterns and so on. They also asked for his name and address, so that they could follow up the questions some other time.

This citizen wrote to me and asked: "Do the girls have a legal right to demand answers?"It was not quite clear whether he meant: "Do they have a right to stand there and question us?" or: "Can we be sentenced to a fine of 1,000 crowns if we refuse to answer?" The point is that to most Swedes, the Central Bureau of Statistics means only one thing and that is the Census, which is run by this Bureau, where you can be fined if

you refuse to answer, and about which there has been a violent debate recently in the mass media. So something of the devilish aura of the Census hangs about anyone presenting themselves as coming from SCB. And it is perfectly clear that the girls had done nothing to dispel this aura, that is to say: inform the citizens that they were not legally compelled to answer.

I cannot help, in this context, remembering what the Director general of the SCB said in November 1974 in the debate about the right of his interviewers to be answered: "We need many statistics to plan our society, and this is only done for our own good. So anybody should have as much confidence in his interviewer as he has in his doctor or lawyer or social worker." I have kept this saying next to my heart, because it seems to me to express so very well how one does not feel towards an interviewer who attacks you, whom you have not asked for, whose professional ethics are absolutely unknown to you, and about whose use of the answers you have no idea.

My third story is simply a letter I got a few days ago, not from an ignorant layman but from a man who has worked with computer systems for many years. The letter is signed but I will not read his name, because it is not necessary. He writes to the SCB and sends a copy to me: "I have received a demand to fill in the Census form, but after having studied the questions I feel unable to do so. I have myself been an active systems programmer for many years and so I know perfectly well what can be done in extracting, linking and re-structuring information. Considering the strong pressures that exist at the present moment to be allowed to open up or to link together many otherwise secret public registers, I cannot exclude the possibility that existing files on people may be linked together to make up new files with quite different qualities than the original ones.

Until real guarantees are given against this kind of misuse - and at the moment, only ways of forbidding it are being discussed, not ways of actually preventing it - the individual has no way of protecting himself except by refusing to deliver reliable data. This can be done either by deliberate falsification or by refusal to give any data at all. I prefer the latter course."

Well, since I was among those who made that Census law I will refrain from comments here. But I can tell you that when I called this man up and asked him: "What do you suggest we do?" he said: "Why can't the SCB mix in 5% noise with all administrative data, when they are used outside statistical purposes and when this means giving out data from interviews?" This is a question I would be very happy to leave in your laps, and I hope you will discuss it.

It is perfectly clear, as Mr. Källner has said here today, that the people who have complained about the Swedish Data Act have been almost exclusively social scientists. That is why it is so very important that a debate like the present is being taken up. I would add: it is very important that you statisticians and social scientists do not conceive this as a quarrel only between yourselves and the Data Inspection Board, but that you see it as a three-way problem. I mean: that you look on the people you interview as your collaborators. Not only should you develop a very fine ethical sensibility for how you treat them, but you should develop proper methods for working in collaboration with them. This is to me the point of what you are discussing today - at least it is the point of what I am saying.

Let me quote here a little more from that British group of researchers that Mr. Källner took up in his talk. They seem to me to say things that are very relevant in all countries and not just in the U.K. They stressed very strongly the fact that research is never neutral. It is very often used to make it easier for one group of people to control another group. In the U.K., the most burning problem is immigrants - this is more important to them than to us.

I will illustrate this by just one of their stories - you may not all have read their report - which was about immigrant women in maternity wards, in Birmingham I believe it was. The study in question proved that immigrant women stayed in maternity wards, on the average, one day longer than white women. This bit of research was then used by some people to prove how lazy and feckless these women were, and how unfairly they took up beds which could have been used by others. The real reason was - as anyone could understand - that the immigrant women had such very bad and unsanitary homes to go back to that the doctors could not send them back as soon as others.

The British research group also mentioned another instance of misuse of social studies which has had its counterpart in Sweden too. I mean prospective studies of potential child batterers. The point of these studies is that they always refer only to the poorest part of the population. Nobody looks for child batterers higher up, and yet they certainly exist there too. But the very fact that a lot of information (most of it of negative character) is gathered about one part of the population but not about other parts means in itself a discrimination against the group that much is known about.

I would like to add another instance of this, which is the same in all countries. Every national bureau of statistics knows a lot about how many working mothers there are in the population, what hours they work, what age their infants or children are, etc. No statistics about working fathers have ever been published. This very fact does a great deal to steer the debate in one direction.

The British group of researchers could not agree on how to handle the problem of group integrity. They say: "If you give certain groups of people the right to forbid certain types of research about them, you naturally open the doors for perhaps self-appointed spokesmen for these groups to attack researchers for political reasons." This might stop all critical research. These fears have been voiced in Sweden too.

Others, however, rejoined: "It is not enough to give individual interviewees a theoretical right to refuse, because this kind of research is mostly done on poor and helpless people with very little political power. They dare not, in practice, say no and they do not know how to stand up for their right to do it." The group of researchers with this view added: "If you were to go up to the manager and ask him the same kind of questions you ask his black workers, he would either throw you out or call for his lawyer."

I would like to illustrate this point with something a Swedish doctor once said to me. He was engaged in a piece of research on a rather obscure disease, not a fatal or a disabling one but one which developed in a certain way during life. So he had gathered data about patients who had had this condition diagnosed at one time of life, and now he wanted to look at them after fifteen years and see what had happened, just in the interest of research. The difficulty was not in finding the people,

but making them come to the doctor for an examination which did nothing for them, only for the young doctor's thesis. He told me: "There is never any trouble with the lower and the middle classes, they always come when you ask them. But the upper classes, they are impossible!" In Swedish this was expressed as "social group III and II", for lower and middle class, and "social group 0,5" for the very top of the top.

I wonder whether you recognize this situation? It is far easier to manipulate people who are used to being frightened of authorities and who obey researchers as if they were part of the same establishment, than to make those people obey who feel they are the establishment themselves.

The British group finally came to the conclusion that the people on whom research was being done should be given far greater real opportunities to discuss and understand the project, including what the results were to be used for. Let me illustrate this with what Margaret Mead said some years ago, at a conference of the AMA on ethical aspects on experiments with patients. Her experience as an anthropologist, she said, was that unless you took the tribesmen into your confidence and told them why you were interested in their customs, you simply made very bad science. Because then the villagers would not tell you the truth, which they did not think you deserved. It was not difficult at all, Dr. Mead pointed out, to make even quite uncivilized people understand that you wanted to study them, and what for. I think this is a story with an important moral and has much wider applications than anthropology and medicine - where the patients should be treated with exactly the same confidence and courtesy as Dr. Mead's black friends.

What one may very well ask in the Swedish context is now: "Is the Data Inspection Board really the best forum for looking after the interests of the interviewees?" The answer to that question is not quite clear. I am perfectly willing to admit this, although I helped to bring the Swedish Data Act into existence, because this is a problem which was never broached either in preparing or in debating the Data Act.

I think it is obvious that we, and you, should now try to create some kind of mechanism for this collaboration between researchers and the people they do their research on. I am quite certain that the Data Inspection Board is not very happy to have to decide these questions, and one could imagine other organs or institutions who might be better fitted for it.

I must, however, stress the fact that I am not thinking of some kind of ethical committee who would sit in judgement from above and decide: "Is this a breach of integrity, is that not?" What I am talking about is precisely a collaborative organ, where you try to find ways really to be able to listen to the prospective interviewees. I know this is difficult, in many cases perhaps impossible. But I do think we should not assume from the beginning that it is impossible. This new way of studying people with computer statistics is such a sharp and heavy weapon that we must really find some new ways of handling it, in order not to cut off somebody's finger with it.

Now I should like to say a few words about this integrity, which no one ever defines. In the old League of Nations there was once a Committee which sat for ten years to debate measures against pornography. It never arrived at a definition of what is indecent. I understand them - we all do - and the definition of integrity is in somewhat the same plight.

My contention, however, is this: we find ourselves in a deadlock as long as we accept a "concentric model" of integrity. This means a mental picture in which we have the individual in the middle of a large circle. Closest to him, there is a very small circle which encloses his integrity. Within that space, he is protected; as soon as he steps out of it, the lights play on him and he is totally visible. The question then becomes: where precisely do we draw this inner circle?

Now some German lawyers and sociologists have done some interesting theoretical work on this and said: with this model, you actually make the citizen politically unfree, because you decide from without where the space of his free actions ends. (Anyone who wants to read this in the original should look up: Numerierte Bürger, ed. Hoffmann - Tietze - Podlech, Peter Hammer Verlag, Wuppertal 1975, page 121: Paul Müller, "Einige soziale Auswirkungen integrierter Information-systeme".)

The opposite, and I think better, model is the sectoral one. This means to divide up the large circle, which means society, into a lot of different sectors, each representing one particular part or facet of society and more particularly the field where the citizen encounters society in one role. In every different sector, it is in the interest of the citizen to be treated with the help of certain bits of information about him, but not all the bits that are theoretically available. He wants to be treated differently according to whether he is at the moment acting as tax-payer, father, voter, schoolchild, prisoner, worker, patient, client, customer, etc. etc. In every specific role he wants to be identified by certain data, but he does not necessarily want to be identified by the same data in the next role.

A functional definition of integrity would thus be: the right to decide what information about oneself one wishes to release in every particular situation. To call one role more "private" than another is to obscure the issue. As citizens, we want control over <u>all</u> our roles.

I may have to point out that I am not intending as an ideal that society should no longer have any right to any information at all against the will of the individual. In every sector, there must of course be a different mix between the interests of the particular citizen and those of his co-citizens, as represented by the authorities. What I am putting forward is that this sectoral model does make it easier to discuss how to weigh the different interests against each other.

The interest of society in general is of course to take the whole circle at once and leave no sectors unoccupied. This makes administration much easier, not to mention research. It would, however, mean that the individual would become at the same time totally visible in space and totally immobile in time (= locked into the attitudes he once had). This cannot be to the common good, if only for the sake of the psychological effects.

Some people will object: but nothing but good can come of a state of things where everybody is forced to be completely honest and consistent! Well, in a society of ideal people this would be ideal, but we do not have that kind of society.

In certain sectors any society needs a very high degree of visibility - income tax returns, for instance. As you know, the Swedish state has recently given itself very far-reaching powers to find out exactly how much money we all had in the bank at any time. The opposite example

is that of the recently discharged criminal. In his case, we do think that he has a right not to have his whole past presented to his employers and co-workers at every stage. We have even passed laws to make this kind of privacy possible. Between these two extremes there is a whole scale of different situations, and I am not going to tell you where I think the line should be drawn on this scale, because that is precisely what I think must be decided in confrontations of all parties.

Generally, one discusses integrity and privacy in these terms: do you think this or this detail is such that it should not be public? What about sexual habits, or religious affiliation, or party affiliation, or mental illnesses, or illegitimate children? I think this is defining the problem in the wrong way. I think we should ask: what mechanisms can we introduce in society to make it possible for citizens to state their own views effectively, not in general but as it applies to themselves? This is much more difficult but also more fruitful.

I noticed in reading some of the papers given out in advance at this conference that one of them (P. Reynolds: On the Protection of Human Subjects and Social Sciences; International Social Science Journal, 1972, pp. 693-719) talks about how to protect the individual against social scientists but never mentions the idea of cooperating with the people he interviews. This astonished me very much. Professor Dalenius, on the other hand, does mention this. He says it is very important that statisticians win the confidence of the people. I agree absolutely with this, but I would add: it is not just a question of confidence in the utility of statistics in general, but in the utility of the special kind of research directed at oneself or the group one belongs to. We must not restrict this debate to how the interviewee feels about the interview situation in itself. There was a certain tendency to this in some of the papers I read. Or else the debate was about the fear that particular items in the answers might be leaked to unauthorized persons. These things are important, of course, and the Data Inspection Board very rightly considers them and decides about them. But it is far more important to know: how will this bit of research affect the lives of the people interviewed? What will be its political relevance? And obviously this means, for one thing, that people should be perfectly free to refuse the answer to all and any questions.

Finally, there is the very vexed subject of imputed answers, and in general of how administrative data should be coupled with interviewed data. What we have to remember here is that data given by the citizen to one particular authority in one particular situation and for one particular reason may not be relevant or useful at all when used by another authority for quite a different purpose.

It is, I believe, a good rule among statisticians that only that organization which originally gathered the data really knows what they mean and what they are worth. The quality of data often declines abruptly when they are used in a different context. This affects the integrity of the citizen, and also the usefulness of the data to the authority who uses them.

For my part I think very little indeed has been done to insure that data are used only in contexts where they are meaningful. I would propose, from the point of view of the citizen and as a preparatory measure, that whenever the citizen is asked by the State to give any kind of personal data, he should be informed on the same piece of paper of all the other State authorities to which this datum will be routinely sent. As

we know, even the data a patient gives to his doctor may in different circumstances be re-routed to no less than 21 different authorities (not all of them at once, of course). The patient generally has no idea of this.

What I have tried to explain is that integrity is not a two-pole problem between the individual and Big Brother. It is a field with at least three poles. The problem could best be stated thus: to whom, among many different authorities and organizations, do I wish to give the advantage of knowing these particular data about me - or about my group? (I have talked a lot about groups here, but that is because group integrity is very important and very neglected, not the least in connection with statistics.) These safeguards will be expensive. The question is: how much is a particular type of knowledge, including the right to collate it with other types of knowledge, worth to the State or whoever gathers it? How much is it worth when you have to pay for it by taking into account the rights of the citizens to <u>their</u> data? This is the kind of budget we must very soon make up, because information capitalism is no less in need of control from below than any other kind of concentrated power.

2.3 INDIVIDUAL WAGE STATISTICS AND EMPLOYERS' OBLIGATION TO RELEASE INFORMATION FROM AN INTEGRITY POINT OF VIEW

by

Karl-Olof Faxén

It is not problems originating in research which form the background of the emergence of the Swedish Data Act and the definition of integrity which has been associated with this Act. Among other things, a research project can never lead to measures against individuals, only possibly against groups to which they belong. The question of integrity in connection with research is different from the corresponding question in connection with permanent all-inclusive files, even though these problems are related, as I will discuss later.

Actually, we do'nt know very much about the concept of integrity, especially if we also consider the questions of integrity in research. From that point of view the Data Act is a temporary arrangement. On the other hand, it is my opinion that the Data Act as well as the Data Inspection Board are necessary and that all of us, researchers as well as ordinary people, must be thankful for the protection it gives. But it is obvious that large government files constitute a threat. The utilization of these files must be controlled by some authority looking after the integrity of the individual, and the Data Inspection Board has an important duty to fulfill.

In regard to research, the interest in being able to carry out large statistical investigations, even longitudinal ones, on the basis of personal data, must be balanced against the demand of the individual for protection of his integrity. It is impossible to draw absolute boundaries. Our knowledge of the meaning of integrity is incomplete. What is considered justified protection will certainly vary from time to time. It also varies among countries and perhaps also among different groups of people within countries.

Nonetheless, the demand for personal integrity is a legitimate and respectable one, for which I think the researcher must show understanding. Here is a limitation of the possibilities to do research which the researcher must regard as a fact and as a starting point for his activity. He must respect the fact that people hold the values they do, think what they think and feel what they feel. He does not have the right to violate the rights of his research objects by going behind their backs and collecting information from existing files on individuals in such a way that this can be regarded as a threat.

There is a connection between integrity in research and in utilizing permanent files for administrative purposes. Research can sometimes point out new ways to utilize administrative files. Let me illustrate with a constructed example.

It is fully possible technically to carry out registration of automobile traffic in order to obtain a register of the movements of all automobiles in the country. Such a register could be of great interest, e.g. for taxation of automobiles. It could also have a research interest by facilitating closer examination of road traffic and, e.g., detailed speed registration.

The potentials for using a register of this sort for other purposes than traffic research, e.g. to find out how people have moved in order to obtain information on the habits of life of individuals, what social contacts they have, etc., are obvious. It is just as obvious that no political authority which becomes conscious of the problem would permit anything of the sort in our country today. However, that does not solve the problem from the individual citizen's point of view. He still feels a threat to his integrity (although not in the sense used by the Data Act), partly because he does not feel completely sure that no one in the administration would use such a system anyhow to study habits of life - without the political authorities being aware of this - partly because in a future political situation of which we do not know very much such a system will be used for such personal supervision. Finally, the technical possibilities themselves can gradually change the political values so that registration of this sort will be considered politically acceptable if it can be justified strongly enough, for example by the need to create a better data base for the fight against traffic accidents or against narcotics trade.

In such a situation registration of this sort becomes a threat against integrity from several points of view, even if it is only a case of a single research project. Thus, even if it is only a temporary investigation during a limited time period, and even if all identifications are removed and all possibilities of going back to a single individual are extinguished, the research project can be regarded as a kind of test of a future administrative system which one does not want on the basis of one's political values. Then the research project itself also becomes a threat.

No one today can imagine that the state would use a register of automobile traffic to systematize and store information on the social contacts of citizens. We would all react just as violently against this as against general wire tapping, for example. On the other hand, such studies of social contacts could be interesting from a scientific point of view. How far can a research project go in this direction?

The difficulty with this is that it is not possible to create protection through non-disclosure rules or by removing identification codes after the investigation is finished, or the like. The crucial factor is the knowledge of the methodology itself, the knowledge that the methodology exists and can be used in a different situation in ways which we cannot imagine today and which are forcefully rejected by all.

The actual registration of personal data increases rapidly as more and more registers are being used by different branches of public administration. In addition, more and more people participate annually in questionnaire inquiries of various kinds. It is a matter of 50,000 people or more. It is understandable if this creates fertile ground for psychological reactions. I do not know what research may have been carried out concerning this particular aspect, that is how specific the reasons for worry are in the data field, to what extent this worry is associated with more precise notions of what constitutes risks for individual integrity, and to what extent it is a question of a more general lack of faith in the environment and in the society which is expressed in this way.

I admit that it is not particularly easy to investigate this type of socio-psychological phenomenon, but it is likely that a research effort could contribute to giving us a clearer view of what the problems are.

What do we really mean by the integrity of the individual in the situation implied by participation in an investigation? Obviously there is a limit both to what kinds of questions can be permitted from the point of view of protection of integrity, and to the possibilities of linking together various registers. The limits are different for research projects than for permanent social information systems. For example, it is possible to go further with questions concerning alcohol habits in a research project than we would accept recording in a register of alcohol consumption used for administrative purposes. But who is to draw this line, and what should the procedure be?

Here we have the work in progress within the Council for Social Science Research on a code of ethics, and we have a proposal from the Data Inspection Board. I will not go into these in detail but will limit myself to saying that the question of voluntariness in practice is difficult. A person being interviewed is in an inferior position in many ways, and very few people can be assumed to understand the implications of later treatment of the answers as such and in combination with information from administrative files. Minorities in the population being investigated might react strongly, even if the investigation presents no problems of integrity for 90, 95 or 98 per cent of the population.

I have no solution, but I assume that the purpose of the discussion at this conference is to deepen the analysis of the problems. Actually, the investigators themselves should be the ones who can best realize the importance of protecting personal integrity.

The investigator must be able to control his instruments of measurement in a different way from, e.g., users of administrative systems. But like everyone ought to respect the data legislation and the instructions of the Data Inspection Board until the time comes when we know more and time is ripe for a revision, so must the research workers - not least for the sake of research itself. It is necessary that a debate be started, but it is unfortunate if it takes place through a confrontation between research workers and the Data Inspection Board. The need for balance increases through the increasing flood of information on individuals.

As is apparent from what I have said, investigators themselves must have a decisive influence on the balancing procedure. The working method of the Data Inspection Board, e.g. remitting proposals to various authorities, is not suitable in this connection.

Even if, among other things, the connections between research and the development of administrative methods of application justify the involvement of the Data Inspection Board, it is not reasonable that the overall balancing be carried out within this body when it is a question of an investigation. But I admit that it is a difficult question, and this is only an early contribution in this conference.

As an example let me touch upon our comment on the proposal remitted to us concerning the political science investigation of immigrants. What we put particular emphasis on was that among the interviewees there were emigrants from countries with dictatorships who were used to an entirely different type of political questioning from the authorities. Therefore, there must be many people who could be expected to regard the investigation as a kind of masked police investigation of the kind they experienced in their home countries. They could not be

expected to have sufficient familiarity with Swedish society to understand the premises upon which the questions were asked. Nor does the possibility to refuse to answer questions offer any protection to an immigrant in such a situation. We could not see that the research interest of putting political and religious questions to such a population, and linking the answers in an unspecified way to information from the population census and tax files, could justify the threat to integrity which we saw in this case. I still think that this investigation is a good example of what the Swedish Data Act is designed to prevent.

This was a special sample. What may be permissible in a normal population is a different question.

After this I will now proceed to treating the integrity questions as seen from the point of view of the Swedish Employers' Confederation (SAF). We have four functions in this matter, partly as producers of statistics together with trade unions and the State, partly as users of research, partly as spokesmen for the employers' own need of protection, and finally as indirect advocates of protection of information concerning employees.

Within SAF we produce wage statistics for blue-collar and white-collar workers. These statistics are individual and utilize the personal identification codes to about 80 per cent for wage earners and to 100 per cent for salaried personnel. They contain information on hours worked, wages and salaries, overtime pay, shift premiums, etc. They also contain information on occupation, employer, place of work, and the like. We do not regard these statistics in themselves as particularly threatening to integrity. The statistics are produced jointly with the trade unions and partly also jointly with the State. Thus, we do not have exclusive rights to use the material for scientific investigations.

SAF and its opponents have a relatively restrictive policy regarding participation in scientific investigations based on this material, when it comes to the protection of integrity. Thus, we make our own evaluation, in addition to that made by the Data Inspection Board.

As producers of statistics we must think not only of our direct suppliers of information, that is the employers, but also of the indirect suppliers, the employees. It is in the nature of the bargaining system that there must be as accurate wage statistics as possible. The best quality is obtained if the wage statistics are based on data for individuals. The personal information at our disposal is collected for specific purposes - for the production of the official wage statistics and for the negotiating statistics for both parties.

In addition to this there are various kinds of scientific investigations. We have collaborated in these by delivering information for investigations of wage formation, e.g. the relationships between earnings and education and between wages and firm failures. But this leads us into a more sensitive area. How far should one go in investigations studying the circumstances surrounding the termination of employment?

I am not sure that the problems are solved by formulating a principle of ethics saying that the persons asked for interviews should be informed that data on the development of their earnings are collected from SAF's files and that they should be given the possibility to refuse.

But let us continue this line of reasoning. How many other registers could be involved in an investigation of this type without chang-

ing the situation with respect to SAF's participation? Can we permit our data on individuals to be linked together with for example the social security files, files on alcoholics or on health care? The situation of our suppliers of information is so delicate that it is not sufficient, in my opinion, that permission is granted by the Data Inspection Board.

Nor am I sure the problem can be solved through some kind of power of attorney from the employee organizations. Of course, the employer can protect himself against attacks from labor organizations by laying down as a condition for his collaboration that the local trade union grant permission for certain treatment of the data. But how much of the real problems are solved this way?

An area of importance to integrity in the activity of employers which has not yet been brought out very much in the debate is the information given by the employers concerning employees in connection with preliminary tax withholdings, petitions for changed withholdings, impounding and seizure of salaries.

We are not happy that the employers are required to handle this material, but that is the way it is. Of course, this casts a shadow over labor relations and enters into the evaluation, e.g. with respect to using the employers' files on employees as a framework for samples in research of various kinds. It contributes to a very careful attitude.

We also have a group of individuals who are employers. In Sweden there are about 70,000 persons employing people in their businesses. Of these more than 30,000 are members of the Swedish Employers' Confederation.

Employers ought to have the same right to protection of their integrity as other individuals. However, the present version of the bill concerning employee participation in the decision making process requires the employers to go extremely far in supplying information to their trade union counterparts.

To be sure, in the introduction to the bill it is stated that "information concerning an employer's private matters which does not concern or influence the conduct of his business is ... outside the right of information". In order to protect the integrity of small employers operating as private firms, a more precise formulation is necessary. Otherwise, the far-reaching right to information concerning the firm can lead, in the first instance, to the local trade union obtaining such insight into the private matters of the owner that his integrity is violated.

But in many other respects, too, there must be protection of the integrity of employers which limits the possibilities of various kinds of scientific investigations. An example is the discussion which took place concerning large customers and large suppliers in connection with the county planning activity in 1974. In the beginning the intention was to require firms to supply the names of their most important business connections in order to facilitate an examination of the geographical network within and between counties. On the part of the business community we strongly opposed this idea, even though in this case it was a statistical investigation under adequate security safeguards. In our opinion, the information was too threatening to integrity. In the end, information on branch and postal zip codes for the respective cus-

tomers and suppliers was deemed sufficient. Thus, this was a case of providing protection against the requirement to give too detailed information on business contacts. It is a different matter that such information may be required in connection with bookkeeping and tax audits for a clearly stated and limited purpose.

I would guess that later today Lars Wohlin will take up the entire matter of the protection of the integrity of firms which are not physical persons and where it is therefore not really a question of protecting private life but rather of protecting the possibilities of operating businesses without close supervision of the authorities, the possibilities of making certain decisions within the firm and of making business contacts, etc. From the employers' point of view there is a clear link between these two problem areas. It is a matter of the possibilities of leading one's own life, operating one's business in his own way, and limiting the form of controls by the authorities which may be required and requiring each intervention to have a clearly stated purpose. The freedom of enterprise is a fundamental value in our society. General, unspecified registration, treatment and demands for information whose purpose one does not understand, and whose possible use in the future one does not understand, generate anxiety and insecurity. For example, why should the personal identification number of the employer be the identification code for his firm?

In closing I would like to summarize my talk in the following way: we at SAF must take a stand on these issues in our capacity as both producers of statistics, as representatives of research interests, and as advocates of the need for protecting the integrity of employers and employees. We have learned that the problems can be viewed from many different points of view, that the judgements may vary, and that it is often a matter of a delicate balance between various interests.

It is clear, however, that the problems of integrity are a reality and that it is understandable that the large accumulation of data on individuals during the last few years and the rapidly expanding technical possibilities of combining information from a large number of registers, of searching for various combinations of behavior from this collected material, opens up gruesome perspectives which can instill fear in us all. It is technically possible to collect information which we do not quite know how to master. This is true for both individuals and groups of individuals.

On the other hand, it is difficult to see how controls can be designed without threatening the freedom of research in the long run. The freedom of research is a fundamental value in our society, not only for the research people themselves. An organization such as SAF has a strong interest in the freedom to carry out research not being curtailed other than when this is absolutely necessary due to a conflict with other fundamental values.

The protection of integrity is such a fundamental value. In opposition to this there is the interest of the research worker to branch out into new questions and the interest of the society in moving the frontier of research forward.

3. Theme No. 2: WHAT ARE THE CONSEQUENCES OF DEMAND FOR DATA PROTECTION FOR SOCIAL SCIENCE RESEARCH? WHAT IS THE ROLE OF INDIVIDUAL DATA IN SOCIAL SCIENCE RESEARCH?

3.1 SOCIAL RESEARCH AND THE INDIVIDUAL'S RIGHT TO PERSONAL PRIVACY

by

Pär-Erik Back

We have earlier today in a number of statements received an especially interesting illustration of those problems implied by the title. Myself, I will deal with them from a certain viewpoint: let's say from the grass-roots level, since I intend to render an account of some concrete and basic experiences which have been made at the institution level at an university since 1973 - against the background of the changing conditions for social science research. Here we will be dealing with observations that can't be generalized. My attempts to obtain more systematic information about the consequences of the Data Act in different respects, still have not given results of sufficient scope.

In order to be clear I would, to begin with, like to emphasize a few other matters. Now, as before, I am of the opinion that we need legislation concerning data. The problem is what form such legislation should take and if research records need to be treated in the same way as other records. Likewise, in the future we have to be conscious of the ethical rules concerning empirical research - like we always have been as long as I can remember. Ethical rules are not recent, which one would be led to believe, when one reads some of the descriptions of experiences with the Data Act. In my field ethical rules have been applied for decades, they have constantly been discussed at conferences. They have been widely accepted in similar forms, even if they haven't been codified and formally approved of, and they have been taught to beginners in education at research level. The reasons for taking them so seriously are not particularly altruistic; the most important of all has been to "look after the field", as we like to say. We have enough trouble with effects of wearing off and repetition to want to obtain further difficulties.

Through the debate that has continued the last years, it has become rather well known facts that Social researchers need access to data about individuals and that deidentification often can't take place without disadvantages, whether it be immediately or after some time. The demands being put by deidentification, during the stage of datacollection and analysis of non-response, lead to several registers becoming files on persons in the sense implied by the Data Act. Several types of investigations require access to information about individuals even during the stage of analysis. Then we have the special problems which are connected with panel studies and with actual longitudinal investigations. We have studies of elite populations where deidentification

is difficult or impossible to make, for other reasons than those dealing with analytical technique. Well, I will now turn to giving some examples of on one hand the deterioration of the quality of data and on the other the increased costs of research.

I'd like to add that not all the consequences mentioned here are directly dependent on the establishment and existence of the Data Act. In part they mostly depend on the general strained atmosphere which the discussion on the Data Act and personal privacy has brought about. This may be an important distinction from legal point of view, since it is stressed time after time in the studies of development made by the Data Inspection Board. But for us who have had to struggle with the largely increasing difficulties of practical empirical research the distinction can be neglected.

First, it can be noted that many young researchers tend to avoid choosing alternatives which imply the establishment of files on persons in the sense implied by the Data Act. If they at all decide to solve or illuminate their original problem, they choose the least appropriate and the most expensive ways of doing this. There are several reasons for this but the time factor is of central importance. In the present situation, considering the insecure conditions prevailing for young researchers at universities today, nobody really has the possibility to wait for a permit for an investigation to take place, and to be able to start it.

In addition to this, one can see clearly that many researchers are reluctant to collect a type of data that they without doubt would have collected 5 years ago. Among other things it is here a question of political and religious variables. The deterioration is immense and for a field such as the one I represent it is catastrophical. Among other things, this is so because it no longer matters if data is collected to be treated by a computer or not, if it occurs in cooperation with the parties concerned or not, and so on.

All that is needed is that some journalist catches sight of the word "poll" or "interview" and it becomes a scandal. "Data" appears in the headlines the next day, like more or less insinuating formulations about intention and so on. Nowadays it does not help how quickly the newspaper publishes corrections and denials - if they actually do it which is not always the case.

Also in the cases where the variables are less delicate and where polls are still being used, one has to take into account considerable deterioration. This especially applies in the populations of the type "parents to school children". In spite of the greatest thoroughness in planning and all imaginable finesse in the forming of the contacts with the investigational field, we could in an investigation involving two similar data-collections, 1970 and 1975, note a decline in the frequency of response from 76% to about 46%. No one had any specific criticism to deliver against the 1975 investigation and no one reacted to any special issue. The general change in the atmosphere was decisive. In contrast to earlier norms it was now considered as a good deed as a citizen not to answer the questions.

There is much more to say about the lowered quality of data. But I would like to also include something about the increased costs brought about by the Data Act at the grass-roots level. They are of two types, partly the permit fee, partly other costs. It may seem petty to con-

cern ourselves with the fee; this is a remark I have heard several times when I have taken up this issue. But I believe that it is important to stress the lack of sympathy which one has for researchers' problems. This really illustrates better than anything else the differences between the world of the researcher and those other worlds concerned in the context, and what peculiar ideas people may have of the researchers' conditions.

As we know there weren't any extra grants whatsoever given to the institutions for covering the fees for the use of already existing registers and for those one wanted to maintain. We, on our part had nine such registers, after a very strict selection (as a matter of fact we deidentified a number of registers which we very much had wanted to keep some time for further investigation). Well, at a conference many years ago I was informed I should not worry. There weren't going to be any sums to talk about. That would be guaranteed by the board of the Data Inspection Board.

Then the first and the so far only decision was made. The fee was set at 1,050 Skr. Supposedly the concerned persons in the board are of the opinion that this is equivalent to keeping the promise, in many contexts the sum is ridiculously low. But if one is the head of an institution at a university one can't look at it like this, when one has 21,300 Skr a year to cover all costs for the institution: all kinds of office material, travel, fees to international organizations, visiting lecturers and so on.

Then one may make the assertion that this still is a transitional phenomenon. From this time on the fee has to be noted in the budget when applying for grants to research councils and the like. But such a remark shows the lack of information about how the financing of research functions. One always has to be prepared to bargain and for adjustments of the requested amounts. When certain expenses now become inevitable and permanent there is less room for other costs and the range of research has to be reduced accordingly.

However, a contributing factor is first of all the even more important costs that beside the permit-fee particularly affect small institutions with very limited administrative resources. According to our calculations the additional work involved in one application to the Data Inspection Board amounts to two weeks effective work for one person. Moreover, forms and variable constructions have to be completed before the application can be submitted. Consequently empirical investigations are delayed with at least the amount of time required by the Data Inspection Board to deal with the issue. Considering our financial conditions, the expenses for photo-copies, paper, etc. are very large, and there has been no economic compensation for increasing costs.

No, these new conditions have principally been dealt with within the framework of the normal grants. Naturally the result has been that the amount of research has had to be reduced. Within the project, concerning the consequences brought about by the Data Act, that I work with, an estimation of this reduction has been made. A very approximate estimation is that the reduction directly caused by the Data Act for an institution like ours amounts to 25-30 per cent.

Aren't there any positive contributions on the present situation from the point of view of an active social scientist? Well, they are not many, but they exist. Even in the Data Inspection Board's own ac-

count for the experiences of the Data Act during the period 1973-75, indications of improvements are given in certain respects; this regards details, but important ones, and they will most gratefully be noted. Thus the fee to obtain permission is to be discussed over again, and possibly large commercial institutions might stand indifferent to this fact, but certainly not researchers.

It's also important that a simplified procedure for getting permission is intended be applied to some of the research records which comprise information about a limited population, which have limited durability and which do not contain too delicate information.

Otherwise there is no improvement to be expected. Rather, one has to take into account a deterioration of the atmosphere for social research, particularly if the law is extended to concern even manual records.

What then would the representatives for social research really like to suggest? What would they do if they could decide themselves? Well, different viewpoints concerning how the Social researchers in general look upon the data and privacy issues can now be found in a number of statements made and positions taken by academic associations.

Maybe the Social researchers' anxiety and hopes can best be found in the petition directed to the Data Inspection Board by the Swedish Association of Sociologists in March 1975. Here a clear and precise analysis is made of the distinction between research records and other records on persons. One also finds a good description of the characteristics making research records less exposed to risks with regard to personal privacy and there are interesting suggestions for the establishment of a safe control of secrecy.

It would be good if the following quotation from the petition of the Association of Sociologists could be kept in mind. "The issue could be expressed like this" it says "that what by law is held for a record on persons, always is a record on matters for Social Research, where the person is an attribute to the 'matters', i.e. to the figure-variables, and an attribute of only momentary administrative interest for non-response- and panel-analysis as well as the combining of different registers. Otherwise the persons are quite irrelevant attributes in the context. This distinguishes the use of personal data in research from most other ways of using it, where the person is the central factor, and where the circumstances are attributes to the person." The Association of Sociologist's pronouncement concludes by expressing in my opinion reasonable demands. If necessary one may approve a simplified procedure of application of the kind that is applied concerning certain administrative records. Review and control of individual pieces of information should not be considered according to the opinion of the Association of Sociologists - "as protection of the integrity of research which in this context should be more threatened than any registered person's privacy".

However, The Association of Sociologists would prefer that only a simple duty to inform the Data Inspection Board be demanded regarding the establishment of research records with identifiable information on persons, since the risk for illegal use and intrusion is so small.

Well, with this my contribution has come to its end. I have tried to avoid polemics. The reason is not that I usually dislike this, on the contrary, I do. No, the reason is that I feel a growing helpless-

ness concerning this issue. For 30 years I have in my profession kept myself informed of the creation of political opinion in Sweden but I've never felt such disappointment of the result as when concerning the issue about research and privacy. On the other hand I'm completely aware of the fact that I belong to a minority, which has very small chances to change the present development. Times are hard for the one who believes in the task of free Social research to guarantee innovation and creativity. For each passing week my suspicions grow stronger that we have the glorious epoch in the history of empirical social science behind us. When a year and a half ago I entered the data debate I was upset but optimistic. Today I'd rather say that I feel sad and pessimistic.

3.2 THE POLITICAL RESOCIALIZATION OF IMMIGRANTS PROJECT

by

Tomas Hammar

A specific case will be dealt with in this paper, a survey of immigrants' political behavior and attitudes in Stockholm. The case will illustrate some of the central questions raised by the Data Act of 1973. What research is legitimate? What organizations and institutions shall be asked to give their opinions on individual applications? What questions shall be allowed in a survey interview? And finally, what is undue intrusion on personal integrity according to the Act and in the application of this Act?

The case: The Political Resocialization of Immigrants Project

Financed mainly by the Bank of Sweden Tercentenary Foundation, this project runs from 1972 to 1978 at the Political Science Department of Stockholm University. The political resocialization of adults who have migrated to a new country has never been investigated. There is a lack of knowledge of the processes involved, and there is at the same time a need for such knowledge, as a number of institutions and organizations in Sweden expend much effort to influence these processes in order to improve the consequences in a relatively large immigration.

A pilot study was conducted in Södertälje, south of Stockholm, in 1973, a survey of more than 500 interviews with Finnish- and Swedish-speaking Finns plus a comparable Swedish group. A report on this study was published in 1975. No permission from the Data Inspection Board (referred to as DI) was required at the time. But in December 1974 an application was sent to the DI to preserve a record of the Södertälje survey for a short period of time.

The main study of the project comprises about 3 000 interviews with four national immigrant groups plus a Swedish group in Stockholm. The interviews, of about an hour's duration, are all done in the immigrants' own languages. They were planned to start in January 1975, but were delayed by waiting for the final decision of the DI till April 1975. The last interviews will be made in May of this year (1976). The data will be statistically analyzed in computers. Reports will be published in a series of doctoral dissertations and a final report will present a comparative analysis of the entire study.

An application for two registers was sent to the DI in December 1974, one for the pilot study and one for the main one. Quick treatment was promised in preliminary contact with the DI; the decision, however, was made only in the middle of March. The civil servant who handled the case did his utmost to speed the procedure, but the case was considered both complicated and important and as it was the first principal decision in this field of research, a delay was unavoidable. A fee of almost 8,000 Swedish crowns was, however, reduced by 50 %.

The DI requested that the interview questionnaire should be attached to the application, but did not ask for a specification of the purpose of the interviews as a whole, of the groups of variables included or of the operationalizing of these variables in various questions. This caused problems as soon as the DI sent out the application asking the following organizations and institutions for their opinions: LO (Swedish

Confederation of Trade Unions),TCO (Swedish Central Organization of Salaried Employees), SAF (Swedish Employers' Confederation), Statens Invandrarverk (Swedish Immigration and Naturalization Board),and Stockholms Invandrarnämnd (Stockholm Immigration Committee). The two last mentioned were also asked to find out the opinions of those immigrant organizations whose members were included in the project. Without notifying the Political Science Department in advance, the DI sent out the questionnaire together with the application. The immigrant organizations in this way received copies with a large number of questions but no explanation of the general purpose of the project nor of the specific purpose of this or that question. SAF critizised this way of handling the matter.

Both the Swedish Immigration and Naturalization Board and Stockholm Immigration Committee were strongly positive and showed great interest in the project. They made efforts to explain to the immigrant organization why they wanted this survey and how it could be of value to the organizations as well. The National Union of Finnish Associations in Sweden, the Polish Refugee Council, the East European Social Committee, the Polish Ogniwo Society in Stockholm, the Turkish-Swedish Association in Spånga, the National Unions of Yugoslavs - all advised in favor of the application. But all had access to the questionnaire, which in this way got a wide circulation. Questions were published in the press, completely out of their context in the interviews.

In March 1975 the DI decided to grant permission for two records used in scientific research and established only for a relatively short period of time. Among the conditions stipulated in the decision was the instruction that five questions about political party identification should be removed (see final section below). Other conditions were that every person interviewed should be given information about the content and use of the record, about the use of computers, etc. Specified dates were set for the depersonification of the records.

Amount of information and nature of data

Permission to establish a record shall be given, according to article 3 in the Data Act, "if there is no reason to suppose that this will cause ... undue intrusion upon the personal integrity of the registered person" (om det saknas anledning antaga att ... otillbörligt intrång i registrerads personliga integritet skall uppkomma). It is left to the DI to determine what the content and the limitations will be of the key concepts in this Data Act, "undue intrusion" and "personal integrity". A number of illustrative comments are given in the opinions and in the DI's decision in the case we are dealing with here.

SAF stressed that the record would include a great number of information of a kind that could be classified as soft data, attitudes, values and opinions. "When all these data are seen together and especially over a period of time, they may be said to constitute a threat against the personal integrity of the individual person registered. To this must be added, that the populations of this survey are mainly immigrants, for whom the survey procedures and possibly the partly unfamiliar questions might be an experience that causes anxiety."

If this statement were interpreted literally and read outside its context, it would imply that social science should be restricted to small investigations, with few variables and no time series, and further that surveys should not be done on persons unfamiliar with survey techniques. As we shall see, however, SAF made this very general comment as an introduction to its statements on certain items in the questionnaire.

TCO declared that in principle "the Data Act should not be applied in such a way that scientific research projects of evident usefulness for the society and for persons affected by the research are prevented or unnecessarily obstructed. Great importance should thus be given to the ultimate aims of the project as well as to the attitudes towards the project among those who were to be interviewed and registered."

The project in this case was considered valuable. But what about social science research that is not considered evidently useful for the society, or that does not gain a positive response from the population that will be studied? This is one of the crucial points in this case. The opinions of SAF and TCO or other interest organizations are used to judge the relevance and value of research projects. There is a definite risk that only projects which, in the opinion of such organizations, are highly valuable will be given permission to establish records according to the Data Act. Interest organizations and governmental institutions might in this way exercise a decisive influence upon what research may be carried out and what may not.

The Swedish Immigration and Naturalization Board and Stockholm Immigration Committee both showed great interest in this special project. They gave in their opinion to the DI their strong support for the application. But it is easy to imagine a situation, where a research project is planned without such support of even in opposition to prevailing ideas within the administration. The composition of the Board of the DI and the procedure of asking for opinions from agencies and organizations have established a control not only of registration but also of social science survey research in general.

Political party identifications - five questions forbidden

The following items in the questionnaire were ordered removed: "What political party has the best solution to (a political problem)? What political party do you prefer a) in Sweden and b) in your country of origin? Were you/Are you a member of a political party in your country of origin/in Sweden?"

According to the Data Act article 4, special reasons must be given to grant permission for a record including data about a person's political and/or religious views. Statistical or scientific research purposes are foreseen as such special conditions, if sufficient guarantees are given against leakage of information.

In this special case, the DI found that the risk of unintentional leakage of information was hard to estimate, but the damage that such leakage might cause an individual could be considerable. As only few persons, the DI said, would voluntarily answer questions about their party identification, and finally as data of this kind were not absolutely necessary in this project, the five questions should be omitted.

SAF wrote that data on party affiliation or identification in this case "were completely in conflict with the intention of the Data Act". In this case data on political views were coupled to an abundance of other data from interviews and registers.

TCO discussed in detail the relevance of the questions about party

identification to the general purpose of the project. Data on political behavior of immigrants might have scientific relevance, but were irrelevant for society's planning of its immigration policy or for measures related to certain immigrant or immigrant groups. "It is of course impossible to predict to what extent knowledge of the immigrant's political party choice would influence the parties' views on immigrant issues, but already the very thought of such an influence brought about hesitation", TCO said. No demand was raised, however, that those questions should be omitted.

A few comments are in order. The DI had never asked for motivation for including the items on political views. In a telephone call it was asked whether prohibition of the political party questions would stop the project. A negative answer was interpreted as if these data "were not absolutely necessary for the purpose of the project". TCO showed that it did not understand the importance of these questions. The effect of the deletion of these questions is that there are blank spots in the middle of the maps we are trying to draw. A local election reform in 1976 gave suffrage to immigrants in Sweden. Great effort is being made to inform the new voters about their rights, their choices and the significance of these choices. The political socialization processes studied in this project cover a much broader spectrum of the political system, but immigrants' knowledge about and evaluation of the election and of the party system are central parts of the study.

Neither the organizations nor the DI asked for the reasons why these questions were included. They made their own judgement on whether or not the questions were relevant for the research project. The DI even went so far as to write that it was technically impossible to pose these questions, as only few respondents would voluntarily answer them. This way of arguing was both incorrect and unreasonable. It is true that many immigrants give "don't-know" answers to questions of this kind. But this in itself is a meaningful and very interesting answer with regard to the political socialization of immigrants. The vague provision of the Data Act article 4:3, that there must be special reasons to permit registration of political and religious views, has thus in this case been used by the organizations and the DI as a pretext for passing judgement on the analytical relevance of asking these questions.

"Undue intrusion"

In conclusion, the DI and the organizations and institutions which give their opinions to the DI in their application of the Data Act control more than the technical use of records. Their broad interpretation of the concept "undue intrusion" has opened the door to control of the relevance both of a research project in general and of variables and items in particular.

An interview lasting one hour and consisting of more than 150 questions of a more or less personal nature must always be an intrusion on personal integrity. To decide whether this intrusion is undue or not, the DI wants to know what the purpose of the survey is, who is in charge of it, and how severely the intrusion is perceived by the person interviewed. Article 3:2 in the Data Act stipulates that in evaluating the risk of undue intrusion the DI shall consider "the existent or supposed attitudes of those who might be registered" (den inställning till registret som föreligger eller kan antagas föreligga hos dem som kan kom-

ma att registreras). As the greater part of the population in this case were immigrants, both SAF and TCO stressed the necessity of special caution. The positive responses of all the many immigrant organizations heard from were essential to the DI's final granting of permission.

But these responses were positive not because the immigrants said there was in this case no or only a small intrusion on personal integrity, but because they wanted this kind of research even if the intrusion was considerable. In another situation other organizations or groups might call the intrusion undue not because it was in itself severe but because they did not want the research project. This again illustrates that research projects under this interpretation of the Act are dependent on the attitudes of organizations and institutions.

Outside the present Act - or inside a revised Act?

Could this project work without permission to use registers? SAF in its opinion to the DI advised the researchers to make themselves independent of the restrictions given in the Act. They should perform all their interviews and collect all their data first, and then abolish all personal identifications and establish their depersonalized records before starting their computer operations.

TCO did not elaborate the idea as far as SAF, but mentioned that the outcome of such a procedure would be a heavy increase in costs and consequently a decrease in the scope and intensity of the analysis.

In this special case one record (the pilot study) existed already before the application. The other was established in order to get access to the National Population Register for sampling of the national immigrant groups, and for this a positive decision by the DI was necessary. The alternatives suggested by SAF and TCO were not available. If they had been available, they would have caused extra costs and technical problems, but they would on the other hand have eliminated the restrictions of the DI decision. Interviews could have included all types of questions on political and religious views. The immigrants interviewed might have experienced this as much of an intrusion upon personal integrity. But the Data Act would not have been concerned, as it is exclusively directed to records. One of the lessons that might be learned from this case is thus that the Data Act might force researchers to find ways, costly and traditional, outside the control of the DI. The alternative might be that the Data Act be constructed with the following aims:

1. to distinguish clearly registers for scientific research from all other kinds of registers,
2. to define "undue intrusion" in such a way that the control under the act is directed only at the technical processes involved in scientific research, preventing misuse of records or leakage of information, but
3. to refrain from all control of relevance, content or approach used in scientific research.

3.3 LONGITUDINAL STUDIES AND THEIR NEED FOR DATA

by

Carl-Gunnar Janson

Truisms have the merit of being true. Moreover, it is often important to remember that they are true. I shall mention a number of truisms and will begin directly with the following one:

"If we want Swedish social sciences not to be only philosophic and exegetic but empirical we have to make sure they are supplied with relevant and not completely harmless data."

Whether or not it is important for society to retain and develop empirically based social sciences is a question which our political leaders must take a stand on. If they believe it is important, then they should remember the truism that empirical social sciences can not survive without empirical data.

Empirical data can clearly be of different types (the second truism). Sociologists, for example, most often use cross-sectional data obtained by surveys, i.e. data that describe a situation at a certain point in time. Most often the information refers to persons or families, even though other units like neighborhoods and local communities occur. If we confine ourselves to information about persons or families, it is generally derived from surveys, i.e., questionnaire and interview. Such individual cross-sectional data are often clarifying, but they also have limitations and weaknesses. These, for example, emerge when one wants to study changes (the third truism). In general it's difficult to avoid interpretations and ways of reasoning that concern changes and time factors. For example, if one finds a greater religious interest among elder people than among young people one naturally asks oneself to what extent this difference can be attributed to differences between generations and to what extent it can be explained by religious interests increasing with age. A mixture of slow and rapid changes may give an extremely complicated cross-sectional picture.

We may then extend the material to embrace independent repeated cross-sectional trials. This will give better results. But differences in time can on one hand originate from displacements referring to individual persons, on the other from the substitution of persons between the cross-sections, the latter through changes in the population and through using separate random samples. With only one observation per person you naturally can't know which persons have changed positions. Repeated cross-sections involving the same persons would be even better, but if we want to know which persons have changed positions, we have to interrelate the individual values over time. Other possibilities are to observe irreversible data when they occur or to pose questions about earlier conditions. Still such retrospective questions, as we know, often have serious sources of errors. Repeated cross-sectional data sets with identical individuals are for instance obtained by before-after experiments, something which has additional advantages, and panels, for example, in election campaigns. But both experiments and panels typically concern only short-term effects. If we want to study long-term processes, development of an individual

during childhood and youth, educational effects, social mobility and deviant behavior, observations over a long time are required, preferably not retrospective observations but prospective ones. The period of observation can be covered by repeated cross-sectional surveys of a given population or by observing relevant situations and changes when they occur or through a combination of these procedures. Such a long-term study is what is usually called a longitudinal investigation. As research proceeds from simple to more complicated problems, when you need higher precision and validity of data, when sociologists after having enthusiastically stated the existence of differences turn to estimating the size of the relevant variables, the need for longitudinal studies rises.

One basic prerequisite for longitudinal studies is an effective and public registration system. The Swedish registration system in its wide sense is surely the foremost asset for the Swedish social sciences, the best compensation for our smallness. The Swedish academic community is as we all know a small part of a small society. The population registration system, in which I include the whole system of official registers of persons, has unfortunately not been made use of by social research to its fullest potential. This I see as a consequence of the general American influence on Swedish social research. In the US the same sources of information do'nt exist. Therefore, American textbooks don't give them and the types of investigation based on them very much attention, but deal mostly with cross-sectional studies based on direct interviews. This brings about a similar tendency among us Swedish sociologists to have the same prejudice according to which secondary data become second-rate data. Nevertheless, something like a Swedish longitudinal tradition has developed. The Swedish contribution among the longitudinal studies is striking. Studies as the ones by Härnqvist & Husén and Boalt contributed to drawing attention to the question of the social class recruitment to higher education. At least during the last decade an increased interest in the longitudinal approach has been noticed with many new projects of this type: the one by Härnqvist & Svensson, The Örebro-project, Project Metropolitan, the one by Bengt-Olof Ljung, historical ones, social-medical ones and so on.

For a longitudinal investigation one first needs to be able to identify and locate the studied persons. One would prefer to keep the population within a restricted area, but from the original reference area it spreads out during the long period of observation. This leads to various technical problems that we will overlook here. The essential thing is to maintain the identification in one form or another during the whole period, as long as data are being collected. De-identification is thus not possible.

Secondly, one needs relevant data concerning the whole period of investigation (a new truism, I don't know which number it ought to be assigned, since I lost count). On one hand, data may come from inquiries, and on the other from different registers. The data in the registers in turn often originate from interviews or questionnaires, but not from research interviews or research questionnaires. They can also be reports on decisions or other actions. To a degree they have other sources of errors than do data from inquiries. Sometimes they are superior, sometimes inferior with respect to relevance and reliability. Regardless whether one makes use of surveys or records or both, data have to be provided for different times, that is be put together from separate or continuous substudies.

Until now there have been three types of data available for Swedish longitudinal studies.

First: Information from population registers and other public records. These data have been useful in sampling and follow-up, since they have been available to researchers as well as to others. Thanks to the Swedish public documents principle and our extensive population registration system in its wide meaning, data of this kind have become remarkably extensive.

Second: Data from investigations involving direct participation of the subject in the study; mostly surveys but also health investigations or the like. It's self-evident that participation in such investigations always has been and still is voluntary. Yet, it has often been possible to make the non-response frequency relatively small because there is a high probability to locate a person selected for a sample and because respondents have shown rather confiding attitudes. An interview might naturally entail an intrusion into someone's private life, an intrusion that responsible researchers always have tried to restrict as much as possible. Here it is essential that participation is voluntary and that this is made clear, so that the persons involved can decide for themselves if they consider that the circumstances speak for or against their participation, their children's or their family's. Of course their decision might depend on to what extent they believe in the promise of confidentiality given by the interviewer and the researcher. The reality of the voluntariness has been put in question, but if it is as fictitious as some claim one would consider it strange that non-responses due to refusals to participate actually do occur and this is obviously more and more often the case.

Third: Data from non-public registers that have been made available with special permission, that is, through a political decision, by the government or a governmental agency. The permission has then been given with certain preconditions: that the research is of responsible, non-commercial nature, that the data-secrecy is ensured and that publication is made in a way that does not allow identification of individual cases. The point of departure has been that it is considered to be good for society that the information in question is made available. It's worth noting that on the one hand these political decisions were made without the registered persons being asked or being able to withdraw from participation, on the other that the criteria on social utility were unspecific, which would make room for critical research. Moreover, it has to be stressed that quality scrutinization of research is generally considered to take place through the treatment of applications to research councils for grants, that is in another context than the granting of permission.

The Data Act brings about a threat of deterioration so that the whole investigational type of longitudinal research is jeopardized. If that threat should become real it will lead to a serious loss for research and thus also for society and its members. Until now the discussion about personal integrity hasn't sufficiently considered the longitudinal studies and their demand for data.

A longitudinal investigation requires a so called research register. Research couldn't be pursued without such a register, but on the other hand it doesn't require more than that in terms of data. It does not require a so called decisional register. By a research register is

meant a register used for research but not for making decisions concerning an individual unit in the register, (a person, a family or the like), the latter with one exception: namely concerning possible decisions to try and acquire data about the person through his direct participation, for example in a survey where participation naturally is voluntary. The most important element in the definition is the demand that, with the given exception, the register is not to be used as a basis for decisions on measures to be taken, be they positive or negative, upon the individual subjects of registration. By conclusions from its results or the like, the investigations may have indirect consequences for society and its members, but it doesn't bring about any particular consequences for the individual subjects of registration in contrast to other persons and families (except from being asked if they want to participate in prospective surveys, etc.).

Thus the definition of a "research register" is essentially negative: the decisive factor is that decisions about actions are not made with reference to the register. The use for research in itself does not make a register into a research register, which means that if in addition the register is used as a basis for decision on measures concerning registered units, it is a decisional register. In contrast, a research register may obtain information from decisional registers. Information is then being used for another purpose than intended at the time of the collection of data, unless you don't take into account from the very beginning that the information in decisional registers may be used for research. Not all empirical social research about persons needs personal files. Personal identification is not needed in the treatment of data in pure cross-sectional studies. On the other hand, certain types of research demand data that can't be included in a research register. This is the case with so called action research. That the researcher more or less directly interferes with the living conditions of the participants, is here an essential factor. In contrast, the results from experiments in a usual sense could be included in a register of research. Current ethical rules like voluntariness, confidentiality, non-detrimental content, dissolvement of possible effects of the experiment after it is completed, etc., are, of course, valid for experiments. To conclude, research registers thus regard such research which doesn't comprise or lead to measures upon participant individual units other than possible efforts to involve the participants' direct, voluntary contribution to the data collection.

Data are often transferred to research registers from decisional registers and from several different registers. As we know in the Data Act such mergers are regarded with suspicion, but here too one should make a distinction between decisional registers and research registers. Merging in decisional registers are often aimed at crosschecking and may give new significance to the information in individual registers. It might become a question of fatal revelation of conditions that the informant has tried to conceal. In research registers, mergings can be a stage in a validity control, but discrepancies are completely void of consequences to the persons concerned. Moreover the objective of merging is most often another one, namely to obtain additional variables, for example to cover several subject fields and thus get data about additional independent variables, of which a dependent variable is regarded as a function.

Research registers and decisional registers also differ in terms of required precision. In research, precision is not required for the

sake of individual decision-making. One deals with errors in measurement and feeds these into the model of analysis and aims at general conclusions about categories. One could even introduce a compulsory noise in research registers, with a random variation adjusted to the context, so that no one reading the data could be sure that the registered information for a person really describes the characteristics and conditions of this person. This should make the information completely harmless even in wrong hands and would replace different forms of "randomized response" and "combined questions". This procedure is, however, inconsistent with the Data Act.

Certain longitudinal investigations aim at prognoses of some outcome for individual persons with given constellations of characteristics or living conditions. Nor does that type of study mean or lead to decisions about actions regarding individual persons in the investigation. In this, the connection is established between results, characteristics and conditions which form the basis of the prognosis, which possibly can be used for other persons later on. In contrast, the outcomes for the subjects of registration have become known already in the investigation. For them there is evidently no question of any prognosis with possible adherent decisions.

Reasonably the question of the right to personal privacy is essentially different for research and decisional registers. In the first case it is a question of data entailing no consequences for the individual. In the other case data can be of great importance to the individual. It can be added that it isn't at all clear what is understood as a threat of ones right to personal privacy as regards data. Limitations are placed on the individual's privacy by society. What citizens may reasonably demand and may reasonably have to accept has to be determined by political decisions. If thereby the consideration of the right to personal privacy is regarded as requiring certain additional restrictions on research, this may lead to an already small minority being able to force their will upon a strong majority as regards what kind of research to carry out. For example if the subject of registration is to give his permission every time you are feeding data to research registers it becomes practically impossible to use such data for longitudinal approach. A survey is then added to the gathering of data. The costs in time, labor and money for this may become close to prohibitive, particularly since the size of random samples from registers often are much larger than ordinary samples for surveys. The non-response of not found persons and persons not wanting to participate may often be expected to be a selection with particularly interesting cases overrepresented. If only 20 or 30 per cent of a cohort have to be excluded from a study of data from a register, by not having given their permission to it, this can make the investigation almost pointless. Then it doesn't help if the rest of the cohort by any chance would wish that the longitudinal investigation be carried out, for instance, because they believe it may result in helpful knowledge about schools, conditions of growing up, class differences, different social problems, etc.

If one takes as a point of departure the sense that has here been given to "research registers" and if one really understands what such a personal file implies it is difficult to see how a research register can reasonably be a serious threat to the right to personal privacy. It

is difficult to understand that it would imply an inappropriate demand on the members of society.

This is valid if an important prerequisite is taken into account, namely that the secrecy is satisfying. This is what should above all be discussed concerning research registers. Here we have a direct connection and a direct support in the professional ethics of the researcher as it was developed long before the Data Act came into existence. The responsible researcher is not in a position to guarantee anonymity but to promise confidentiality. The information is given in confidence and is not divulged to authorities or other persons. We have here a non-codified correspondence to the relation of the doctor and psychologist to their patient and the one of the lawyer to his client. It has long been clear that such an attitude is an absolute prerequisite for surveys and longitudinal studies. To my knowledge, there weren't any ethical grievances about Swedish social research lying behind the demands on protection by law. There wasn't a radical change in conditions with the ADB. Surveys and data in registers both within and without longitudinal studies existed much earlier than ADB. In certain cases it is or was easier or at least quicker with ADB. In Project Metropolitan many types of data have been obtained by manual search of files and excerption of dossiers, and the data we have acquired through computers are of the kind that older longitudinal investigations got too. I suppose that coded data and data on tape imply an improvement of the protection of secrecy as compared to uncoded data on forms. Contrary to this we have to admit a deterioration in this protection through the Data Inspection Board's list of records, so that now one can easily get to know where to look for what.

One of my last truisms will be that there is never any 100 per cent guarantee for anything. There are always risks. This might be the case for aircraft or other traffic, power supply or participation in conferences, but the question is the relative size of the risks of different alternatives including renouncing from activity and the possible profit of the different lines of action. The same applies for ADB. I presume that all data can be cryptographically analyzed if one has sufficient will and enough resources, whatever the precautions taken. (Data in text en clair don't have to be decoded.) Here one has to judge the risks against the incentives for such violations. Of course, somebody who wants to prove that decoding is possible may have incentives for such exaction, but otherwise it should be better to spend one's energy trying to exact material from decisional registers. Criminal records, social records, records of venerial cases and other medical casebook data, records of party members, etc. should be more useful. As a rule, with reasonable means of protection research registers are, or can at least be made, more complicated to decode and give less results.

To conclude I'd like to stress the fact that I'm convinced that the problems I've dealt with here can be solved if there is enough good will. I will finish with a last truism alleging that researchers have such good will.

3.4 IS THERE A NEED FOR LONGITUDINAL STUDIES IN THE SOCIETY OF TODAY?

by

Allan Svensson

I will here present an orientation on a longitudinal project, called the Individual Statistics project, and I will mention something about its importance to research and investigation work made in the field of education. Moreover I'll consider some of the consequences that the Data Act and the so called ethical rules may bring about concerning these types of longitudinal investigations.

The outline and aim of the Individual Statistics project

Since the Beginning of the sixties the Swedish Central Bureau of Statistics and the Institute of Pedagogy at the University of Gothenburg have gathered information for the so called Individual Statistics project. The first time a compilation was made was during the spring-term 1961 and then concerned pupils born the 5th, the 15th and the 25th of any month in 1948. This information for about one tenth of the generation was then annually completed with data until 1969. In the spring-term of 1966 a compilation of information was started in the same manner concerning pupils born the 5th, the 15th and the 25th of any month in 1953 and this information has annually been completed until 1974. In the first test sample the number of individuals amounts to about 12,000 and in the second to about 10,000. In both test samples about 90 per cent of the pupils were, on the first occasion of compilation, in the sixth form within the compulsory school system.

Among the information gathered for the project one can mention:

I Basic information

a) Information about school attendance, for example form, type of class, character of the class, school reports.

b) Information about certain personal conditions like the profession and education of the parents.

c) The results of three ability-tests, one verbal, one spatial and one inductive.

d) The results from the standard tests in the mothertounge, mathematics and English which are given to pupils in the sixth form.

e) Answers on certain questionnaires that illuminate the pupil's attitude towards school, his leisure interests and plans concerning future studies and profession.

II Annual information

Information about schoolconditions of the same type as above under I a). The information is collected as long as the individuals are undergoing education.

III Information of enrollment

The information is obtained when the male part of the pupils enroll for military service and consists among other things of educational data, results from four intelligence tests and answers on certain questions concerning adaptation to the home, to school and work.

IV Information from U-68

According to instructions from the Investigation on Education of 1968 (U-68) inquiry-data have been collected for about a third of the pupils who are included in the test sample of 1966. The inquiry deals with among other things the attitude of the pupils to higher education and their views on different professions.

V ATV-information

In the years of 1970 and 1973 an inquiry was sent to the approximately 2,000 men who are included in the test-sample of 1961 and who don't have any theoretical education above the compulsory school level. The collected information, among other things, illuminates the attitudes to adult education among these persons.

The aim of the Individual Statistics project and the data bank established by the project may be seen as threefold:

1. To enable follow-up-studies of large representative samples of pupils and see how different geographical, social and psychological factors influence the choice of education and profession, and to examine what changes in these respects the shift to the nine-year comprehensive school has brought about.
2. To give a basis for studies concerning the significance of different environmental factors to displacements of intelligence, partly within a sample of pupils tested at different age-levels (13 and 18 years of age respectively), partly between different samples of pupils tested at the same age-level but at different times (1961 and 1966 respectively).
3. To deliver data to investigations aiming at illuminating how different types of demographical and personality-related factors are related to success and adaptation at school.

Within the first sector we have carried out several investigations concerning the choice of higher education. We have studied the tendency to continue with different kinds of secondary education among pupils coming from different school-forms, from different parts of Sweden and from different strata in the society. Success in studies have been related to home-milieu, ability-profile, study-ambition, leisure activities etc. Recently we have given special attention to the problems of adult education and we are at present studying the covariance of different background factors with choice of and success in different types of post-secondary education.

Within the second sector we have studied, with the help of the basic information from 1961 and the enrollment information from 1966, relative displacements of intelligence between 13 and 18 years of age. It becomes apparent that the displacements of intelligence that occurred during the five-year period are systematically related to differences in theoretical education and to a certain extent to divergent home-

backgrounds. This concerns mostly a quantitative factor which measures the level of general ability, but also to some extent a qualitative factor which gives a measure of the structure of the ability.

The pupils born 1953 and tested 1966 went through the same intelligence test as the pupils born 1948 and tested 1961. The average points of the test have increased with a few units between 1961 and 1966. The increase is largest in the verbal test but quite noticeable in the inductive and spatial tests too. The results indicate that there has been an actual rise with respect to intelligence for 13-year old pupils between 1961 and 1966. It's also worth mentioning that this rise is higher among girls than among boys, somewhat higher among pupils from lower social groups than among upper social groups and clearly higher among pupils from thinly populated areas than in densely populated areas.

Within the third sector we have investigated the significance of home-milieu to school-work and we have among other things made thorough studies of over- and underperformances at school. In this connection we have found that pupils from higher social strata succeed better in school than one would expect from their intelligence, while the contrary goes for pupils from lower strata. The connection between social background and relative school-performance still varies systematically, partly between different subjects, partly when standardized proficiency test and the teacher's marks respectively are used as a variable on criteria.

HAS THE INDIVIDUAL STATISTICS PROJECT HAD ANY ESSENTIAL SIGNIFICANCE FOR THE RESEARCH OF THE BEHAVIOURAL SCIENCES?

I can answer this question with an unreserved "yes". I then first think of the pioneering studies by Kjell Härnqvist concerning the significance of different environmental factors for the development of intelligence during the years of youth. But even in another way the project has given certain scientific gain - among other things it has as a result three doctoral dissertations and eight licentiate dissertations.

HAS THE PROJECT BEEN OF ANY INTEREST OUTSIDE THE NARROW CIRCLE OF RESEARCHERS? HAS IT HAD ANY FUNCTION THAT WAS OF BENEFIT FOR SOCIETY?

I'd like to give positive answers to these questions too. Let me give two examples:

When the Investigation on Education (U-68), in the year 1968 was supposed to study the choices of studies and profession after the ninth form in the comprehensive school, it was of great value to have the Individual Statistics material available. By completing this with a smaller collection of data, one could in very subtle ways illuminate how different background factors like sex, home-milieu, pre-requisites concerning ability, earlier performances at school, direction of interests etc. influenced the choices of profession and education.

The project has also been essential to "the Investigation of the Inner Work of the School" (SIA, Utredningen om skolans inre arbete). In the directions of SIA it was particularly stressed that one should pay attention to the situation of the weakly performing and the under-

performing pupils. Our studies concerning the problems of underperformance thus become of central importance to the investigation and the suggestions of measures that were given in the research reports of the project are to be found in the main report of SIA.

WHAT CONSEQUENCES FOR THIS TYPE OF LONGITUDINAL STUDIES DOES THE DATA ACT INVOLVE?

From the point of view of research it would be unfortunate if it was not possible to carry out these kinds of investigations in the future. Among other things this would mean that Sweden would have played out her role as a pioneering country on this sector of the pedagogical, sociological research. The different types of longitudinal studies that have been carried out have as a matter of fact drawn great international attention and the results of their research have frequently been connected in manuals.

If it in the future would be too demanding and too many bureaucratic obstacles to overcome to make it possible to carry out longitudinal studies, the risk is too obvious that the qualified researchers will leave this field and look for fields where it is easier to work, where you don't have to risk clashing with the Data Act and the Data Inspection Board.

If thus from the point of view of the researchers it would be regrettable if this type of activity ceased, the consequences should be even more serious when it concerns the general planning work within the area of educational politics. Then it will be considerably more difficult to get answers on such questions as:

HOW DOES THE SECONDARY SCHOOL FUNCTION?

WHAT IS THE CAUSE OF ALL THE INTERRUPTIONS OF EDUCATION?

HOW DOES THE REEDUCATION IN THE SECTOR OF HIGHER EDUCATION FUNCTION?

WHAT MEASURES SHALL WE TAKE AGAINST THE SOCIAL SELECTION?

HOW IS THE NEW UNIVERSITY REFORM TO BE EVALUATED?

If you want exhaustive, thorough answers to these types of issues, you have to allow large follow-up-studies in the future too. Personally I'm convinced, that with some good will, we will find solutions that satisfy both the researchers, the investigators and those who are to see to it that the Data Act is applied.

To conclude I'd like to touch upon another important factor that relates in certain ways to the Data Act. I'm thinking of the so called Ethical Principles (of Research) for Psychological and Pedagogical Human Research that were approved one year ago on a trial basis by the Swedish Council for Social Science Research. These principles will possibly be approved of by the National Board of Education which recently has circulated them for review and comment.

The main principle in these ethical rules is as follows:

"The self-evident basis for research-ethical considerations, within psychological-pedagogical research is that the participants in the investigation are not to be exposed to physical or psychological damage or humiliation and that they have to be protected to the greatest possible extent against discomfort and inconveniences."

Of course I stand completely behind this main principle. No researcher should be of any other opinion, and no one should consciously go against this basic principle. What I'm on the other hand very doubtful about are some of the points derived from this principle. Among other things they say:

1. Participants in the investigation have to be informed beforehand about all the features of the investigation that quite likely could influence their willingness to participate.
2. After the participants have been informed their individual approval has to be received. This should take place under conditions that give them actual freedom of choice. For those not of age approval also has to be received from the guardian.
3. Participants should be informed beforehand about their right to interrupt their participation at any time.

If these points are to be practised in the future, not only longitudinal studies but almost any form of larger pedagogical field investigation will with great likelihood be rendered impossible.

As regards investigations of the same type as the Individual Statistics project, I'm not of the opinion that there is a need for such strict directives. According to my opinion the suggested "security directives" are only needed in cases where the persons of the experiment are exposed to physical or psychological influence lying beyond the normal school-situation, or when the pupils are asked to answer questions concerning things that the school normally has no reason to investigate. These points of view have been thouroughly motivated by the Institute of Pedagogy in Gothenburg in a committee report to the National Board of Education.

3.5 CONTRIBUTION TO THEME NO. 2

by

Lars Wohlin

The Institute for Economic and Social Research (IUI) engages in empirically oriented economic research mainly about companies, with information from the companies. We also do quite a bit of research on the economic behaviour of individuals. We are involved in consumption analysis, budget data for households, migration tendencies of labor, etc., but I will here restrict myself to our research involving firms. One can hold different opinions about the role of the firm in society. Some see it as a social institution which has certain production tasks to manage. Other see it as private corporate body, conceptually almost like a person, having the right to a considerable private sphere, without any obligation to society to divulge information. Myself, I take a utilitarian attitude towards the issue of the duty of the firm to provide information. This duty of the firm shouldn't be greater than what is motivated by the economic welfare of the country and then you may ask the next question: How far does it stretch? First, attention has to be called to the fact that it isn't always easy to draw a line between the individual and the firm. This is the case, for example with small companies. We have come across this problem when investigating the financial situation and earning-power of a small company: it is not possible without investigating the economic situation of the owner. Then you immediately see this difficulty in drawing a line.

The second problem regarding the duty of the firm to provide information or the use of information provided by firms is that this information sometimes concerns employees and the management does not always have the right to devulge information about its employees. This can be the case with individual wage-systems in the company and the desire to link this to the profit-earning power of the company. One would think that profitable companies pay higher wages and so on. We have worked with simple hypotheses like this one, but as I mentioned, here we are faced with difficulties in drawing the line.

A third aspect of the duty to deliver information is that the firms compete with each other and don't want to provide information fearing it might become available to its competitors. The profit generated in a firm is often determined to a great extent by its exclusive know-how. Society has understood the importance of the right of firms to retain this kind of information which is essential to their ability to compete - for example, laws governing patents. To stimulate the companies to create new knowledge they have to be given a right to make exclusive use of this knowledge during a number of years. Otherwise, there would be too little incentive to obtain new knowledge.

When we have gone out and posed questions dealing with such sensitive issues as the firm's competitive position, even though we were otherwise well received, we have received no answers, for having come too close to very "hot" issues. For example, this was the case in an investigation about what generates innovation in firms. Here, we were quickly approaching central and sensitive issues in the Swedish engineering industry, where one is working with plans lying 5-10 years ahead, about the products to be sold then. If they are to succeed with their new products, then it is essential that no other company

introduces an equivalent product. So we always get to a point where the company will simply refuse, even if they have full confidence in us. One has to accept that the companies will not divulge information that is vital to them.

There isn't any Swedish interest in forcing companies to provide information that can make it more difficult for Swedish companies to compete with foreign companies.

I would also like to mention the problems we had when we made a large investigation into the direct investments of Swedish companies abroad and tried to obtain information about their subsidiaries abroad. Here the fact is that if we start demanding information about the subsidiaries of Swedish companies, it becomes a very sensitive international matter. You can imagine that one would find oneself involved with weighing very delicate matters if American, Japanese and other governments would demand that their subsidiaries in Sweden should give out information about Sweden. We in Sweden believe that ITT's Swedish subsidiary which produces equipment for the Swedish national defence, should not be required to give out this information or any information whatsoever, to the American authorities, and vice versa. Thus, considerations of international relations limit the possibilities to demanding information from the foreign operations of Swedish companies. Of course we have carried out investigations, but we haven't been able to penetrate as deeply with the investigations as we would maybe have wished from a researcher's point of view.

When collecting information about firms we of course partly make use of existing material which in this country has mainly been collected by the Central Bureau of Statistics (SCB); but there also exist other institutions and growing number of such institutions which are collecting their own information. The Swedish Price and Cartel Office (SPK) for example are not modest in their ambitions to collect such material. The same is true for the Swedish Board of Occupational Safety and Health. We try to use this material, and here we have practical difficulties. There has been a growing reluctance on the part of SCB of permitting access to their records. Previously we were permitted to make use of a richer collection of material directly from the SCB that made it unnecessary for us to seek it from the firms. Now this has changed. Now we can hardly get any information from the SCB, the research paragraph that is said to exist in the regulations of the SCB seems for us to be very unclear and seems to be applied in a very hazardous manner. Sometimes it works, sometimes not. We feel that we are somewhat handicapped in relation to investigations conducted by the government. I know that the SCB probably would not admit this, and I don't know if it's actually true, but we have experienced it that way when we have discovered the kind of material made available to them.

When we cannot get access to the SCB files we have to send an inquiry to each firm to find out if we may get access to its primary material delivered to the SCB. Thus, this is our normal procedure nowadays, but of course it is complicated for two reasons. It delays the project maybe a year to get permission. In addition, when getting the permission, one has to pay quite much to the SCB to make copies of the material given access to. We started a test investigation, which showed that only to get copies of the data given to the SCB from one company would cost us 5,000 Skr. You will then understand that these charges amount to sums that are not possible to deal with.

Another problem is to get information about sampling frames. To know what firms to investigate we need a sampling frame. The secrecy at the SCB has now become so great that we can't even obtain frames from the SCB. That is, we can obtain a sample, but we can't obtain a record of the firms containing more than names and addresses. We will not get information about the number of people employed or any additional variables. This means that we don't have any possibility to control the quality of the SCB's material before the investigation. Earlier, in the sixties, when we had permission we could go through the population in question and find a great number of errors in it.

Hence, as regards populations and the duty to report, I think that the firms should have such a duty to report or that one should be given access to certain general back ground variables, which are necessary to identify the firms, in order to establish decent frames of sampling. Certain general information like for example the number of people employed and if possible sales statements would be of great value if it could be regarded unclassified material.

I'd like to emphasize the significance of access costs. There is a very subtle difference between being given access to material, let's say it costs a million, and actually not being given access to the material. I mean, that many times the question of economical availability is of central weigh to the researcher. If you are faced with very high costs to get access to the material, naturally research will take place on the large institutions with economic resources. I'm happy to say that IUI has better resources than many others - we are not handicapped - but we have very high expenses in this respect.

The institute itself collects information for research, and data about the firms. We've been the first in this country to collect information about the expenses for research in Swedish industry, about the foreign investments of Swedish firms, the lorry traffic, etc. Then the SCB has taken over these statistics and turned it into recurrent regular statistics. We thus initiate new studies which are built on new theories and which are based on inquiries, and these are of course of a voluntary kind. There is no authority we can appeal to in order to oblige the companies to provide information. We don't represent the state. We consequently have to convince the firms of the suitability in participating in our inquiries, and it should be noted that every researcher really has to put a lot of effort into convincing the companies on the one hand of the general interest of the investigations and on the other that the researcher himself has enough capacity to analyze the collected material. I think the demand on secrecy is possible to deal with. However I'm of the opinion that the SCB and all the others who collect material via special units, have to better motivate the collection of information. When answering an inquiry one wants to know: Which problems are to be solved by the information provided? Which people are responsible for the material being analyzed in a qualified way? Like the firms, I myself don't only have demands on secrecy but also on a qualified usage of the given material.

The firms have very high costs for answering our inquiries. The last investigation carried out cost the institute close to half a million crowns, but at the same time we asked how many weeks had been spent by a qualified clerk to obtain the information concerning the firm, and with a somewhat stereotyped calculation of the costs per hour we arrived at the conclusion that the collection of information itself within the firm cost about the same amount of money. Therefore one really has to

remember not to pose questions that make the information costs too high for the firms, because then you will not obtain any answers or you will obtain answers that are close to pure guesswork. The whole quality of the investigation is then considerably lowered.

Convincing the companies to answer our inquiries is a difficult task since we many times pose new questions, new types of analysis, that the firms are not used to, and then you can't make use of general investigations but have to start with explorative studies, in which you first convince 2-3 firms; then make an analysis of their material, and then show how this may give generally interesting results and finally to provide a foundation for more extensive collections of material.

Of course another problem is that we many times are interested in time-series of information from firms, stretching over a long period. In this respect we are dealing with longitudinal studies when we are concerned with individuals. Maybe we are also dealing with longitudinal studies when concerning ourselves with firms, but when it regards firms it is almost impossible to define what a firm was in 1956 and what it was in 1976. A firm in 1956 could be something completely different twenty years later even if it has the same name. Therefore it is difficult to interpret information from time-series concerning individual firms over long periods of time. The difficulty does not lie in the issue of secrecy but in identifying the object itself.

Information about firms consequently very easily becomes out-of-date. It should be possible to demand less secrecy when dealing with material that is over fifteen or ten years old, than when dealing with recent material. There is no firm which does not experience five year old information as history, as something not worth dealing with, while an individual naturally may think that it's very delicate to deal with what he was doing 25 years ago.

The IUI thus collects material on its own and this means that we are almost in the same position as the SCB when it comes to meeting the demands of other researchers on access to our material. The IUI owns the material and the principle employed is that the researcher who has collected the material may use it, that is, in order to write his thesis, and try to exhaust the material before handing it to anybody else, and this is most often a question of a period of 5 to 10 years. No qualified researcher would otherwise collect material, and one needs the most qualified researchers to collect material if there is to be any material worth analyzing. That's why you all the time have to support those talents who are willing to sacrifice four years of their life for the collection of material. This is why I'm very restrictive with letting other researchers have access to the material, before the investigator himself has completed his project. But of course other researchers must have the right to be taken into consideration later on, among other things, in order to check the quality of the research made. Nevertheless this implies a considerable time-lag and if they are to be given access to this material we too have the same demand as the SCB has, namely that they have to ask the companies to be given access. But for material older than 25 years, we have omitted this procedure. In this case we have modified this principle.

As far as information about plans is concerned I think that the duty to report is meaningless. No company can be held responsible for not fulfilling its plans. You could always put the blame on the change of external conditions. The answers are easily adjusted to what is thought to be expected by the inquiring authority.

I would like to end this contribution by expressing a wish for the development of a dialogue with the Central Bureau of Statistics about how far this secrecy about the information is to reach, with what right you will be able to borrow data tapes and carry out compilation and programming on your own. When merging two tapes at the SCB to sort out certain aggregated information, the SCB has to account for the programming and that programming-cost may, as we experience it, sometimes be very high compared to what we would spend ourselves on that procedure. A much more thorough debate is needed on the principles for using the material of statistics on firms kept by the SCB. Furthermore, it is not meaningful to press the firms to give information concerning plans or information that give such high information costs that you really cannot expect to obtain any answers or information that is so essential for their competitive situation that they are likely to tell tales.

3.6 DISCUSSION

Chairman: Staffan Helmfrid

Sune Åkerman:

I would like to point out more strongly than has been done in our earlier discussion how important the long, comprehensive data series are for research. As you know, in Sweden we happen to have such data series from the mid-eighteenth century. Interference with these series means damage to the life nerve of research. It is therefore extremely unfortunate that in the Swedish riksdag (parliament) bills have been put forward time and again in all seriousness to the effect that the censuses of 1965 and 1970, among other things, should be decoded in order to prevent identification of individuals. These politically motivated proposals have been made with reference to the integrity aspects.

If they should be successful, this would be a dramatic impoverishment of information systems which are unique in the world and which permit longitudinal studies based on individual data, which give us possibilities to carry out really deep analyses of various changes in the society and their effects on the population.

It is still not clear how real the threat to the collection and maintenance of data which is fundamental to research should be considered to be. But there are signs indicating that scientists must beware. Otherwise they risk being confronted with a very painful fait accompli. As for myself, I therefore find it urgent for us to organize a resistance and an enlightenment campaign articulating our views which have hitherto been so little taken into consideration.

Thus, to me the long data series appear much more important, not least to future research which will try to reconstruct what has happened in the post-industrial stage, than the panel studies based on small samples of the population upon which the discussion has centered. Here we cannot escape our responsibility by defining away some of the files by calling them "research files". Without doubt the most exciting research must be based on long, fundamental data series which can then, of course, be supplemented with interviews, etc. Unfortunately, we can never in the future repair the damage which a short-sighted and unwise policy in the present can cause by blocking and deforming the collection and storage of the form of raw data under consideration here.[1]

Örjar Öyen:

Allow me to say, Mr. Chairman, that I am pleased to have the opportunity to participate in this meeting.

The Norwegian Ministry of Justice now is considering a proposal for a Norwegian data act. The proposal has been evaluated by various agencies and organizations, and it is expected that a proposition for a law will be written in the near future, and that the proposition will be considered by the Norwegian National Assembly not during the present

---

[1] For a more detailed argumentation, see Forskningens framtida datatillgång (The Future Availability of Data for Research), prepared by C. Winberg and S. Åkerman and commissioned by the Coordinating Committee for Long Run Research (June 1976), especially pp. 7-23 and 58-78.

session but probably during the next session of the assembly, that is, in about a year or so. Nobody knows what this law proposition will be like in detail, but the recommendations, developed by a committee chaired by Mr. Helge Seip, now Secretary General of the Presidium of the Nordic Council, have some striking similarities in many points with the Swedish Data Act, and it is probably no coincidence that the text of the Swedich Data Act has been printed as an Appendix to the recommendations of the Seip Committee.

According to the proposal research will fall within the domain of the law, and the question of whether or not to grant permission to develop a personal data system for research purposes will be decided on the basis of two main criteria. The first is one of relevance: the other is one of usefulness. Only information may be included in a data system that is considered to have relevance for the solution of a particular public agency's tasks. Research projects which include personal data will have to be evaluated on this criterion. Also, as a basis for the decision about permission there shall be a balancing between the potential usefulness of the results of the research and the negative consequences foreseen as a result of a utilization of the system. Much may be said about this: I shall not go into detail. The arguments which have been used by Norwegian researchers in relation to the proposals are arguments that are well known from the Swedish debate. What I should say is that one general comment from researchers is that if one were able to answer fully in advance the questions of relevance and usefulness, there would be very little reason to conduct the research at all. This is in the nature of research. So, you will appreciate, I am sure, that we Norwegians have a lot to learn from the mistakes which in this case have been committed by "Sweet Brother and Sister".

If you will allow me to make a personal comment to the debate we are listening to today, it would be that maybe even in Sweden the debate is concerned with something other than what it pretends. It is perhaps a paradox that on the one hand social research is receiving a great deal of support and is recognized as useful and applicable - there is a dramatic increase in funds available for social research - while on the other hand there is a deliberate effort to restrict the social scientists themselves. One such restriction concerns researchers' access to data. The constellation of the two tendencies - the support of research itself, and the restrictions placed on the social researchers - ought to be a matter of great interest. And it ought to be recognized as a rather crucial political matter. I think it is remarkable that it has been left to some special pressure groups to conduct the battle for the protection of personal, individual integrity. I believe that the driving force behind this campaign is a desire on the part of some groups to reduce the influence of social research, which is perceived as becoming a significant source of power in society. It is not only that the social researchers are radicals - they are indeed radicals, at least in Norway they are seen as radicals. It is also important that social science research findings often seem cumbersome, provocative, and a nuisance to the interests that particular groups in society seem to have. I was involved in the attempt to launch the Norwegian part of the Inter-Scandinavian Project Metropolitan several years ago. There was a very strong reaction toward this attempt. And even the most moderate data needs triggered a powerful debate that was carried on through the conservative Oslo newspapers for over a year. It led to a debate in the national assembly, and it activated parents to hold protest meetings in

some of the Oslo schools. This gave me an opportunity to attempt to identify the rationale behind the reaction. It was decided by the municipal school board that parents were to have the right to withdraw their sons from the sample of the cohort born in 1953 from being objects of research. We received, unexpectedly, good research data about the reaction itself, which had we had a data act we might not have been allowed to analyze!

The reaction was indeed a reaction toward an image that has been created, an image of the social scientists and their research. It was a reaction that evolved in the highest income brackets and came from the conservative side of the political spectrum, from the areas of Oslo where people are best off. So, I am very sceptical about some of the talk we are listening to here about the need for the protection of individual integrity, particularly when this argumentation is tied to the point that it is a matter of protecting the weak and the poor in society. I feel that one should not be misled by this argumentation, and I am much surprised that politicians in Sweden have not been struck by the quotation we heard today, the quotation from a statement by the Swedish Employers' Confederation. It is possible that a reaction will come.

In conclusion, allow me to state that it is rather remarkable that empirically- and positivistically-oriented social researchers know so little about the distribution of these attitudes in the population, so that the representation of these attitudes may be left to small but very eloquent pressure groups, as can be read from the support declarations which come in through the daily mail.

Ulf Himmelstrand:

Optimistic as well as pessimistic conclusions regarding the implications of the Data Act have been aired in our discussion today. Members of the Data Inspection Board have expressed the opinion that the dangers involved have been exaggerated, and that we social scientists have misunderstood what it is all about. References have also been made to the modifications of the Data Act which are in the pipeline. A new government commission is going to be appointed to look into this matter, and we have been told to expect improvements of the Act which will be to the benefit of social scientists as well. No special consideration was given to the predicament of research in the work preceding formulation of the Act now existing; research is not even mentioned in the Act and the instructions appended to it.

For the sake of our discussion I will assume that these cheerful references to a brighter future are well founded. I will also assume that the suggestion made by the Swedish Association of Sociologists regarding a simpler application procedure (in Swedish: enkel anmälningsplikt) will be designed for the future. By making these assumptions - and they are only assumptions - I am in a better position to discuss certain other matters involved. I am particularly interested in some issues of research policy. Örjar Öyen touched upon some of these issues in an earlier discussion. This perspective is necessary if we are to understand the meaning of our discussion concerning the Data Act and its application.

If all science - including social science - could be considered pure science in the sense that the consequences of scientific endeavors pertained only to the scientific community itself, and to its theoretical models, hypotheses and research methods, then our discussion would

be unnecessary and out of place. The need for discussion of these matters arises from the fact that research methods and research findings may have consequences outside the scientific community. That is the source of our problems. In that context, and with reference to our discussion about data and personal integrity, we must consider what kind of consequences are involved. We must also consider the question of consequences for what and for whom in what context of power and social control. I am sorry that Kerstin Anér is not here today. She touched upon several of these questions but expressed some rather simplistic notions, I am afraid. Research can lead to social control, she said. Research is never neutral, she continued; it leads to control and that is why we in our turn must control research. I think we ought to discuss these matters more thoroughly: consequences for what and for whom in what contexts of power and social control. I hope others will also take up this discussion, since I will only be able to mention a few aspects of this broad and significant topic.

Karl-Olof Faxén pointed out that research rarely has consequences for single individuals. For instance, research can be carried out in preparation for the introduction of some new administrative steering system. The main effect of such research consists in changes in such administrative systems, and it affects individuals only secondarily (even though these individual effects were not mentioned by Faxén). We can also conceive of various kinds of counter-research which when applied contributes to strengthening the consciousness and resources of various collectivities, for instance trade unions, thereby also changing the constellation of power in society. These types of macro-sociological consequences are not considered at all in the Data Act. So the question, Mr. Chairman, is whether I should be allowed to go on talking about these things since we are supposed to discuss implications of the Data Act as such. Just as Kerstin Anér did, I will allow myself to go somewhat outside our main topic. I will do this by making a rather blunt and categorical statement about the responsibilities of research - a statement which must later on be qualified and amended.

The attempt to make researchers responsible for the use of their research findings, I say, is a diversionary manoeuvre by some politicians, civil servants, administrators, bureaucrats and business-leaders - an attempt to divert attention from the fact that it is people in these latter categories, in most cases, who should be held responsible for the uses and abuses of research findings. Some of our evening newspapers seem to find it profitable to engage in the same diversion of attention; they convey an image of research as a threat against people, and as a target for suspicion when, in fact, the responsibility for the application of research should be allocated among politicians, planners, bureaucrats, and leaders of the business community. To this attack we should respond by saying that we researchers produce research. How the rest of you use our research findings is your responsibility.

Yes, this is a very simplistic view of researcher responsibility. It can be qualified and amended by taking into account three types of cases where research has consequences pertaining not only to the scientific community and to researchers as such but also to their responsibility as citizens.

The first case is research using experimental subjects. Usually experimental subjects do not belong to the scientific community. They are objects of research but also individual human beings with their own moral and legal rights. Within the scientific community we have had an

extended debate about the ethical implications of experimental research, and I will not here try to summarize the recommendations made by various ethical committees within the scientific community. Let me only point out the rather misleading way in which the term "experimental subject" is used at times, and even by researchers themselves, unfortunately. This morning there was a reference made to a report by some psychologists who claimed to have carried out a study with, I think, 40,000 experimental subjects (in Swedish: försökspersoner) here in Sweden. As far as I understand this use of the term, it is inadequate and based on a much too broad definition of "experimental subjects". In the strict sense of "experimental subjects", no one could be said to be an experimental subject unless he or she is exposed to experimental manipulation intended to influence or change the subjects so that they are different in some respect after the experimental manipulation has taken place. This is what is called an experimental effect, and this could be defined in physiological, psychological, mental or any other terms. As far as I understand it, this study of these 40,000 persons was carried out through interviews, and even though there may be some negligible interviewer effects in the sense that the respondent in such an interview may come to think about matters which he has not turned over in his mind before, these effects cannot be labelled experimental effects, and these respondents should not be called either experimental subjects or "mice" (in Swedish: försökskaniner) - unless it is considered legitimate to use misleading terms simply for the purpose of giving the impression of a threat to the personal integrity of a person who is interviewed. But of course our responsibility for experimental subjects in the strict sense remains.

Secondly, it is obvious that the researcher cannot legitimately withdraw from his responsibility for such research applications which cut only in one foreseeable way, and which induce some kind of damage to people. Let me take one example which perhaps is not the best possible one, but I discussed it with Karl-Olof Faxén this morning, and I also think he referred to it in his introductory speech today. I am thinking of so-called prediction studies where the purpose is to find a number of indicators which can help you to predict who will become, for instance, a drug or alcohol adict or criminal or something of that sort. What is the use of such an instrument of prediction? I would like to assert that such prediction studies are of little or no value scientifically - except, perhaps, as a methodological or statistical exercise. In more substantive research such instruments of prediction are rather uninteresting since they conceal more than they illuminate. An ideal instrument of prediction should involve a large number of different variables or indicators which are weakly correlated among each other but maximally correlated with an external criterion of what you wish to predict. Then you add up the values of the several indicators involved. Since the parameters of the predictive equation are determined on the basis of multiple regressions with reference to the criterion without much concern for interaction effects, and since parameters often turn out to be variables rather than constants under changing structural conditions, it adds little to our body of generalizable scientific findings. However, sometimes such instruments of prediction are used for purposes of selection or treatment of individuals or groups. They can be used with regard to individuals only if the instrument is based on a very large number of items and is sufficiently reliable; less reliability is needed if selection and treatment procedures are targeted on certain groups of people who are singled out as social risks

by such an instrument of prediction. Without going into any further detail, I think we should ask ourselves whether this kind of application of instruments of prediction for selection procedures is the kind of consequence which we should accept. Whatever our answer to this question, I think the researcher who has constructed such an instrument of prediction should be aware of these consequences and be ready to take responsibility for them, since the instrument itself is built so as to allow only a very restricted set of applications, particularly if most of the predictors refer to non-manipulable personal characteristics. If we accept the limitations of such applications, we should also accept an interpretation of the Data Act in this case which prescribes removal of personal identification codes from the data. These data would then be useful only for internal methodological explorations within the scientific community.

A third type of consequence to be considered emerges in cases where our research findings in principle are value-neutral in the sense that the value relevance of findings is about equally large for different interested parties outside the scientific community. This would be the case, for instance, if a particular research finding could be utilized by employers and employees or perhaps by their respective organizations. I think about nuclear power as another example. It can be used both for peaceful purposes and as a terribly destructive means of war. Now the question arises: should research workers take responsibility for such findings? Is it not those who use such findings who should carry the responsibility rather than the research community? But even in cases like these I think a researcher must be careful about the structure of the context in which he works and in which his findings are applied.

Assume that we have a society where various interested parties are relatively balanced in terms of power and access to information on research findings. In such a society our formula could be applied without any question. The responsibility for application of research results should there be allocated to those who use or abuse research findings rather than to the researchers who produce such findings. But not uncommonly the researcher finds himself in a historical situation or in a society in which the balance of power and access to information about research findings are very one-sided. In that kind of situation the whole problem of responsibility for research findings is rather different. Under such circumstances the production of so-called value-neutral research findings will benefit only one side - the more powerful side, which also usually has more access to information about research findings. The effects of such an imbalance in the power structure should be taken into account in the moral calculus of any research worker involved.

May I say in conclusion that what I have discussed so far may seem to have little relevance for a discussion of the Data Act - except with regard to what I said concerning experimental subjects, perhaps. Nevertheless, some of the questions I have looked at are often introduced in the debate about the Data Act without any clear conception of the distribution of responsibilities in each particular case. Also Kerstin Anér, in spite of her intellectual brilliance, seems to allow herself to move quite freely from more general discussions of science policy and the responsibility of the research community to very specific demands for control over research without looking carefully into the problems of responsibility in each single case. There seems to be a tendency to use the fear and apprehension which understandably

are related to the uses and abuses of research findings to motivate more stringent controls and a more restricted freedom of research within the scientific community. This transition from a discussion of the uses and abuses of research to production of research should not be undertaken in such a lighthearted manner. I suspect there is an unholy alliance here between some politicians, civil servants, and, perhaps, leaders of the business community who have an interest in censoring certain kinds of research findings, and weaker groups of people in the community who tend to be apprehensive about any threat - particularly when it is blown up out of proportion by our evening newspapers. The weakest and the strongest in society may thus easily form an insuperable majority block directed against research.

Those in the research community who have a talent for popular writing have a great responsibility not only to make our Data Inspection Board understand some of the problems involved here, but also to convey to the mass public a clearer conception of the fact that a lot of research "intruding" into the lives of individual citizens in fact may do more to strengthen the position of the weak, the sick, the exposed and exploited than any Data Act can in effect do.

Karl-Olof Faxén:

I don't know if this is the right time to return to a question that I dealt with in my speach before lunch. How are we to set the boundary for the objectives of research in a particular concrete situation? When are the considerations for the individual's right to privacy of primary importance?

Those researchers who have expressed themselves on the subject have, as far as I understand, been reluctant to deal with this problem. The principle of complete freedom is asserted, in any case the right of the researcher to set the boundary himself is claimed. This attitude to the problem I consider to be unrealistic. I believe that the public reaction to a large extent can be derived from the fact that about 50,000 people are annually interviewed, even if the real reason for anxiety is not that research is made, but the consciousness of the large administrative records and the threat against personal privacy brought about by their existence. The public confuses this problem with the research problem.

I imagined that this was a problem that actually needed to be the object of debate at this conference, and that one wouldn't only illuminate the weaknesses of the Data Inspection Board and the difficulties to combine reviews of proposals in connection with research and, what is the main task of the Data Inspection Board, namely the supervision of the permanent records. As an example I will mention the strong reaction from the Swedish Employers' Confederation (SAF), when a much respected researcher made an inquiry among the employers on information concerning the political sympathies of their employees, concerning how many were liberals, social-democrats, communists and so on, according to the estimate of the employer. First of all, we considered this to be a question which the employers should refuse to answer. Secondly, we were extraordinarily upset by the fact that a researcher could even think of posing such a question in an inquiry. The researcher in question of course understood the situation when it was explained to him and omitted the question in the inquiry but I think the remarkable thing is

that in Sweden today a researcher at all can come to think of the idea of posing this kind of question in an inquiry to employers.

Concerning Öyen's discussion of SAF's view on the immigrant inquiry, I'd like to say that this was not at all based on political interests. We experienced no interest from employers in this respect. We were simply personally upset by the fact that inquiries with this aim were made among immigrant workers, which implied that these delicate questions would be posed, even to political refugees. Later on this personal emotional reaction was phrased in bureaucratic terms consistent with interpretations of the Data Act. I can understand that it was difficult to clearly understand from this text how honestly damned angry we were.

The Swedish Employers' Confederation has a strong interest in the greatest possible freedom in research. We don't support the establishment of any kind of public institution, which might restrict this freedom. SAF will always find itself in a subordinate position in relation to a public institution of this kind. Like others in the same position, we have an interest in maintaining the greatest possible freedom, and we realize that every prospective form of measures taken by society in this field may be directed against us.

At the same time, it is necessary to realize that if social research is to obtain the freedom that we consider important, there has to be some way of setting the boundary, in order to avoid too delicate investigations about people's sexual lives or family relations. If there is no such boundary set by researchers, all of them will have to suffer for the occasional mistakes in going too far, that are bound to occur.

Ulf Himmelstrand:

The metaphor used when you speak of weighing the value of personal integrity against the value of the freedom of research could be a quite misleading metaphor at times. It assumes that we are confronted with two contradictory interests, the interests of personal integrity and the interests of research, when in fact both of these values in a large number of cases could be seen as being placed on the same side of the scale and counterbalancing the values of economic growth and bureaucratic rationalization. Let us not be misled by these facile and misleading metaphors which often serve to make us see contradictions where there are none, thereby diverting our attention from much more severe and threatening contradictions and conflicts in our society. Again, I think that no discussion of this kind can be fruitfully pursued without an accurate analysis of the structure of power and conflict within any given society. I could speak for quite some time about that particular topic, but I hope you see my point anyway.

Claes-Göran Källner:[1]

I have a feeling that research workers are in some ways ignorant of how administration - or bureaucrats, if you like - work. If you pardon me, this may be true also for social scientists.

Authorities as well as courts have to make decisions. They are enjoined by the state to make up their minds, to consider and verify whether there are "special reasons" when the law prescribes that for

[1] This is an edited tape recording which the author has not had the opportunity to review.

a certain decision. However difficult it is, we have to make a decision. This is precisely the dilemma of every judge. Thus we cannot, though we would often like to, push the problems away and let others solve them.

In order to carry out the duty to make decisions with which the state has charged us, we must, of course, have as good a foundation as possible. Dr. Hammar complained earlier today that we used a submission procedure in his case concerning the investigation of immigrants in Stockholm and Södertälje. He presented several objections to this. The fact of the matter is that we are charged by the law with the responsibility of finding out the opinion among people affected by the project. And I ask whether it can be done in another, better way. The bureaucrat, if you want to call him that, is clearly enjoined to act in a certain way, precisely in order to ensure that the basis for the decision is as detailed and sound as possible. The system of letting all parties have their say did not come about by accident. One party should have the opportunity to express its opinion on the arguments of the other. If possible, one should ask those concerned economically, physically, emotionally, integrity-wise, or the like. You must not criticize us bureaucrats for acting this way. We have to, and there are good reasons for it.

Already this morning I touched upon the question of how much university departments know about the responsibility we have for publicity, secrecy, and order. I am somewhat doubtful concerning the research workers' knowledge of the existing regulations. I could illustrate this by citing a few lines from a letter which has been discussed earlier today, namely the letter from the Swedish Association of Sociologists to the Data Inspection Board. I think Pär-Erik Back has referred to it and Ulf Himmelstrand has signed it, so it must be permissible to cite from it. Without mentioning the source I refused this morning the assertion of this letter that the Data Inspection Board has not thought through the special situation of research. There is also something else in the letter which may be of some interest in this context: "Against this background[1] it can not be regarded as difficult to understand that more than one social scientist, both young and old, has been heard articulating the suspicion that the Data Inspection Board's control of the instruments of social science research in questionnaire investigations can be a way of diverting attention from more serious integrity problems which are difficult to handle and which are linked to the administrative individual files maintained by public authorities and commercial organs. A less sinister interpretation is that the Data Inspection Board and its board of directors have never really thought through the special situation of research, its needs and its both critical and innovative role." The intention with this is to reproduce this suspicion as hearsay without advocating it.

Still let me say that the first interpretation, the more sinister one, to me is an expression of a most fantastic conspiracy theory and I don't understand how one can get such an idea. But evidently it is possible, even though it is later retracted. I still interpret it as an expression of poor knowledge about how Swedish public authorities

---

[1] The Association of Sociologists refers, among other things, to the fact that the Data Inspection Board and the National Central Bureau of Statistics have wanted to examine possibly sensitive questions in the investigation of living conditions in the society. (Editor's note.)

work. Then I might add that the Data Inspection Board has certainly not interfered with public authorities and commercial registers, on the contrary. We have at least as big controversies with the government and the central authorities, as you know, as we do with research. It is my opinion that, by and large, research is not at all as difficult a problem for the Data Inspection Board as are the public authorities and the commercial interests. This morning I tried to express the view that the integrity problems of research are really not so great and will be still smaller under the condition expressed by Himmelstrand earlier today. I share his hope that the coming revision of the Data Act will lead to a reduction of the integrity problem of research.

Since I am referring to Himmelstrand anyway, perhaps I may continue and comment on his previous remarks. According to my notes he said that the responsibility for the results of research could not be put on the research workers but on the politicians and the bureaucrats. I interpret this to mean that it is not given that it is the research workers who are responsible for how the research results are used because then they are suppressed, etc., etc. This is possible to say, even if it is not undebatable.

What I have meant is rather that the research workers must be responsible for their methods to the extent that these make far-reaching intrusions into the private, e.g. economic, conditions of individuals. We heard Faxén and perhaps above all Wohlin touch upon this problem. This morning when I appealed to the research workers to take responsibility themselves for the ethical evaluation on account of their knowledge and broadmindedness, I was thinking particularly of the methods.

Finally I want to comment on Öyen's assertion that the interest in personal integrity is something characteristic of a conservative power elite. He did not use those words, but I understood the statement to be roughly that. However, that does not correspond to the experience we have had in Sweden. When I say this, I refer to both letters and telephone calls which the Data Inspection Board has received. These come from all groups in the population. Also I can refer to the obvious interest shown by the large trade union organizations and the measures they have taken in this area. I presume that Öyen does not consider them part of the conservative power elite.

Edmund Rapaport:

Sune Åkerman has broached the subject of de-identification of primary source material. I would like to say something about the background to this approach. Discussions about de-identification as a method of protecting individuals against violations of their right to privacy began in the fall of 1974 in connection with a decision by the Data Inspection Board concerning a certain, rather important investigation that the Central Bureau of Statistics (SCB) was engaged in. According to the permit issued to us by the Data Inspection Board, our use of material gained by access to data banks was conditioned on that we would de-identify it. This we found inappropriate and consequently we appealed their decision, which lead to the government ordering that we temporarily be issued a permit relieving us of the obligation of de-identification. Their decision does not represent a final position on this subject by the government. Rather, in part deferred final judgement until an planned commission could investigate the question in general.

The commission has just about concluded its investigation and a report will soon be published. Without forestalling the accounting for this investigation I would like to mention a few common aspects on these problems.

De-identification, as well as a couple of other measures which were examined by the commission, namely, destruction and the ensiffering of primary source statistical data, should be viewed here as methods for protecting individuals against what is called violations of their right to privacy. De-identification implies a considerable limitation on the possibility of using the material in the future; destruction totally prevents its future use. What, then, are the kind of needs and situations which must be considered when trying to judge these and similar measures? We in Sweden have for a long time had a general archive principle which in principle states that most existing material, at least with regards to that which belongs to the public sector, is to be saved primarily for the use of future researchers. This principle is quite well established, although some modifications due to economic considerations and concern over the individual's right to privacy are being considered. Today's discussion shows that it is no longer sufficient to make some general reference to the needs of research and to rely upon the general archive principle. In the debate today, many good points have been made as examples of arguments stressing the researcher's need of data. The way in which we have chosen to argue here demonstrates we have seen the need to motivate the nature of our research. Researchers' need of data is no longer an absolute need that does not need to be questioned, rather it is something that has to be motivated and defended. The question then becomes: what is it that weighs in balance against this need. The individual's right to privacy, which is what today's discussion centers on, is unfortunately a difficult concept to define and not much has yet been done to elucidate it. Presumedly, it is a question of certain interests which those persons, about which information is kept in an identifiable way, are assumed to have. Almost by definition, one can note, that the registering of material which is identifiable is not without risk to the associated individual. What the nature of these risks are, and how large they are for the individual is a question which has to be considered in the light of different possible perspectives concerning the future. In other words, a subtle analysis is required. The justification for research is, I believe, deeply felt by the public at large, and not just by some power elite. The large resources expended by society on research is a conceiving proof for this belief.

I remember well how we in connection with the discussions accompanying the enactment of the Data Act and its applications by SCB had reason to impress upon many researchers that they had showed absolutely too little interest for the problems that had occurred. I can now state to my pleasure that this interest is now present and that the discussions are in full swing. The interests which researchers represent are enormously important. The difficulties that exist here lies in the fact that the perspective of research is so illusive that it is often difficult to document, for example, the need of a certain collection of data in 20 or 50 years. It is here that a large part of the problems exist and the difficulties depend in large measure, as well as I can understand, in that we are unable to foresee the needs that will materialize. Nevertheless, one does not take care of the situation by merely stating that research must be given an absolute freedom and should not be forced

to document or give an account of its needs. This will not work because people's different interests are involved. Here one and the same individual can very well represent both the interest of research and other interests as well. It is not necessarily different people or different groups which stand opposed to each other.

I have the feeling that it is with research as is often said of democracy, that one must continuously reconquer it. Research must continuously fight for its freedom and the problems are so complicated that it is difficult to find general, final solutions; rather what is needed here is continuous picking discussions and the subtle weighing of different interests. In today's situation, the right to personal privacy is an important factor that researchers ought to take into consideration. But the actual concept is, as said, obviously unclear and it has to cover many and conflicting interests. Sometime it has also become a cover of personal opinions without reason - this is one of the risks in the situation. My hope is that this discussion as well as future discussions and investigations can lead to a more modulated view and a more structured perspective on these problems.

I would like to take the opportunity to refer to another proportion of this discussion which perhaps does not really belong to that which I have spoken of previously, but which was brought up by Lars Wohlin from the Industrial Institute for Economic and Social Research during the afternoon session. He is not here now, but I can perhaps anyway bring it up. He directed as well as I can understand a rather severe criticism against the Central Bureau of Statistics for its niggardlium toward researchers with respect to industrial data. The criticism is illustrated in part by the fact that even such simple statistics as the number of employed - an important variable for research - could not be obtained, and also in part by the fact that one strictly enforces the current twenty year secrecy period. He was of the opinion that industrial data become antiquated very quickly and that firms would not have anything against allowing data which were 5 to 10 years old to be used more freely. I'm doubtful with respect to these claims and I will explain why. But first I want to state quite generally that the Central Bureau of Statistics, for reasons that I don't think need further explanation, is extraordinarily restrictive with the releasing of primary data and is enormously anxious about maintaining good relationships with their data sources in precisely this respect. But when it concerns the number of employed the SCB has on several occasions suggested that this information be made generally available and the criticism from the Industrial Institute for Economic and Social Research seems to be unjustified when by rights they should have turned to their directors who on one occasion after another have refused to accept the suggestion that these statistics be made open. With regards to the time period for required secrecy of data on firms, the Committee for Legislation of Publicity and Secrecy recommended a secrecy period of between 30 to 70 years for statistical data. With respect to data on firms the SCB stated in its reply that the present limit of 20 years was completely sufficient. To my knowledge none of the organizations that represent the interests of industry has suggested a shorter period than that suggested by the committee. Quite generally I stand askance to the notion of equating personal privacy, that is the protection of physical persons, with that protection one wants to give to firms, even though there is a grey area for those firms, for which it is difficult to distinguish between the owner of the firm and the firm. Otherwise the situation is completely

different and one should, by my understanding, not try to transfer the positive connotation associated with the notion of the personal right to privacy, to another issue to which it does not belong as far as I can see.

Ulf Himmelstrand:

To Mr. Källner I would like to say this. We know very well that it is the job of bureaucrats to make decisions in accordance with valid law, and that you must try to find the best possible basis for such decision-making. In using the term "bureaucrat" I am not necessarily using it in a negative sense. One of the greatest of sociologists, Max Weber, has analyzed bureaucracy and the role of bureaucrats in ways which are not at all completely negative but which point out the important role of bureaucracy in modern society. Nevertheless, those of us who do not hold positions in the bureaucracy, have other tasks as citizens and private individuals. One citizen obligation and right is to critically reconsider laws already in the books, and to make demands and suggestions for changes of such laws. Laws are not given by Our Lord in Heaven. Laws are made by legislators elected by the people, and laws can be changed. We are thus entitled to discuss whether certain laws are justified or not, and that is what we are doing. Of course I do understand that until a law is changed civil servants in public administration must apply the law as it stands - even though there is some leeway, you must admit, in how laws are interpreted. I can even understand that a public servant at times may feel a bit uneasy about discussing the value and effects of a law which he has to apply every day, but the rest of us certainly are entitled to carry out such a critical evaluation, particularly since this particular law was worked out without a proper consideration for the role of research. The ethical problems of research mentioned by Mr. Källner this morning are also important in this context, but I can assure you that these problems are far from new in the scientific community. We have discussed these problems for decades, and what Mr. Källner said this morning thus is more or less commonplace to most of us who work in the scientific community. The important problem here, however, is to discuss how these ethical considerations are to be interpreted and applied in law and practice at different junctures of the relationship between the research community and the rest of society. This is what we are trying to do here.

Pär-Erik Back:

I want to comment on two issues which have been touched upon here tonight and, in addition, take up an issue which has not yet been raised. First the assertion that the scientists do not understand the ways in which bureaucrats and civil service departments work. It is true that the Data Inspection Board has to keep within certain rules, directives, and laws. But I am not willing entirely to retract the criticism which I have levied for many years precisely at the Data Inspection Board. It is a fact that the civil service departments have a considerable degree of freedom. They can behave in different ways within the given rules, and it is my opinion that the Data Inspection Board, perhaps especially in the past, behaved in a more unfriendly way towards research than necessary.

The other comment concerns Hammar's investigation. Tomas Hammar is not here and cannot defend himself against the criticism to which he has been subjected. As a matter of fact, I have a certain insight into the matter, since I was an expert in the Bank of Sweden Tercentenary Foundation (Riksbanksfonden) and observed the project for several years before it became a matter for the Data Inspection Board. Karl-Olof Faxén says that there was no criticism of the project based on political interests. Instead, it was simply a matter of people immediately becoming irritated at the questions contained in the questionnaires.

The observation is interesting because then we can only acknowledge that here is a case where there is a difference of opinion. This whole project was subjected to an extraordinarily thorough review and discussion concerning its design, question by question. The project was gradually changed and finally became one of the most interesting social science projects which we have had in several decades. For my own part I did not find the questions asked particularly remarkable. Nor did the immigrant groups concerned find them troublesome. Director-general Källner said that the questions were remitted to the respective groups concerned and wondered what was wrong with that. How else should the Data Inspection Board act, given that it has to obtain material for evaluation? But the mistake was, as Hammar himself pointed out, that it was not the most important material which was considered. As far as the scientific relevance was concerned, decisive regard was paid to the points of view of interest groups in the labor market and which, in my opinion, are absolutely untenable. That is the essential point which has to be made.

I then come to the third point which I wanted to raise here. It concerns the 11th paragraph of the Swedish Data Act and its rules on personal information to be used for automatic data processing abroad. Here is one of the most unfortunate things which has happened the last few years. During the 1960's we had an interesting methodological development in many areas of social science in regard to international comparative studies. There was also growing cooperation in Europe among young scientists. There were several projects started in which 5, 6, or 8 countries were represented. The idea was to use similar data from the various countries, and much planning was carried out. All that is now almost completely wasted: Sweden is left out here. We are practically an underdeveloped country. If research groups in other countries find out that Sweden is to be included and that the Data Act even looms at the horizon, they simply say no thanks and goodbye. This creates a lot of trouble for us, and it would be interesting to hear to what extent there is any reason for hope in this regard. We are living in an era of internationalization and cooperation across national boundaries. But on this point there seems to be a kind of neo-nationalism. For a scientist, especially a social scientist, national boundaries are often no more holy than the old county lines.

Tore Dalenius:

I want to draw your attention to three topics about which I feel very strongly:

i) researchers' unsatisfactory relations with the public and with members of Parliament;

ii) the tendency of many champions of privacy protection to discuss the privacy issues in terms of "capability" rather than "intentions"; and

iii) their related tendency to depict the computer, rather than people, as the villain.

I will take up these topics in turn.

1. I think that we researchers seriously overestimate the appreciation that the public and the members of Parliament have of research. Do we "market" our product well? My answer is NO.

I want to suggest that some social scientist make an analysis of the reports submitted to the Swedish Council for Social Science Research on projects carried out. Mr. Bruhn-Möller knows how critical I am of the quality of this reporting. If a significant improvement is not made we may experience a parallel to what has taken place in the United States. Senator Mike Mansfield and others have focused attention on some projects in a derisive way which is likely to make tax-payers (and other senators) reluctant to support research. I want to add that in my opinion some researchers have made it far too easy for critics to ridicule their research; this problem is, incidentally, not unknown in Sweden either.

2. My second point concerns the regrettable tendency to frame the discussion of privacy protection in terms of the great capability provided by today's information technology of invading privacy rather than in terms of our intentions in this area. As an example which (I hope) is extreme, I refer you to the discussion of the contingency of a military invasion of Sweden. The discussion of this specific case is in many important respects a parallel to the discussion of international security, and this is a parallel which may lead us astray.

3. My third point is related to the second one but deserves to be discussed in its own right. It concerns the role of the computer in the debate about privacy and the need for privacy protection. In my firm opinion, the debate has gone wrong, partly due to the great emphasis on the capability of the computer in the realm of data processing (including linkage of records, data storage, etc.). Against the background of the preponderance of this argument in the debate in Sweden in recent years, it should be no surprise that the Swedish Data Act is geared towards computerized records (in the words of the Act: "...files, records or other items of information being maintained by data processing...", as translated in P.G. Vinge: Swedish Data Act, Sveriges Industriförbund, Stockholm, 1973); it does not apply to files, etc. that are maintained manually! From a technical point of view, this seems to me to be unsatisfactory: privacy protection can be achieved more easily and at lower cost in a computerized information system than in a manual system. It is worth repeating Senator Sam J. Ervin's thought-provoking words: "The threat to privacy comes from men, not machines."

A conceivable consequence of the Swedish Data Act is that researchers may choose to use manual systems rather than computerized systems, thus enhancing both the cost of their research and the risks of invasion of privacy.

Barbro Westerholm:

It has been our experience, in our study in the county of Jämtland in which we are recording individual drug purchases since 1968 for a part of the population, that the subjects have few objections to the recording of such data. I think their positive attitude is due to the fact that they were informed very early on about the purpose of the registration, who has access to the data, and the value they themselves could get from the registration.

The other question concerns international collaboration in research projects with individual data. I would welcome a more precise description of what the difficulties are and why the data cannot be handled in an unidentified form. The National Board of Health and Welfare is involved in international collaborative research on adverse drug reactions and malformations. Here it has been possible to work without individual data (although at first we thought this would be difficult). There may be solutions to the problem, and we should discuss them before it is concluded that the Data Act hampers research.

Ole Engberg

In this discussion I feel like a technocrat among social researchers. Also, I am present here as an observer, as we have not yet had much public discussion on the integrity issue in Denmark. From this doubly isolated point of view I want to add four comments on the problems being discussed here tonight.

1. The background of the problem

I agree with Öyen that a power-struggle lies behind the integrity issue: who should have access and who should be denied access to information and the facilities to use it? Who decides what is relevant for me? And on what basis?

As we have witnessed, the answers to these questions are difficult to find and will vary with local conditions even inside Scandinavia - and inside Sweden.

And a further complication: the only power that is to gain from the answers is the third largest industry in the world - the computer industry. They need our answers to sell more and they will use our answers to sell more. Can we control the technological and commercial development of computer usage just by setting up "Admittance prohibited" signs? A sign you can always get around. Databanks in other countries (with other more lenient data laws) are an example.

2. On law-making

This brought the international facet of the problem into the picture. Maybe we have to look at use and misuse of information - produced by private industry and public administration - as on that of all other consumer goods. Take the pollution/environment problems. Here we find two competing schools of thought: the nurses who want to protect us from all bad by setting up the "Danger and prohibited" signs and the insurance-people who say: "The producer must pay all damages - that will make him extra careful". This may work for detergents or tobacco. But for automobiles? where the car may be perfect and only the driver crazy!

My observation is that the many groups represented here are trying to solve problems created by a new technology by very traditional and

unimaginative methods and with a narrow horizon. Alas, I have no constructive models for trying out new ideas.

3. A new specialist group is interested

Here we are, social science researchers, scholars, statistical experts, men of law and medicine and probably a few bureaucrats (in the positive Weber-sense). Each has his own interests, norms and ethics.

Only modestly represented here is the profession that designs the systems and keeps them running: the edp system-people. They are the ones who still make most of the decisions: What can be done?/What cannot be done?/What will not be done?

My postulate: More and more of the edp specialists feel this responsibility and are looking for ways to handle it. Many professionals believe that rules of ethics should be established (the British Computer Society has a proposal). IFIP (International Federation of Information Processing) has just established a new task group (no. 9) called "Computers and Society". This group especially asks social science researchers for help: Where is our technology bringing us?" or still better: "How should we try to form our society?" It could be hoped, that something useful could come out of such a cooperation.

4. Ten years from now

Last but not least: I envy Sweden this debate on what data should and could be used for. It will focus on the technical problems of using data for secondary analysis.

In Denmark it has been suggested that all sensitive data should be destroyed after five years - then the data cannot be misused.

The technical answer to this is that destruction normally is not necessary. After five years nobody can find the tapes. If they are found by chance they cannot be read - and if that obstacle is finally overcome, and the magnetic spots are changed to data, nobody can make the data into information whose statistical significance can be evaluated.

In other words, until new "datadocumentation norms" - not to be mistaken for system and program documentation - have been developed, adapted and integrated in daily work, you will only have a "misuse-problem" five years from now if you archive not only the tape but also the "edp-specialist" who made the job and the administrator who ordered the job and knows where the data came from - sometimes even more than one administrator has to be archived.

Anders Klevmarken:

I would like to support Tore Dalenius' plea for explicit examples of threats to someone's personal integrity. What kind of threats are there? Would it be possible for some research institute to survey the nature of different kinds of threats, who is threatened and also how seriously these threats are perceived? Would it be possible to measure the intensity of a threat or the probability that a threat is carried out? Social scientists might be able to answer these questions and in this way bring some facts into a debate which is now very emotional.

Another contribution of ours might be to ask ourselves, as consumers of data, for what purposes we necessarily need individual data.

It is sometimes possible to use transformed data i.e. averages, although at a loss of information. For instance, in my own research about the relationship between education, age and earnings I needed estimates of age/earning profiles. Ideally these profiles should be estimated from longitudinal data, panel data, but I had to use successive cross-sections of averages. In this particular case the assumptions made seemed justifiable and the loss in efficiency was not harmful. In general, we should investigate what losses in quality and efficiency are acceptable and try to give measures.

For some purposes, for instance in simulation studies, one may even deal with data on "synthetic" individuals, i.e. individuals who have been made up from pieces of information from individuals actually observed. Each synthetic individual thus possesses characteristics which do not correspond exactly to any observed individual, but in principle there might exist an individual with such characteristics. Databanks of synthetic individuals would thus not conflict with personal integrity.

Finally, a comment on international joint projects and the Data Act. This issue was brought up by Barbro Westerholm who would welcome a more precise description of the difficulties. Sometimes we find it more efficient to process data abroad because we may not have the necessary computer capacity or, more often, the relevant software. We may also have to consult experts abroad. In most instances it is probably possible to meet the requirements of the Data Act and process data in Sweden or, alternatively, process data in unidentified form, but at a higher cost and at a time loss.

Ingemar Fägerlind:

As a colleague of Prof. Torsten Husén at the Institute for the Study of International Problems in Education I have been involved in two studies where large amounts of data have been stored. The first one is the longitudinal Malmö study which began in 1938, in which 1,500 persons and their children have been followed up to the middle of the 1970s. The second is the IEA (International Association for the Evaluation of Educational Achievement) study where cognitive and attitudinal data from about twenty countries were collected.

The experience of myself and my colleagues in working with these two studies leads me to conclude that some of the speakers at this symposium are reacting against imaginary problems. In working with the longitudinal data since 1963, I have sometimes been frightened to find how easy it was to obtain personal data. When the data registers were computerized this task became even easier. As a citizen I think it is important that there should be some checking of the nature of the data collected. I also think that it is important for researchers to give careful attention to the manner in which the data are stored and utilized.

On the other hand, it is also important to relate that in the Malmö study, which began in 1938 and has continued with data from three generations, there has never been any complaint by the participants that the researchers have misused the data.

With regard to international research, I consider that Prof. Back was too negative. Of course there are problems in studies using personal data. In the data bank for the IEA study, there are no names or

addresses for the respondents whose data are stored. From the data thus available many valuable cross-national studies have been performed, which leads us to hope that this type of study will be continued in the future.

It was mentioned in this discussion that there is a constant fight between different groups in our society. I agree, and when it comes to social science research I have noticed a battle between the researchers and the "Ämbetsverk" (government agencies). It is not only the National Central Bureau of Statistics that wants to perform all the research on their own data within their own institution. You will notice this tendency also in other "Ämbetsverk" which do not want to make their data available to outside researchers. I find this to be an important problem.

Dr. Anér suggested that researchers should devote more effort to developing methods for cooperation between researchers and participants in research projects. I agree that this is very important for the future and should be discussed further at this conference. This is just as important as the need for researchers and administrators to meet, as we have done here.

3.7 THE INTERESTS OF THE SWEDISH DATA ACT AND THE PRODUCTION OF STATISTICS - AN ATTEMPT AT ANALYSIS

by

Edmund Rapaport

1. The issue

In this lecture I intend to try to systematize and analyze more closely the relationship of statistics production to the interests protected by the Swedish Data Act as interpreted by the Swedish Data Inspection Board. The review is based on the practice starting to be established by the various decisions handed down by the Data Inspection Board and by the changes in these decisions instituted in some cases by the Government.

The interests which the new Swedish data legislation are to protect consequently provide the starting point. In Swedish legislation and discussion these interests have been summed up in the term "personal integrity". However this concept is defined - a difficult question I will discuss later - it obviously covers people's defensive interests. It is therefore a question of protecting the individual. Regarded in that way, protection of the interests of the individual can be roughly translated as protection against different kinds of risks. The concept of risk is not used in any accepted statistical sense, as the occurrences that constitute these risks can hardly be described in probability terms.

It should at once be emphasized that other interests may run counter to those which the Data Act is intended to protect. The possibility of conflicting interests in this field is well known, and both the Act itself and its preparatory study recognize and treat in some detail at least some of these conflicts. With reference to the production of statistics the conflict might be preliminarily formulated as follows.

A modern postindustrial society is heavily dependent for rational management and development on the supply of relevant, correct and sufficient statistical information. In a wide perspective, general access to statistics is fundamental to the public debate which in turn is fundamental to a living and working democracy. In society's decision processes, statistics supply essential materials. Access to statistics is a necessity in other important cases too, e.g. in social science research. If this need for statistics collides with the notion of personal integrity, which is to be given preference?

No general answer can be given to this question. Various types of personal integrity risks have to be balanced against various types of community losses caused by a reduced amount of information. This balancing of various types of risks and losses can be done most concretely in an actual situation, viz. when a given statistical survey is up for consideration under the Data Act. However, in my attempt at analysis it is suitable and necessary to generalize on a certain level of abstraction. It should be noted, too, that the interest of protecting personal integrity is not necessarily opposed to the interest of producing statistics. On the contrary, these two interests coincide to a great extent.

2. The risk of unauthorized access to primary statistical data during normal social conditions

The Data Act and its application to the production of statistics as regards protection against encroachment on personal integrity means that data about private individuals (primary data) are not to be used for purposes other than the production of those aggregated anonymous data that constitute statistics. This circumstance and the reasons for it are so well known that further comments are unnecessary. It is sufficient to note that the interests of personal integrity in this case coincide entirely with the interests of the statistics producers, and the question is mainly one of protecting the primary data against unauthorized access. In Sweden, the Data Inspection Board exercises supervisory jurisdiction in order to ensure that protection is satisfactorily arranged, and this supervision constitutes valuable support of the efforts of the National Central Bureau of Statistics (SCB). As part of these efforts, the conventional safeguarding of buildings, storerooms, transports, etc. has been examined and, wherever necessary, strengthened. The protection of computer installations and of computer operations in general has also been reviewed and improved. This review work has been performed in contact with the Data Inspection Board and the Swedish Police Board, and in close cooperation with the government consulting agency in this field, Statskonsult Ltd. The risk of unauthorized access in these (so to speak) conventional ways has probably been reduced by these measures as far as can be justified when the cost of protection is balanced against the probability of assault attempts. It is part of the picture that there is no known case in the history of the SCB of a successful attempt to gain unauthorized access to primary data.

Improved protection against unauthorized access can also be effected through measures directed towards the primary material itself. Three kinds of such measures, destruction, de-identification and encoding of statistical primary material, are at present the subject of a special investigation by the SCB in consultation with the Data Inspection Board and the Swedish Record Office. This investigation, which was started slightly more than a year ago and whose findings will be presented shortly, originated as a result of directions about de-identification issued by the Data Inspection Board.

The directions in this matter were the removal of all identification terms shortly after the data collection in order to provide additional security for the individuals who had furnished the information. The SCB objected these directions on two grounds. First, the de-identification would make longitudinal studies on the collected material impossible, and this would collide with important research interests. Secondly, from a legal point of view the collected material would probably be deprived of its confidential classification, while the possibilities of so-called backstairs identification, viz. the identification of private individuals by means of the remaining data, would still remain. The Government suspended the decision about de-identification by the Data Inspection Board, pending the results of the above-mentioned investigation. This investigation also discusses the suitability of entirely destroying the statistical primary material in certain cases, which naturally would give the most far-reaching protection against unauthorized access, but at the same time render future processing quite impossible. With de-identification, on the other hand, continued statistical processing remains possible, although without any chance of

supplementing the material with new variables, as mentioned above, of longitudinal studies. The third type of measure discussed in the investigation, the encoding of identification terms and perhaps of other data as well, is a protective measure, having, in principle, no effect on the possibility of continued processing or supplementation of the material.

The conventional protective measures against unauthorized access and the special measures now being investigated seem quite sufficient to provide perfectly satisfactory protection under normal conditions. Probably it can also be assumed that the special measures of destruction and de-identification of the primary material, which in principle might collide with statistics interests, in reality hardly need become controversial. Probably they will be contemplated only in very special cases, when both the susceptibility of the material and the degree of interest in preserving it for the future will produce a concerted opinion to employ one or the other of these measures. The normal, basic safeguarding will usually be sufficient and can if necessary be reinforced by e.g. encoding.

Protecting data about private individuals from unauthorized access must also include the prevention of disclosure through published statistical results; i.e. published statistics must de facto remain anonymous. Intensive development work has taken place in this field, particularly within the SCB. It might be said that Swedish statistics production in this respect is well in hand and that there exist adequate methods both to discover and to prevent disclosures. The discussion has been extended to so-called probability disclosure, in which field research and methods development work is going on. The problem of disclosure in tables has also long been discussed by the Data Inspection Board. The directions issued for various statistical surveys regularly include a stipulation that the statistical results are not to expose private individuals. For obvious reasons, the interpretation and application of this stipulation have been left to the specialist responsible for the publication.

The legal framework for protection against unauthorized access is naturally of great interest in this connection. By this I mean the definition of the protected field in relation to the general Swedish principle of public accessibility, in other words the secrecy regulations. The attitude of the Data Inspection Board as revealed in various decisions pertaining to the SCB seems to some extent influenced by the view that, at the present time, secrecy regulations are not entirely satisfactory. As the subject is too complicated to be discussed now, I will limit myself to stating that essential improvements and elucidations of the secrecy regulations are expected to occur in the near future.

This review of the interests and various types of risks during normal conditions may in conclusion be said to indicate that there is no fundamental conflict between personal integrity and the need for statistical information. The prospects also indicate ample possibilities of adjusting the protection level to future personal integrity risks without any appreciable conflicts with the risk of public information loss. A review of interests and risk situations during conditions other than normal is consequently of greater interest.

3. The risk of unauthorized access to primary statistical data during conditions of war and danger of war

The risks of unauthorized access discussed in the previous section were termed risks under normal social conditions. The types of risk discussed below are characterized by occurring under conditions other than those of an orderly development of a democracy, e.g. conditions of war and danger of war. In Sweden the protective measures to be instituted in such situations and their preliminaries are constitutionally regulated. These measures also cover the information in traditional documents and other data media. It is, inter alia, the duty of each government authority to have emergency plans for these situations, including the measures removal and destruction of information media. The objective seems to be to avoid loss of information in military operations and in addition to avoid its capture by the enemy. The two interests might of course collide with each other in the same way as they might collide in other fields than information. It is to be assumed, however, that such information about individuals is to be destroyed which can be used in the hands of an enemy power to the detriment of the persons concerned. In other words, in the difficult balancing of interests, solicitude for the individual takes preference over general, long-term interests. Such information is predominantly found in the countless number of so-called administrative registers on individuals, but the above pertains to collections of data for statistical purposes as well.

An interesting question in this connection is whether legislation intended to protect personal integrity, such as the Swedish Data Act, should consider the risks in war and danger of war. In regard to the interpretation of the Data Act the Swedish Government, in a much-noted appeal not concerning statistics, has taken the attitude that the treatment of such questions falls outside the Act's area of application. The Government's formulation deserves to be quoted: "The Data Inspection Board gives three main reasons for the decision against which the appeal has been made. As first regards the risks in political upheavals or war, the relevant questions pertaining to e.g. safeguarding, removal and destruction of registers of individuals fall primarily within the jurisdiction of authorities other than the Data Inspection Board. By means of the Data Act and the work of the Data Inspection Board it is possible to assess the different registers of individuals that exist and the new ones that appear."

Another risk in a situation of war danger arises if the government at that time were to demand access to primary statistical material of special interest at the crisis. A well-known example of such a situation is the attempt in the United States during World War II to utilize population census material to find persons of Japanese descent. This attempt failed. It seems very difficult to anticipate future situations similar to the one just mentioned. If, against all surmises, such a situation were again to arise in some country at some time, it is to be assumed that the statistics authorities, as in the United States, would have sufficient courage and strength to defend their material against such attempts and that other bodies in the country will not succumb to the temptation of making short-term gains at the price of great damage to their fundamental long-term interests.

In conclusion I would make the assessment that in Sweden the risks possibly inherent in the production of statistics and the storage of data in times of war and danger of war are already adequately covered

and consequently do not require any special treatment or further proceedings within the framework of the protection of personal integrity.

4. Personal integrity risks in the case of national upheavel, etc.

The discussion about personal integrity has also dealt with the possibility of unauthorized access to and use of data about individuals in the case of a coup d'état or similar upheaval, when the usual rules about inter alia the use of primary statistical data have ceased to function. In regard to such risks, too, it is doubtful whether they naturally belong to the field defined by the concept of personal integrity. Irrespective of this, it can be ascertained that the only satisfactory protection against the risks now in question is the total or partial destruction of material that can be put to improper uses.

The first question, then, is whether filed primary statistical material really presents any risks for private individuals under such extraordinary conditions. Available historical experience hardly gives any guidance in an evaluation, except possibly hypothetically in regard to very special limited materials. In general it can be shown that primary statistical materials are often distributed on rather small samples. As far as census registers are concerned the information is usually rather limited and usually available in more detail in administrative registers kept by other bodies. Also, the information is often "old". It is consequently hard to see statistical primary materials as presenting any great risk, even during extraordinary conditions. It is quite likely that they rapidly lose their value for purposes other than statistics and research, even during the conditions now discussed.

However, the wider and more difficult question is whether it is at all meaningful to limit the review to primary statistical materials when discussing society's preparedness and defence against coups, as seems to be the case in the Swedish general debate at present. To me, that is beginning at the wrong end. A rational preparedness plan must cover not only the problem about information storage and primarily information storage which has nothing to do with statistics, but also a string of exceedingly difficult questions about society's organization, the power of various bodies, attitudes, etc.

Thus, there is the risk that the type of preparedness plan which begins and perhaps exclusively deals with primary statistical material might be the cause of great information losses without essentially strengthening the protection of the individual. In special cases it might be quite reasonable to destroy, partially or totally, very delicate material. However, a concrete decision for a given primary material must be taken with a sense for distinctions and within the framework of a total assessment.

5. Fair play for the respondents

In Sweden the Data Act gives persons the opportunity to check whether the data recorded about themselves are correct. Incorrect data are to be corrected if the error is significant from the point of view of integrity. In the same way some of the instructions issued by the Data Inspection Board in accordance with the Data Act suggest an interpretation of the concept of personal integrity as of the individual's interest in being correctly treated. In regard to the SCB, this has resulted in directives about information to be given to the respondents, about their

rights during the data collection stage and about verifying the contents of the computer registers which form a processing link in the production of statistics.

According to the directives regularly issued by the Data Inspection Board, the information to the respondents is to be explicit primarily with regard to the rules that govern the collection of data and with regard to whether or not participation is voluntary. The respondents are also to be informed about the confidentiality classification of the collected data and are to be told that collection of supplementary data from other sources might occur and which these sources might be. Directives have also been issued stating the right of an interviewed person to discontinue an interview or to change information given previously in the interview.

For the SCB these directives represented no changes in subject-matter in respect to the information given to the respondents and their rights, but did involve certain adjustments in the presentation. A detailed and correct presentation of the relevant aspects of a survey at the time of the data collection has always been considered important by the SCB, and no conflict of interest vis-ã-vis the Data Act in this respect consequently exists.

The Data Inspection Board has taken the attitude that an individual is in principle to control the information about himself, and this has resulted in an injunction against so-called indirect interviews. Indirect interviews, which are defined as the collection of information about one person from another, occur only sparingly in SCB surveys. However, in the labour force surveys they are considered permissible in view of inter alia the simple contents of the interview, and they form an important means for carrying out these surveys speedily without an excessive non-response rate. The Government revoked a decision by the Data Inspection Board and permitted the SCB to continue with indirect interviews, although with certain limitations on the circle of permitted respondents and on the contents of the interview.

Another directive of seemingly ethical purport, an injunction against imputations, is of great significance for Swedish statistics production. This directive was also issued in the decision concerning labour force surveys. Imputation is a procedure for attributing a value of a variable to a person or object on the basis of related information about this person (object).

Until last year a method for supplementing the information about non-respondents was used in the SCB labour force surveys, in which "twins" were found between respondents who had participated in the survey and non-respondents, and the values of the respondents were attributed to the corresponding non-respondents. This procedure was termed "total imputation" in the SCB report to the Data Inspection Board. In its decision the Data Inspection Board issued an injunction against imputations in labour force surveys. In its appeal to the Government the SCB distinguished between total and partial imputations. By partial imputations the SCB meant the imputation of a few non-reported variable values for a respondent who had otherwise participated. The SCB stated that the total imputations in the labour force surveys could easily be replaced by procedures approximately equivalent from the statistical point of view, while partial imputations were of great importance for the production of statistics in general. However, in the labour force surveys they occurred but very rarely. The government's decision meant inter alia an injunction against imputations. The contents of the SCB

appeal, together with the Government's explanatory statement of its decision, permit the interpretation that the Government has in principle vetoed total imputations but has not adopted a definite attitude to partial imputations. Formally the decision pertains only to the labour force surveys. It should be noted that the Data Inspection Board has indicated that it is not opposed to imputations per se, as long as they do not mean that "faked" data are recorded for identifiable persons during statistical processing. In other words, there are no objections to e.g. imputations on aggregated levels.

6. Conclusion

The compass of this lecture permits only a very summary analysis of the field of interest formed by the production of statistics on one hand and the interests expressed in demands for personal integrity on the other. The exposition is not only summary, it is also incomplete. For instance, there has been no mention of the great benefits statistics production may doubtless expect, at least in the long run, by an effective protection of personal integrity. Nor have the specific conditions for research created by the Data Act been discussed. My intention has mainly been to argue in favour of the possibility and advantage of discussing and analyzing the concepts of personal integrity and statistics production in terms of the interests and risks involved. I maintain that the discussion sofar carried on in Sweden, although extensive in itself, has to some extent neglected to systematize and concretize the problems; an urgent task, particularly in view of the approaching revision of the data legislation.

3.8 THE RIGHT OF PRIVACY AND THE NEED TO UNDERSTAND

by

Vincent P. Barabba

It's good to be here in Stockholm - not only because it's my first opportunity to visit Sweden, but because I wanted very much to take part in this symposium. The subject under discussion is one of deep interest to me personally, and to the entire statistical community of the United States.

In any society which places primary value on the rights of its citizens, the role of a data-gathering agency is naturally one which comes under close scrutiny. Statistical agencies in free societies are charged with two tasks - to provide the data which are increasingly vital to the decisionmaking process, and do this job without infringing upon the rights of individuals. I feel these tasks are not as much in conflict with each other as they might seem at first. I would add, however, the degree to which they are a problem is a function of the degree to which the average citizen perceives them to be in conflict.

Because all of us in this room realize the value of statistics to the functioning of society, we also share similar concerns. My purpose today is not to make direct comparisons between the approach taken by Sweden and that taken by the United States in an attempt to determine which is most proper. My purpose is to examine the problem in the context of the United States experience - and share with you how legislation has been shaped to guarantee the confidentiality of individually identifiable records and yet keep a free flow of statistics available to decisionmakers. Specifically, I want to tell you how this legislation affects the Census Bureau, the primary data collection unit in the quite decentralized U.S. Federal statistical system.

The original title of this talk was to have been "The Right of Privacy and the Need to Know". I changed it to "The Right of Privacy and the Need to Understand", because I see a different connotation between the words "know" and "understand" in the context of this problem. Surveillance systems need to know information about the individual. A statistical system needs to understand society as a whole. In a surveillance system, information about the individual always remains identifiable, regardless of the data format. In a statistical system, information about an individual is amalgamated with information about many other individuals. Personal identification is lost in the process of creating summary data.

The specter of government intrusion into the affairs of individual citizens in democratic societies has always been a highly emotional subject - mainly discussed in the framework of newspaper headlines. Events of recent years in the United States have brought the issue into sharp focus: military surveillance of civilians, wiretapping, the bugging of offices, and industrial and political espionage.

My files are crammed with clippings on the subject, with headlines such as:

"Reversing the Rush to 1984", or "Big Brother Society Feared". Such headlines reflect the fear that information gathered for legitimate purpose will be used later in a different context which could injure either the individual or his family.

This fear was summed up very well by the noted Soviet author Alexander Solzhenitsyn in his novel, Cancer Ward. He wrote:

> "As every man goes through life he fills in a number of forms for the record, each containing a number of questions. .... There are thus hundreds of little threads radiating from every man, millions of threads in all. If these threads were suddenly to become visible, the whole sky would look like a spiders web. .... They are not visible, they are not material, but every man is constantly aware of their existence. Each man, permanently aware of his own invisible threads, naturally develops a respect for the people who manipulate the threads."

This fear of the misuse of personal information is exaggerated by the popular image of the computer. That image often casts the computer in the role of a villain. It becomes the tool of the all-pervasive, yet unidentified "They". Whenever we find fault with some action of government, business, school, or any other segment of society, it's always "they" who did it, and increasingly the computer is blamed for making it possible.

Though every technological advancement does have the potential for concentrating power in the hands of those who control the equipment, this power can be diffused as that technology spreads - at least in a democratic society. For example, let's look at the computer. Is the computer only a tool of the powerful, whether it be big government or big business? My answer is no. I base my contention on the fact that as computer technology has advanced, computers have come increasingly within the grasp of those who want to use them. Inexpensive computer kits are available, desktop models are offered for sale, and entire systems are sold by firms replacing them with newer models. You don't even have to own a computer to be able to use one because of time-sharing. Additionally, the power of information can be diffused through the use of computers.

A good example is the network of Summary Tape Processing Centers established around the United States on local initiative. Summary statistical tapes not containing any individually identifiable records from the 1970 U.S. census were sold by the Census Bureau at cost to these centers, which use them to make special tabulations for data users in the public and private sectors.

Another factor is that the public generally over-estimates the abilities and the applications of computers - thanks in part to futuristic movies and television. Together with recent headlines, this image of the computer has led to vague fears of an ominous National Data Bank, which would store every facet of our personal lives for instant retrieval by any government agency which requested information.

For this fear to be realized in an open society - for a society to move that close to the nightmares described by Alexander Solzhenitsyn and by George Orwell in his 1984, these societies would have to abrogate not only current law, but their entire democratic tradition.

Let me underline this thought as far as the U.S. is concerned by quoting to you a portion of a paper written by Otis Dudley Duncan which concerned plans for the 1970 U.S. census:

> "... In this country we have proved that a statistical system can incorporate rigid safeguards of confidentiality. The institutionalization of these safeguards has proceeded to the point where it is inconceivable that they would break down, except in the catastrophic event of a breakdown in our whole system of institutions protecting the rights of the individual." Then he adds these key words: "In the case of such a catastrophe, my guess is that much more direct ways of infringing these rights would be found than that of making inappropriate use of statistical records secured ostensibly in confidence."

However, I would be glossing over the subject if I didn't say there is an inherent conflict in gathering data from individuals. That conflict is between the individual's right of privacy on the one hand, and, on the other, government's use of mandatory processes to obtain the information it needs for valid purposes.

Basic to this discussion is the question: what is the right of privacy? It is a very easy term to use, but a very difficult one to define. Legal and academic scholars in many nations have wrestled with the problem for the better part of a century. Obviously, privacy does not exist in an absolute sense, any more than freedom does. Oliver Wendell Holmes, a famous Chief Justice of the U.S. Supreme Court, included this thought in a celebrated opinion: "Freedom of speech does not include the freedom to yell 'fire' in a crowded theater."

Privacy, as freedom, has meaning only in the context of human society, and society changes as time passes. As society becomes more complex, it needs to know more about its composition in order to establish priorities and properly allocate its human, financial, and natural resources.

The right of privacy is often expressed as "the right to be left alone". But that concept is inconsistent with the individual's responsibility to society.

Each man, woman, and child in a society reaps benefits from being a member of that society. Of course, these benefits vary from place to place and within the subgroups of society. Yet the individual obviously derives benefits from dwelling among other people.

It is axiomatic hat we never get anything for nothing. What, then, is the trade-off when it comes to the individual and society? The obligations of an individual living in a highly complex, densely-populated industrial civilization are greater than any in history. Sometimes the price the individual pays is in money - such as taxes; in other cases, it is time, such as serving on a jury, a jail sentence, or duty in the armed forces when required. Sometimes it is establishing qualifications to do certain things - such as driving a car, or practicing certain occupations.

If we grant that we all operate in the context of human society, and that we have a responsibility to that society, we can arrive at a definition of the right of privacy along these lines: It is the right of the individual, to the extent possible, to control what information about himself he releases, to whom he releases it, and under what conditions.

All of which is a roundabout way of saying there is a right of privacy, but it is a right which may be circumscribed to allow the expression of other freedoms. Obviously, any limitation of our right of privacy must be made with extreme caution and only after careful consideration of the consequences.

Former United States Senator Sam Ervin - for many years a most respected Constitutional authority in Congress - once said: "Somewhere a balance must be struck between the individual's desire to keep silent and the government's need for information. If it is proved necessary to invade certain rights, clearly it is the Constitutional duty of Congress to establish precisely how and under what circumstances this may be done."

It occurred to me as I thought about this quote that there would be great difficulty in establishing the Census Bureau as it exists today with its present authority. If the administration asked Congress to create an agency with the power to decide what questions it would ask of the American public, the power to compel a response, and the power to impute characteristics for non-respondents - such a request would be drowned in a howl of protest from both the members of Congress and the media.

I believe the fact that the Census Bureau exists in its present form and has the authority I described is because a unique contract of trust with the American public has come into being as the Bureau evolved.

This contract of trust has several unique properties. It is based solidly in law, but its main force lies in intangible aspects such as tradition, practice, and professional pride. While it took many decades to build up, it is nonetheless fragile and could be easily damaged. Finally - and perhaps most important - even though it could be easily damaged, the entire weight of the U.S. statistical system rests squarely upon it.

We know that a census or a survey is only as good as the contract of trust with the people who answer the questions. If there were general public disbelief that we were keeping our word that their answers are confidential, these answers would not be as accurate, or given as willingly. Indeed, this situation would take place if the public felt that even the potential for such violation of their trust existed.

If this situation ever occurred on a large scale, the quality of the summary statistics would deteriorate. As a result, the United States would lose its main decisionmaking tool, and society would be the loser. But our experience over many years does not justify this concern.

In our evaluation of the 1970 Decennial census of Population we estimated an undercount of 2.5 percent of the population - an improved performance over 1960 and 1950. Of course, the census is mandatory. On the other hand, a study of five long-running household sample surveys shows very high voluntary response rates. The oldest and largest of these is the Current Population Survey, taken monthly of some 50,000 households. Outright refusal to take part in this survey stood at 1.5 percent in 1965 - and 2.2 percent in 1975. That is a remarkable record when you remember all that occurred which had the potential of increasing the mistrust of authority on the part of the average citizen during that decade.

The contract of trust which makes response rates such as these possible has been a long time in forming. Its foundation - and that of the census itself - lies in the U.S. Constitution, written in 1787. Article One stipulates that seats in the House of Representatives will be apportioned according to the distribution of the population, and that an enumeration of the population will occur at least every ten years. Beginning with the first census in 1790, that provision has been carried out, and forms the basis for representative government in the U.S.

Through 1830, confidentiality of private records was not a factor. In fact, federal marshals, who supervised the first five censuses, were required to post copies of the list of names in public places in their districts.

Between 1840 and 1870 there were no legal restrictions, but census takers in the field were instructed to treat all information they gathered as confidential. Those instructions became law in 1880. All enumerators took an oath not to disclose any personal information. Oddly, this requirement did not extend to their supervisors.

That loophole was closed for 1900 - when all census employees were made subject to a 500 dollar fine for violation of their oath.

In 1902 the permanent Census Bureau was established. It had become apparent that the need for statistics was a continuing one, and that the size of the job was such that a full time trained staff was much more efficient than setting up and disbanding a new organization every decade.

Up until 1910, census law required the Director to furnish on demand to governors of States or heads of local governments certain parts of an individual's return - name, age, sex, birth place, and race. The act for 1910 changed that wording to read that the Director could - at his discretion - furnish information for genealogical and other proper purposes.

The year 1910 also marked the start of another tradition - the presidential proclamation. The one issued then by President Taft told the American people their replies to census questions were to be used only to compile general statistical information, and that their answers were protected by law. In part, it read:

> " The census has nothing to do with taxation, with Army or jury service ... or with the enforcement of any National, State, or local law or ordinance, nor can any person be harmed in any way by furnishing the information required."

The current law under which the Census Bureau operates is Title 13 of the U.S. Code, most of which dates from 1929. This census law is very specific when it comes to personal information. It requires that information obtained from an individual be used only for statistical purposes. It also requires that published data be in such a form that it is not possible to identify an individual or a single business establishment. The law stipulates that no one other than sworn officers and employees may have access to individual information, and each census employee has signed an affidavit of nondisclosure to uphold the law. Each employee is officially reminded of this law twice a year.

The current law still has wording much like that of 1910, which allows the Director at his discretion to provide copies of individual information for genealogical and other proper purposes. The key word here is "discretion". Over the years, the application of this power has become restrictive rather than permissive.

Beginning immediately following the strengthening of Census law in 1929, a number of events occurred to strengthen the Bureau's operations under that law. In 1930, the Attorney General ruled that even the name and address of an individual collected during the census are confidential. His ruling touched upon two requests for personal information from census files - the Women's Bureau of the Department of Labor had asked for names, addresses, occupations, and status of employment of workers in Rochester, New York. These were to be used in connection with a planned survey to determine the economic effects on family and community life of women working in industry. The second request came from several sources and asked for the names and addresses of people who could not read or write. This was turned down even though the requests were for the purpose of aiding public education and the drive to eradicate illiteracy.

At about the same time, the Secretary of State asked the Census Bureau for data about individual farms in one county in the State of Washington. Clouds of sulphur dioxide gas from a smelter located across the border in Canada were causing extensive damage to crops in the U.S., and the matter had been handed over to an international tribunal. The Census Bureau refused to release the information, and the tribunal decided not to press the point. The reason? Because it would have caused the U.S. Government to breach a promise it had made to its citizens as well as to violate the census law.

Now we jump to 1942. It's hard to imagine now, but following the attack on Pearl Harbor, there was near hysteria about the Japanese - Americans living on the West Coast of the United States - emotion which led to one of the most embarrassing moments in U.S. history, the internment of large numbers of these loyal Americans. At the height of this feeling, the War Department requested that the Census Bureau supply the names, addresses, and ages of all persons of Japanese extraction living on the West Coast who were counted in the 1940 census.

The Census Bureau sent one of its officials to California to act as liaison. He, in effect, was conscripted by the War Department for the project. Considering the wartime emergency, this official could have asked the Bureau to comply with the request for names and addresses. Instead, he worked to get the job done by being responsive within the law as it was then understood.

The Bureau made special tabulations for the War Department for certain States, counties, and municipalities, but did not release individual names and addresses in conformance with census law.

In 1947, during the rising concern about possible foreign infiltration and sabotage, the attorney general requested information about certain individuals in Census records on behalf of the FBI. Again, the request was denied.

A loophole in the law turned up in a case in the early 1960's when the courts ruled that file copies of census forms retained in company files could be subpoenaed. This resulted in Congress amending the law to extend confidentiality to include even copies of census questionnaires.

Just last summer, IBM had requested a U.S. District Court judge to order that census questionnaires from firms selling computer products and services be admitted as evidence in a trial. After reviewing the law and the history of confidentiality, the judge refused to do so, saying in part:

> "Maintenance of confidentiality facilitates the functioning of Government by encouraging the submission of full and free census data. Secondly, it protects the privacy of members of the public who are required by law to submit information, often of a confidential nature. The court believes that these policies outweigh the defendent's need for the information, and that, accordingly requirements of due process do not compel disclosure."

That briefly is a summary of how confidentiality grew to be an integral part of census taking in the U.S. Keeping that information in mind, and my earlier remarks about generalized fear of the computer, let's look at how the 1970 U.S. census was processed.

After all the forms were collected, the data on them had to be transferred to computers. It used to be that the data on each form were manually transferred to punchcards, and the punchcards fed to computer tape. Now, we bypass this laborious process. The forms, specially designed for the purpose, were microfilmed on highspeed page-turning cameras. This is the last time each original form was handled until it is destroyed.

The first page of the census form was not microfilmed. This page contained the address of the household. So, the rolls of microfilm, which have names and personal information, contain only a geographic code relating that information to the area on which the household is located. In order to gain access to personal information, the name and address must once again be linked to the data, a process requiring the efforts of many people. This separation acts to reinforce confidentiality.

Another electronic device produced by the Bureau reads the microfilm. It transfers directly onto computer tape the dots formed when respondents filled in the circles by the appropriate answers. This machine cannot read handwriting, so the names of individuals are separated from their personal information at this point for the rest of the tabulation process.

Even this is not enough to guarantee that a person could not be identified in the statistical summaries. Some areas have such a small population that it would be possible by deduction to identify personal characteristics in the tables. Our computer processing program is set up so that if this would be the case, that information is suppressed - both on computer tape and in the printed publications.

When it comes to suppression of data from the economic censuses even the disclosure rules are confidential - because that information by itself could be used for deduction, since the number of firms involved is so many fewer than the number of citizens.

We are looking into adopting other techniques being developed by other countries for protecting confidentiality. These include rounding numbers to the nearest five, and a "random noise" system, in which values of one and negative one are scattered throughout the tabulations, balancing to zero at certain geographic levels. Such a system would have no substantial effect on statistical analysis.

I hope it is clear that confidentiality of personal records can be enhanced, not necessarily weakened, by the use of computers.

When the tabulation is finished, the original paper forms which have been stored in guarded buildings on a government facility are destroyed. They are shipped in sealed boxcars and recycled, with Bureau officials watching until they drop into the pulping vats. With this step, easy access to personal information disappears.

That leaves the microfilm copies. Where does it go after we are finished processing the data? The rolls are sent to a Bureau facility which we refer to as the Age Search Service. This is a unique, self-supporting operation which has helped millions of people. Every day the Bureau receives more than a thousand requests from people who need to verify some item of information about themselves. Most requests are for substitutes for birth certificates which either never existed, or have been lost or destroyed. People need them to qualify for retirement, for Social Security benefits, for government medical programs, to obtain a passport, and many other uses. For a small fee, this facility will search old census records and issue a certificate which is officially accepted for such purposes.

This service is provided only at the request of the person himself. For example, a son cannot ask about his father unless he has a power of attorney or a death certificate. This operation is the only use made today of the Director's authority to release personal information at his discretion.

Finding information for those who request it is not an easy job. It takes an expert to utilize the microfilm. The Age Search Service is a completely manual operation, with no computers involved. Since the census is based on addresses, there is no such thing as a computerized master list of records arranged alphabetically by name to make the search easier.

The very size of the U.S. population helps to guarantee confidentiality. It took more than 8,000 kilometers of microfilm to process the 1970 census. To locate the correct reel of film the person making the request must supply information about where he or she lived at the time of the census.

For us to make this process of working backward to arrive at personal data any easier would be an enormously complex and costly undertaking. Indeed, some suggested laws regarding privacy would have required us to organize our records in just such a way in order to comply with accountability procedures. While designed to protect personal information, in the case of Census Bureau records, it would have had the effect of making personal data more accessible.

At one time, there were more than 60 proposed laws dealing with privacy before the Congress, many of them very restrictive in nature. The version which was finally passed - the Privacy Act of 1974 - is, in principal, a very good law. It calls upon the Federal Government to follow fair information practices, and widens the area of Federal conduct which is open to public inspection and accountability.

The act applies only to personal information maintained by Federal agencies in a record system - whether computerized or not. It does not apply to private organizations or State and local governments.

It does not regulate the method of obtaining information. However, the law does establish specific requirements for advising individuals in advance of certain items: the law authorizing the collection of the information being requested; the uses to which the data may be put; whether response is voluntary or mandatory; and if there is a penalty for not responding.

The Privacy Act permits the individual to review his personal record and request amendments within a specified time. If the agency does not amend or correct a record upon request, the refusal is subject to review and ultimately, to judicial review.

However, the Privacy Act of 1974 takes into account the reputation and the needs of the Census Bureau. It does so by recognizing the difference between administrative and statistical records. Administrative records are maintained as a basis for decisions about individuals, while statistical records do not directly affect the person whose information forms part of an overall statistical picture. The initial dialogue about the need for privacy legislation did not make this vital distinction. That the final version of the law did so has allowed the U.S. statistical system to continue to function in its present form without measurable effects on the quality or costs of statistical products.

The key provision in this regard is one which allows the heads of government agencies to exempt certain records systems from individual access and correction. Such an exemption is authorized where the records are required to be maintained by law and are used solely for statistical purposes.

The Act also permits Federal agencies to disclose individually identifiable records to the Census Bureau for the purpose of planning or carrying out a census or survey under the provisions of census law. This transfer of information to the Bureau may be done without the consent of the individuals involved.

This provision recognized two points. The first is the Census Bureau's reputation for safeguarding personal information. No employee has ever been prosecuted for violating his oath regarding disclosure of personal information. Even at the height of the Watergate revelations, the name of the Census Bureau never appeared in even the most speculative stories about abuse the personal data.

The second point is a practical one. The exemption acknowledges that certain statistical programs can make use of existing administrative records both to save money and to avoid requesting individuals to provide data they have already provided for other agencies.

We at the Census Bureau believe the exemption from the most restrictive provisions of the Privacy Act for the Bureau stems from three factors:

First is the constitutional mandate to take the decennial census every ten years. This means that the Bureau has the best national population sampling frame in existence, one which, by law, cannot be borrowed. Some of the work done utilizing this sampling frame, requires administrative records from other agencies which may have to be transferred to the Bureau.

Second is the obligation placed on the Federal government by law to avoid or minimize duplicate data collection mechanisms.

Third is the fact that the Bureau has been a permanent statistical agency for more than 70 years. During that time it has pioneered in the development of sampling theory and statistical methodology, and has provided the expertise and research capability to dramatically improve the quality and accuracy of statistical work in the past few decades.

As for accountability, the Privacy Act directly affects the Bureau in several respects. As with all agencies, we must publish a notice of the existence and character of every system of records we maintain.

The Bureau may also be affected by the procedures adopted by those agencies which provide information to us for statistical use. Although individual consent is not needed to make such transfers, the supplying agency may have to keep an accounting of the transfer of certain records and make this information available to the individuals named. The agency would have to do this to be able to forward to us any correction or notation of dispute which becomes attached to the original record in accordance with the Act.

It is still unclear how far this provision might be carried, but it appears doubtful that isolated corrections to administrative records should have any significant bearing on statistical operations.

Two provisions the Act requires were already being carried out by the Bureau on its own. One stipulates that when records are transferred from one agency to another, the receiving agency must treat them as if it had originally compiled them. For the Census Bureau, this has been a one way street, with all transferred records coming under the provisions of the Census law, and no Census records transferred out of the Bureau.

As I mentioned before, the Bureau is also accountable, along with other statistical agencies, to inform respondents in writing at the time of data collection. We have sent such advance letters for some time, but the Act will cause us to make sure that the message is as clear as possible.

Perhaps the section of the Privacy Act with the most potential impact is that which creates the Privacy Protection Study Commission.

For example, the privacy commission is expected to examine the developing practice in the private sector of matching and analyzing statistical data - such as census data - with other sources of personal information. The goal of such commercial organizations is to reconstruct individual responses to statistical questionnaires for commercial or other purposes. The result can violate either implied or expressed confidentiality. While the state of the art in this area is relatively primitive, the Bureau will have to devote increasing attention and concern to the manner in which we prepare and disseminate aggregate statistics.

The Commission is also expected to determine specific categories of information which Federal agencies should be prohibited by law from collecting to avoid violating the individual's right of privacy. Even if the Commission does not explore this area, it is guaranteed to receive more attention in the next few years. I say this because such concerns always are voiced before and during the decennial censuses.

With four years to go until the next census, there are legislative proposals which would require Congressional approval for census questions, or limit the number of mandatory questions, or make all ques-

tions voluntary. Even without new laws, it appears likely that proposed census questions will receive a more thorough examination by Congressional committees than before.

Clearly, the Privacy Act and the concerns it reflects make public officials more accountable for the collection and handling of personal records. But even with the increased accountability, and the strict provisions of law, the Census Bureau continues to exercise a large measure of discretion in fulfilling its role as the Nation's prime fact finder.

One example is imputation. Since individual data records are used only in combination with many others to form statistical totals, the Bureau has traditionally refined incorrect or missing data collected in the field through direct and indirect allocation.

A direct allocation is made when a response is imputed on the basis of information given in another item for the same respondent. For example, if there was no response to the question of marital status for a woman, but she was reported as the wife of the head of the household, the computer is programmed to automatically assign the response "married" to that question.

Another form of direct allocation is the consistency edit. This is used when the information reported in a given item is inconsistent with other information reported for the same respondent. A reported age of 10 years for a grandmother of the head of the household is not allowed to stand, for instance.

An indirect allocation is made when a response is imputed on the basis of information in items reported for a different respondent. For example, missing income may be imputed from the computer record of the last person processed with the same sex, race, household relationship, and similar age and employment characteristics.

Sometimes - because of nonresponse or mechanical processing problems - it becomes necessary to impute characteristics for respondents for whom almost no information is available. In this case, the record of a nearby respondent is duplicated and substituted for the missing record. Of course, this procedure is not appropriate for certain items such as place of birth or occupation.

Since individual statistical records cannot be used for any administrative determination, these imputation procedures do not harm the person who responded to the original question. We believe that allocation and substitution provide better data than could be obtained otherwise. For one thing, these procedures reflect local variations in characteristics.

Another example of the Bureau's freedom of action involves the interpretation of its authority to collect data. Historically, this authority has been viewed broadly, rather than in a restrictive fashion.

The Constitution of the United States, written in 1787, establishes the census for the purpose of apportioning seats in the House of Representatives to assure equal political power. That language reads, in part:

> "The actual enumeration shall be made within three years after the first meeting of the Congress of the United States, and within every subsequent term of ten years, in such manner as they shall by law direct."

Congress has delegated the authority to take the census to the Secretary of Commerce. The most recent words now read:

"The Secretary shall in the year 1960 and every 10 years thereafter take a census of population, employment and housing (including utilities and equipment) as of the first day of April which shall be known as the census date."

The same law also contains these words regarding surveys:

"The Secretary may make surveys deemed necessary to furnish annual and other interim current data on the subjects covered by the censuses provided for in this title."

In turn, the Secretary has delegated these authorities to the Director of the Census Bureau, one of the agencies in the Department of Commerce.

Much as the contract of trust with the public has developed through the decades, so has the trust Congress places on those it has charged to take the census. This trust has allowed a broad interpretation of the enabling language, with the result that the census has grown into a vital statistical resource on the whole spectrum of social and economic characteristics of the American people.

The same is true regarding surveys. This has allowed us to add supplemental subjects to authorized surveys so that a wide range of social and economic trends are identified between the benchmarks of a census. These include school enrollment, educational attainment, income, marital status, birth expectations, and a number of other characteristics.

Once again, this practice saves time, money, and reduces respondent burden. It allows the most efficient use of a data collection network which is already in existence.

That is the background of how the Census Bureau became the unique organization that it is in the United States, and how it relates to the Privacy Act and concerns about the right of privacy.

However, just because our record is good, and we have been exempted from certain portions of the Privacy Act, does not mean that all problems are solved and all questions answered. One problem deals with the one way street of data transfer I mentioned before, in which the Bureau can accept records transferred from other government agencies, but can not release any of its own records. This creates a definite hardship for many agencies in performing their own required, legitimate statistical work.

There are also a number of questions we need to continually deal with if we are to maintain the unique position which has evolved. For example - are we taking every reasonable precaution with the physical protection of personal information? Are we taking an undue risk of adverse public reaction with a particular set of census or survey questions? Have we insured that in making available summary tapes we have taken all necessary and desirable precautions for preventing potential disclosures by inference? Have we sufficiently advised all employees, especially field interviewers, as to the safeguards, regulations, and policies that govern their conduct? Have we educated data users to maximize the utility of existing data as an offset to collecting new information? How many years are census records to be kept confidential? These are just some of the considerations we must keep in mind.

Earlier, I quoted from Alexander Solzhenitsyn about the threads which radiate from every person - representing the questions he has filled out on forms during his lifetime. To balance this ominous perception, and retain the analogy of Solzhenitsyn, I would like to offer a portion of a poem by Edna St. Vincent Millay:

"Upon this gifted age, in its dark hour,
Rains from the sky a meteoric shower
Of facts ... they lie unquestioned, uncombined.
Wisdom enough to leech us of our ill
Is daily spun, but there exists no loom
To weave it into fabric ...".

The most magnificent tapestries are made up of countless threads. To get the full impact of a tapestry, you don't examine a tiny portion with a magnifying glass or a microscope. Instead, you stand back at a distance and see it in its entirety. And when you do so, the individual threads are no longer discernible. To protect each thread to the point of being able to keep track of where it is in the tapestry, remove it upon demand and, change it - would simply mean that there would be no tapestry.

In the same way, society's need is for the whole cloth of statistics for rational decisionmaking. The statistician and the decisionmaker are concerned with information about individuals only insofar as it forms part of the general statistical portrait of a given area.

The Census Bureau's perception of its mission is to gather accurate, timely, relevant and complete data from individuals, businesses, and governments and to make available to the public general statistical summaries of that data while protecting the individual's privacy by keeping personal information confidential.

In this role, the Census Bureau can serve as the loom in Edna St. Vincent Millay's poem, producing the statistical cloth increasingly needed by today's decisionmakers. It is my sincere hope that this loom - and others like it around the world - will be allowed to continue their vital task.

3.9 DISCUSSION

Chairman: Ingvar Ohlsson

Ingvar Ohlsson:

We have heard two lectures on the conditions of statistics production, and it is obvious, as has been said many times here, that science and statistics production face the same problems. We all think it is natural to distinguish between administrative files on the one hand and scientific and statistical files on the other. We only want information on individuals as building blocks in a statistical process or a research process. For that reason these lectures have dealt with the problem of protection, both legal and technical, and with the need for maintaining the confidence of the suppliers of information. We have about an hour, and the first speaker to whom I give the floor is Professor Gastwirth from George Washington University. He is also chairman of the American Statistical Association's Committee on Confidentiality.

Joseph L. Gastwirth:

First let me thank all of you from Europe for the gracious hospitality I have had in Sweden and I apologize I can't speak your language. Professor Dalenius asked me to make a few comments about the work of our committee and some of the problems that I foresee in the States that may also happen here and indeed may already have happened here ahead of us. The basic problem stems from what Mr. Barabba mentioned in his paper that it is the right of the individual, to the extent possible, to control how information about himself is used, to whom and for what purposes it is released. Because our government is made up of a variety of Departments, each of which has slightly different regulations and codes of fair information practices, some statisticians felt that the American Statistical Association should study the issues and the difficulties existing and proposed laws create in the statistical community. Most of the members of the committee have had wide experience in government and although I come from an academic background, I did spend a year with Office of Management and Budget and have testified in a variety of law cases. To help get the discussion going, let me just mention some of the problems that our survey of government agencies has led us to foresee. A major problem of our act is the requirement that the public be notified of the uses of the data at the time it is collected. This may cause an unintended burden on statistical uses since it is difficult to communicate all the possible uses of, say, the Census data. By this I mean it is difficult to communicate succinctly the several hundred tables generated from the basic Census data. Thus, our committee will try to develop a reasonably accurate and compact statement of uses of survey data. Secondly, it is impossible to notify respondents of new uses of data. For example, you may wish to study the effects of the oil embargo on employment by comparing current (1976) data with the data taken in 1972. It would be extremely costly to try to find all the people interviewed four years ago. Our committee is considering recommending some type of public review board if a drastically new use of data is to be made.

There are two other related topics which you may find interesting. A fundamental problem we are facing is the merging or linking of differ-

ent data sets, usually using social security numbers (in the States). Indeed, our Privacy Protection Commission has been asked to study this issue. It is especially complicated in our country because we have mandatory data sets derived from tax data on draft and military service records in addition to survey data which is usually collected on a voluntary basis. I perceive an ethical problem when researchers combine voluntary data with administrative data when, to the best of my knowledge, the survey respondents are not told that the information may be merged with administrative data. In particular, if respondents refuse to answer a few "sensitive" questions (e.g. income) in a voluntary survey, is it honest and ethical for statisticians to obtain that information from administrative records? These problems speak to the major fear the public has in the United States, namely the fear of a massive data bank storing "everything" known about them in a master file. Here is where the statistical profession will have to help the public understanding. We will have to stress that good statistical practice does not require the linking of all Census records with all tax records with all drivers license data, etc. Statistical purposes may require the linking of survey and administrative records in small samples. Provided that the probability of any person being in a statistical file is of the order of one in 500, you can feel confident that the file is not going to be subverted since there is so little chance of finding the subject being investigated in it. Moreover, the most sensitive items of personal information, e.g., medical data, are usually not found in general purpose statistical records.

The other area statisticians should explore is how can we use micro-aggregation or merging grouped data efficiently. In some of my research on income inequality, I was able to obtain quite accurate estimates of several measures of inequality from grouped data. Currently a Ph.D. student of mine is working on estimating the correlation coefficient of a bivariate normal distribution from grouped data and the preliminary results are quite promising. Recently, Haitovsky has written a whole book on regression with grouped data. Thus, with further developments in statistical techniques, we can probably resolve the confidentiality problems inherent in merging different record systems.

Finally, we know that working with large masses of complete data sets (e.g., the entire tax records) is not only more cumbersome but usually leads to less accurate results than working with data collected from a carefully designed sample of respondents when the data is collected by trained interviewers and is processed through careful editing procedures.

Sune Åkerman:

This will be an unusual and rather unexpected confrontation for me with the decision makers who participate in drawing the lines for the research which will be possible to carry out in countries like our own or the USA by future historians and other scientists who want to devote themselves to changes in the society in a longer perspective.

It has struck me how important it is to make clear to oneself that in this regard the United States represents an entirely different tradition from that of the Scandinavian countries. It is well-known fact that you are strongly influenced by your Anglo-Saxon background when it

comes to the organization of the system of justice, the political structure in general, and in regard to many central values. Such a value is the suspiciousness vis à vis the state. But in the Scandinavian countries we hardly regard the state in general as something threatening; on the contrary, it can be thought of as the protector of the small and underprivileged. For that matter, that is the whole idea of the welfare state. This difference, I think, is important for the way in which the integrity debate is carried on. Against that background I would like to warn that we not only buy your technology but also uncritically accept your values in issues like the integrity issue. It is alarming to me that to such a large extent the discussion has been led into technical problems in connection with so-called file splitting, with codes designed to prevent access to information, and deformation and destruction of material. Instead we ought to devote much more energy to underlining how necessary personal information is to both the individual and the society and how indispensable it is for research.

And then a more direct comment to Vincent Barabba's elegant presentation. I was rather shocked by its content as a historian. What you are doing in the United States influences our research possibilities in the future, too, and I want to ask Vincent Barabba if there hasn't been a rather sharp reaction on the part of the scientists. How on earth will it be possible to conduct longitudinal, individually oriented studies a hundred years from now in the United States? It is an unusually interesting country in many ways, and evidently we will have no improvement in the future in comparison with the very difficult situation which we have now when we try to reconstruct the very fragmentary information from the censuses of 1850, 1860, 1870, and 1880. The census of 1890 has burned up, as you may know. After a great deal of hesitation the census of 1900 has been released recently. I want to present a plea from scientists on this side of the Atlantic, too: Show more consideration to us in this connection. Approach the Scandinavian mentality more if possible. Scientists constitute no threat here, and least of all those who will conduct research in the distant future about what has happened in post-industrial society in the United States. Unfortunately, their task will not be a particularly easy one, if I have understood the contents of Mr. Barabba's presentation correctly.

Vincent P. Barabba:

You may be correct that our expressed concerns in this area may be found in our Anglo-Saxon background. However, the researchers in the United States, even with their Anglo-Saxon background, hold the same position researchers here in Sweden have expressed when it comes to access of personal records. In fact, in the U.S. at this very moment there is a rather vigorous debate taking place related to access to our historical census records. As I indicated in my talk, prior to 1910 a promise of confidentiality was not made to the American people and, as such, those records have been maintained and made available by the Archivist of the U.S. The critical question that arises is what to do about the records of 1910 and following decades. The Census Bureau feels that the promise of confidentiality is essential to continued high response rates. In our promise to the American people there was no distinction made as to how long that period of confidentiality would be. The debate that is taking place right now in the U.S. focuses on how long should it be before the Archivist of the U.S. is granted the authority to release those records

for research. Part of the debate involves the extent to which the Census Bureau can still be responsive to the needs of the individual researcher, relative to the archival data, by releasing the information in anonymous individual sample records or with the consent of the respondent.

Our position is that the Constitution says there will be an enumeration of the people and the primary purpose for that enumeration is to apportion the members of the House of Representatives among the States. If releasing the individual information causes a concern in the American people which results in lower response rates, then we feel that the cost of access to the individual records is too high a price to pay even for research.

Now, that's the general position we take. We realize there are approaches we can take which do not affect our ability to enumerate the American people. For example, there was a study dealing with Scandinavian immigrants to the U.S. where the National Institutes of Health wanted to measure the incidence of a particular disease among Scandinavian immigrants in the U.S. as compared to their siblings who remained in Scandinavia. In this instance, the Census Bureau drew a sample of the immigrants as recorded in the census, searched out those individuals, interviewed them, and received their consent to contact their siblings to participate in this survey. The information was then turned over to the National Institutes of Health with the consent of the individual.

Another way that we try to meet the needs of researchers is through the use of public use sample tapes. We draw a sample of one-in-a-thousand people, or one-in-ten-thousand people depending on the need of the researcher. We eliminate any personal identifiers and further depersonalize records by not naming any geographic area of less than 250,000 people. We make that micro-record available to the research community in machine-readable tapes, and there is no way that one can identify the individuals. So, to this extent, we do a service for the research community by giving them a machine-readable tape for further analyses. There are series of projects of these types which we try to do while always reminding ourselves that our first responsibility is to enumerate the total population.

Ulf Himmelstrand:

A lot of the data collected and statistics recorded by the census bureaus in most countries of the world are collected and recorded in order to serve governments, that is as a basis for government decisions, local, regional, and federal. Some decision-makers in the business community may also find some of these statistics relevant. The time horizon of such decisions probably is rather limited in most cases. It could be a period up to the next election, or it could be a somewhat longer period. For researchers in this field the time horizon is also sometimes very short, very short indeed; but with the advent of futurological concerns, and also of course in a historical perspective, in looking backwards toward history, you have a much wider time horizon. Yesterday our discussion touched upon questions like these with reference to the access to longitudinal data.

There is also another consideration here which makes the use of census data appear in a different perspective, namely the fact that

governments naturally are trying to attain certain goals, and they thus need data which is immediately relevant for what they are planning and for the policies they want to pursue and implement, whereas at least some researchers involved in policy research may question the very basic assumptions underlying such government policies, and may wish to prove or falsify such assumptions as well as alternative assumptions providing quite different premises for the making of policy and for decision-making. Assume now that the testing of such premises requires the study of changes in individual variables over time. The question is then to what extent census data are available to test such underlying assumptions, or to prove that a certain line of policy and decision-making renders effects which are not wanted, perhaps, even by the policy-makers themselves. Such effects do turn up. I am thinking here of technology assessment not only in a limited sense but also of assessments of legislation as well as "non-decision".

Another relevant question is whether such a critical perspective and data analysis threatens personal integrity of citizens more than do the policies of the government itself. I suppose government officials and some politicians take it for granted that the personal integrity of citizens is less threatened by government decisions and laws - including the Data Act - than by researchers who want to "intrude" and find out about individual cases in order to be able to aggregate certain trends in changes of individual conditions over time. It is one of the duties of research workers to test what is taken for granted. Therefore I think that we should consider also in this discussion this basic question whether the critical approach of social scientists in analyzing and evaluating tendencies over time is more threatening to the personal integrity of citizens than some of the decisions taken by government, or more threatening even than an application of, say, a data law which implies a refusal of access to some critical data. Perhaps in some cases people's integrity may in the long run be more threatened by the absence of research findings of such a critical nature than by research based on controversial personal data.

Joseph L. Gastwirth:

I am sorry that I can't pronounce the name of the last speaker. I agree with his major point. One of the suggestions of the privacy report of the Health, Education and Welfare Department was to erase the individual identifying number as soon as possible. Then public use tapes, as noted by Mr. Barabba, could be issued, making possible various re-analyses of the basic data. This would allow researchers to study the effects of economic policy, say, without jeopardizing the personal integrity of the people interviewed. Also Professor Boruch is re-analyzing educational data. The problem facing the research community is how can we encourage independent analysis of the data used for government decision-making and preserve the anonymity of the respondents.

Vincent P. Barabba:

If I heard your question correctly, that wasn't a complete answer. Behind your question is the notion that government data collectors concern themselves only with data requirements to do an adequate job for government officials. But, that what is adequate for the government is not necessarily best for the people. And further, you feel that the

answer to the question you just raised is a much more important question of "integrity" than the maintenance of confidentiality provisions. I would grant you that we have stressed at this conference that confidentiality is a critical part of a data collecting system. There is no empirical evidence to my knowledge to support that contention. It is based solely on our own experience. We have, however, recently contracted with the National Academy of Sciences to investigate whether it's even possible to design a methodology to determine the impact on response that a promise of confidentiality has, as distinguished from that which a statement relative to no promise of confidentiality might have. That project is initially under way; and our position is that if the promise of confidentiality is not as important to the collection of statistics as we believe, then I think we will take a different look at how we make information available.

Your question is an excellent one because in any society these questions of personal integrity always seem to revolve around the point that information be generated for some good purpose. I think there are too many examples that even though adequate statistics have been provided that the well-being of the individual in a society has not been improved. Most statistical agencies spend considerable resources determining whether the data they have collected are accurate and reliable. We normally depend on the leaders in the government to determine whether we are collecting relevant or meaningful information. If the government is not correct in its function, then these very accurate and reliable data are less useful to society. As an example, in the U.S. there is an increasing interest in neighborhood governments - i.e., small groups which are, in a sense, outside the realm of the traditional municipal governments. These are people who live in an area which is described best as a neighborhood and it may not have the normal political boundaries that we are familiar with. Some participants in this movement have come to the Census Bureau and have pointed out to us that we have as much an obligation to provide information in a form which is meaningful to them as we do to provide data to the Federal, State, or local government. It is now up to us to plan for the dissemination of information to this public as well as to the Federal Government. If we can provide meaningful information to these groups, it will give them a chance to compare their "well-being" against the "well-being" of other neighborhoods as well as the larger communities of which they are a part. So I think the central point of your question is a very meaningful one and I wish we had better evidence to support our contention that confidentiality is as important as the amount of time we have allocated to it during this conference.

If we are wrong in our concerns and priorities about confidentiality, we have perhaps inhibited some valuable research from being taken. However, if we are correct, and we released the individually identifiable information, we have created a situation from which we cannot recover without serious damage to the data collection system. We would have broken a long-time promise. So that's one of the reasons we seek empirical evidence. I would point out that our position on this matter with the Archivist is to oppose the release of data for which we have made a promise of confidentiality. We intend to keep the promise. If we can demonstrate that information can be released after a certain period of time without affecting response rates, then, in future data collection activity, we will make the promise with the condition of release clearly stated.

Tore Dalenius:

I would like to elaborate on a point which I consider most important but nonetheless somewhat neglected: the public image of statistics, and especially official statistics.

I think that it is safe to state that we are presently witnessing a development toward less public appreciation of official statistics. The increased rate of refusal to cooperate in surveys is but one piece of evidence for this. It would be irresponsible to do nothing, arms folded, and hope that this development will soon come to an end and perhaps be reversed. I am convinced that the forces responsible for the present development are far too strong to be given free rein.

In the past, the National Central Bureau of Statistics (SCB) has been fortunate in facing situations which have offered it an opportunity to act in a way that may contribute to improving its public image. Such a case occurred when SCB refused to turn over schedules from the recent population census to the law-enforcement authorities to serve as a basis for searching a certain suspect. I hope this agency will get more such cases and deal with them in the same way!

But it is by no means sufficient to be passive and wait for opportunities to appear; the SCB must exercise leadership in this area. More specifically, it must initiate an educational campaign to educate the public about the benefits of official statistics. For such a campaign to be effective, it is necessary to have better information than now available about what the public knows about official statistics and about its fears and values in the realm of privacy, etc. I am glad to learn that the SCB has at last decided to devote some resources to a survey in this area.[1]

An educational campaign is clearly only one way of tackling the problem I referred to in the beginning of my contribution. I will point to another supplementary way. The SCB is - let's face it - a contributor to the "paperwork burden" in Sweden. It is interesting to note that we speak about "paperwork burden" (or "response burden"). Observe, please, the negative attitude induced by the word "burden". I suggest that it is feasible to take a positive approach, emphasizing the important role that survey respondents play by providing the factual basis necessary for sound decision-making.

Finally, I want to focus your attention on the role of "linkage of records". Many ardent champions of privacy protection take a negative view of "linkage of records"; in my view, they put far too much weight on the potential threat to the privacy of people stemming from uses of this technique. What they seem to forget is that limiting the use of "linkage of records" makes it necessary to collect the information sought in other ways, thus increasing the volume of paperwork and increasing the costs. I hope that Dr. Anér and her colleagues in the Parliament will keep this in mind in their endeavors to protect our privacy; they must not be allowed to forget that what it is all about is to strike a calculated balance between the citizen's right to privacy and the need to know.

---

[1] A report on the results of this survey is expected to appear in the fall of 1976.

Edmund Rapaport:

I would like to continue with the line of thought that Sune Åkerman followed in his contribution. He was anxious over the possibilities in one hundred years of doing research on the individual oriented data that is collected today. I believe that this is a justifiable worry. In our discussions through the last years here in Sweden we have often had a very short-sighted perspective on the use of data, with the long term perspective falling by the way-side. Future research is in a rather special disadvantageous position with respect to current demands regarding the use of data for it is today so difficult, if not impossible, to foresee what type of problems and what kind of methods researchers will be dealing with one hundred years from today. The only thing one can do is to extrapolate from the cases in which old archive-stored material was used with great success in research. However, if one were to pose oneself the question "How much did those who collected the data foresee of its eventual use when storing it?" Then the answer must surely be that they did not have the slightest idea about this.

A parallel can here be drawn with environmental protection. Today when one speaks about protection of the environment, one is usually focusing on the more short term aspects, i.e. our own environment, but also to a degree on the long term aspects, i.e. the consideration for the coming generations. The situation can be said to be the same when we are speaking about protecting the research that will be conducted many years from now. This is not to say that this weighing of the short term interests against those of the long term must always be decided in favour of the long range historical perspective. We must even here make subtle judgements and weigh the different interests against each other. The results will have to depend upon this balancing of interests. The information problem in this weighing procedure is extremely large. How are we to argue in order to convince people of the need of material which will occur in three, four, five generations? Even if I am in principal looking at the problem in the same way that Sune Åkerman did, I would like to mention one difference we have concerning the comparison of the Anglo-Saxon and the Scandinavian political traditions. Åkerman implied more or less that the values, especially as expressed in their attitudes towards society and governmental organs, are different, that they have different attitudes towards the private sector. I believe that this point of view held for perhaps ten years ago, but I strongly doubt that it is accurate today. On the contrary, I believe that one can see obvious signs of a rapid diffusion of ideas and their penetration of the national boundaries, even to the extent that we hear the echoes of each other's opinions,attitudes,and arguments across the national boundaries.

In this discussion about the role of information I would like to take up something that Kerstin Anér mentioned yesterday. Unfortunately, she is not here today. She focused on the word group-integrity and meant by that the need for the protection of interests of groups of individuals. Extending her arguments as they are we are forced to conclude that information about groups may be dangerous and therefore needs to be censored. I believe that this understanding of the situation is quite contrary to that possessed by most social scientists. When we as statisticians speak about protecting the privacy of information we mean information about individuals. This concern is however part of an attempt to make it possible to bring forth and disseminate information about groups. This is after all the aim of a statistician and it is the purpose for which statistics is studied. To legislate protection of group informa-

tion contains the risk of censoring that kind of information that is necessary for the continued development of a modern democratic society. The results can be an impoverishment of open debate, this we should strongly resist. This is not to say that in extreme cases there cannot exist information about groups that must be handled with caution. I do not want to imply that no problem can arise, rather that a basic principle must be established: in a democratic society statistical information should not be censored and withheld.

Ulf Himmelstrand:

I would like to address myself very briefly to the problem of informing the public about the meaning of statistics, and also of how to assess public understanding empirically; what people mean by confidentiality and so forth. Such studies are very urgent indeed, as a basis for informing the public more adequately about the uses of statistics and the meaning of confidentiality.

I was a bit surprised about Edmund Rapaport's concern with the fact that certain data on individuals were collected, say, fifty years ago or perhaps eighty years ago, without these induviduals understanding that Sune Åkerman, say, today would like to look at these data; would these individuals in fact agree to such a use of personal statistics fifty or eighty years later? I think such considerations are altogether beside the point if you want to inform the public correctly about research based on such statistics. It is beside the point because, as pointed out by Mr. Barabba, what a researcher does is not to zoom in on each single individual, but rather to take a look at the whole web of aggregate statistics over time. As Chairman of the Swedish Association of Sociologists I have written a letter from this association to the Data Inspection Board which you have in your conference file; here we have pointed out that the kind of registers that we are interested in are not essentially registers of persons but of variables. Individuals are incidental attributes to these variables. Information about individuals is needed only to administer the whole thing before we zoom out to get a broader view. So I think that the kind of information that Tore Dalenius was asking for is extremely important. We must make our way of using personal statistics perfectly clear. Any references to the rights of individuals, whether they lived fifty or hundred years ago or are still alive, are simply irrelevant in this particular context, that is in the context of research. That is an essential difference between the uses of data by researchers and by decision-makers. Of course, there are borderline cases when researchers work for governments and decision-makers in the business community, for instance. Such cases must be handled with special care. But the main distinction, between the ways in which government agencies and business firms use personal statistics and the ways in which researchers use such data, must still be maintained.

Örjar Öyen:

The remark I would like to make is concerned with the request for public relations on behalf of social research. Sure enough, a great deal of effort could be made to sell the projects; I am talking about the need to sell research through some kind of public relations activities. We could do much better.

But the issue of public relations is not a straightforward one. The impact of public relations may be very different from discipline to discipline. In medical research it is always possible to argue in terms of the basic motive of prolonging life. It is a universally endorsed goal. Everybody would be in favor of any project designed to gain knowledge about prolonging life, or at least this has been universally accepted until quite recently when troubles are introduced even in selling research on the basis of this kind of argument even in medical research. But in social research it is very often an entirely different matter. Almost every social science research project in some way is controversial to some group in society. I think that almost any project in some way is perceived as a threat to some groups and therefore it would be in the interest of some group or other to make an effort through whatever channels are available to prevent the research from being done. In this situation public relations easily becomes an attempt to adjust a project, or the rationale for a project, to whatever ideology is in control and this is to me one of the gravest dangers inherent in the existence of data laws and data inspection. It is probably quite true, as it has been said, that the Data Inspection Board is restricted in its activities and its judgments by the existence of the Act itself, and the activities are carried out within the margins set by the Act. We know this has to be true. But we also know that there are wide margins of judgment, and then public relations on behalf of some social science research project easily becomes an effort to estimate the ideological preferences of the members of the board in control of the data inspection, and I am sure that the Swedish social scientists have been able to make some fairly intelligent estimates of the ideological composition of the Data Inspection Board. So this is a system that will, I should say, encourage the presence of a dishonest kind of public relations, and I think we should guard against any kind of public system that creates conditions for that kind of dishonesty. Thank you!

Arne Grip:

I will only make a brief comment on Rapaport's somewhat distorted description of Kerstin Anér's thoughts on personal integrity and group integrity. I take the liberty of reading from my book ADB-system och kommunikation (Administrative Data Processing Systems and Communication), (Lund, 1974, p. 62) which deals, among other things, with this question.

> "Concentrating the discussion to personal integrity implies an unfortunate individualization. Turning social problems into private ones impedes the understanding of social relationships. Each individual is part of a society. Only a few data are specific to the individual. It follows from this that the problem of personal integrity is a social problem which concerns large groups. Which are the groups whose social integrity is in danger? Which are the groups whose integrity is threatened by "improper intrusion"?
>
> Speaking of 'improper' implies problems of legitimacy. Prisoners, the socially disabled and the mentally ill are groups where 'intrusion' becomes proper more easily. But what about people looking for work, those receiving work rehabilitation, other hospital patients, those registered with the national health insurance agency, school students, etc.?
>
> A fruitful way would be to concentrate the discussion to social integrity, i.e. that certain groups should not be more subject to arbitrariness than others, and to determine the meaning of proper in-

trusion. After the meaning of proper intrusion has been determined, the meaning of improper intrusion can simply be determined as that which is not proper."

With this I want to point out one of the difficulties in the discussions on data and integrity. People talk about data as form but forget what is most important for research: that which is contained in the data, what the data are used for and what they should not be used for.

Edmund Rapaport:

I will try to briefly discuss the question of group-protection. If by group protection we mean that individuals belong to various groups and that the groups in a social perspective have different interests which should be taken into consideration, then I am in agreement. If one looks at it in that way then it is important that information about groups becomes extensive, that it illuminates more aspects, more groups and more problems than it has until now and thereby contributes to subtle and well thought out solutions to different problems. But unfortunately the concept of group-protection as presented in a few discussions or suggestions has amounted to something that requires the suppression of information about groups. The underlying reasoning appears to be that information may be harmful to a group and therefore it should be suppressed. This I find alien, but I would be very pleased if the problem with group protection was solely a question whose nature was as the previous speaker presented it, namely, one of creating nuance-filled and well balanced information. Then I would be completely satisfied.

Ulf Himmelstrand:

Örjar Öyen mentioned the problem of possible ideological contamination in the public relations attempt. I think that it is necessary for researchers, particularly in the social sciences, in their appeals not only to the public but also perhaps to the various agencies that make decisions about our access to data files, to emphasize a kind of overarching ideology of democracy which we all profess. That kind of ideology implies a freedom of speech and right to information. We must try to get institutionalized the right to research, the right to access to information - exactly as Edmund Rapaport said. The very decision to try to conceal certain information and make it unavailable - and I am thinking here of aggregate data based on individuals - is in itself an infringement on such rights. We must appeal to these rights in the name of democracy.

Vincent P. Barabba:

I find it interesting that discussion of the two points of view tends to identify areas of agreement upon many items relative to access and the availability of information. It even occurred to me when I was listening that, perhaps, in addition to the many recommendations that were made, a more meaningful dialogue might occur between those of us who collect information for general purpose statistics and those of us who seek more detailed and furter analyses of those data. This dialogue

might lead to methodologies that could be used in such a way to allow better access and allow us to maintain what we perceive to be a contract of trust with those with whom we deal. This is something that is beginning to evolve within the U.S., as legitimate pressures for additional access to these records are coming forward. It also occurs to me that the computer and the various methods of analysis that are now available to us can perhaps offer partial solutions to this very complex problem. I can only assure you that, as it relates to the institution that I represent, this is a dialogue that we have found to be most meaningful and important - because data that are not used, are not very valuable data. Thank you very much!

Ingvar Ohlsson:

With that the discussion is ended. I just want to add a few points of view on what Tore Dalenius and Ulf Himmelstrand have said. There is the conflict, mentioned by Dalenius, between the integrity of the individual as supplier of information and the need for statistical information. Various groups in the society need information, not just governments and other decision makers, as Himmelstrand noted. It is important, as Tore Dalenius said, that the problem is to get people to understand the user side, to realize that science is important to us all, and that statistics can give statistical insights into the society. But it is difficult to reach suppliers of information and others with that information at the right time. We are working on this a great deal at the National Central Bureau of Statistics, for example in connection with the latest population census. We devoted a lot of work to informing people about it and its use. I can also mention that the Central Bureau has just appointed a relatively comprehensive investigation of the questions of loss of statistical information due to lack of response or erroneous information and information problems concerning individual statistics in order to improve the situation of suppliers of data. When suppliers of data are to decide whether or not to answer, they also ought to know what is lost if the statistics are not obtained. Ulf Himmelstrand spoke of the need for democratic checks on what is going on in the society. It is important to make clear what role science and statistics can play. I will not dwell upon this and will refrain from trying to summarize. We thank Mr. Barabba and Edmund Rapaport for their talks and also those who have participated in the discussion.

4. Theme No. 3: A REVIEW OF CURRENT METHODOLOGICAL DEVELOPMENT

4.1 RANDOMIZED RESPONSE

by

Jan Lanke

1. Background

Much research in social sciences leans heavily on information provided by individuals in some population under study; most often these individuals are selected by sampling but sometimes even the whole population is investigated. The information is often collected by means of personal interviews and we shall in what follows concentrate on that case.

In those frequent instances when the questions concern matters of a personal or even intimate nature, it is reasonable to expect that some interviewees will be reluctant to participate; this reluctance may manifest itself in refusals to answer or, even worse, in false answers. In such cases some benefit may be obtained by introducing anonymity-preserving measures; one may e.g. equip the questionnaires with detachable identification tags which are removed prior to the data processing.

On the other hand, such devices clearly do not overcome the reluctance caused by hesitation to inform a fellow human being, the interviewer, about the true state of affairs. An ingenious device for reducing such embarrassment and thus increasing the interviewees' willingness to cooperate was introduced by Stanley L. Warner in 1965.

The Warner technique, which has become known as Randomized Response (RR, for short), has as its main feature a deliberate and controlled introduction of randomness into the answers. For simplicity, consider a question that can be answered by 'yes' or 'no'. Somewhat loosely one may then say that in an RR interview, a yes-answer does not necessarily mean 'yes' - rather, it means 'yes' only with a certain probability; similarly it is only with a certain probability that a no-answer stands for 'no'.

The original Warner scheme was constructed for the simple situation in which the object of the investigation is to estimate the relative frequency of those individuals in the population that have a certain property. Since 1965 a host of variants of Warner's method have been developed, both for the simple problem of estimating a relative frequency and for more complicated problems such as that of estimating the mean, say, of a quantitative characteristic.

The object of the present summary account is to explain the main idea behind RR; the ambition is thus by no means to provide information on different variants of RR or to discuss those important problems of a practical nature that arise when an RR investigation is to be performed.

2. A simple example of a technique for RR

Suppose that we want to estimate how many drug users there are in a certain group of individuals; thus we assume that some reasonable and easily understood definition of 'drug user' is agreed upon. Since use of drugs, although not criminal - at least not in Sweden - is generally looked upon with disapproval, some embarrassment may be connected with answering the question 'Are you a drug user?" in particular if the true answer happens to be 'yes'. Then we may apply an RR technique to diminish the reluctance to take part in the interview.

Instead of giving the interviewee the straightforward question 'Are you a drug user?' we may for instance give the following more complicated set of instructions:

> Here is an ordinary well-balanced die; throw it a few times to convince yourself that it has in no way been tampered with. - Now throw it once more in such way that the result can be observed only by yourself and then say either 'yes' or 'no' according to the outcome: if the die showed 1, 2, 3 or 4, then answer the question 'Are you a drug user?'; if the die showed 5, then say 'yes'; if the die showed 6, then say 'no'.

Since a yes-answer does not necessarily mean that the interviewee is a drug user, this procedure may be considered less annoying than a usual interview.

Then the question arises how to extract the information obtained by this randomized procedure. To be specific, let us assume that in 600 interviews we have got 180 yes-answers. Out of the 600 interviewees, we expect 100 to have answered 'yes' because the die showed 5; thus only 80 of the 180 yes-answers can be supposed to be genuine. Furthermore, only 400 interviewees can be expected to have got the outcomes 1, 2, 3, 4 when rolling the die and thus only 400 are likely to have answered the question on drugs. In conclusion: out of (probably) 400 interviewees that have given information on their use of drugs, (probably) 80 answered 'yes'; the estimate of the frequency of drug users thus is 80/400=0.20, i.e. 20 %.

The randomized procedure just described has clearly made it possible for the investigator to obtain valid information on the incidence of drug consumption although the interviewer cannot identify with certainty any single interviewee as a drug user. However, this protection of the interviewees' privacy must be paid for in a certain sense; a simple way of describing how this happens is to say that the estimator is less precise than it would have been in an ordinary interview investigation, had all the interviewees been willing to answer the straightforward drug question truthfully. On the other hand, owing to the sensitiveness of the matter we believe that an ordinary interview would result in untruthful reporting, causing an unknown systematic error in the estimate and leading, most probably, to underestimation of the number of drug users. Thus a randomized procedure can be said to change an unknown and inestimable amount of bias into ordinary statistical variation, the effects of which can easily be assessed by standard statistical procedures.

3. Some statistical comments[1]

In the type of randomized interview described in the preceding section it was quite clear on intuitive grounds how one should find the estimate π_A^* of the unknown relative frequency π_A. The result $\pi_A^* = 0.20$ in our numerical example can of course also be obtained by means of the following straightforward statistical argument.

Let λ denote the probability of receiving a yes-answer in an interview of the type considered. Clearly the conditional probability of a yes-answer is π_A if the die shows 1, 2, 3 or 4 while it is 1 if the die shows 5 and 0 if it shows 6. The theorem of total probability then gives

$$\lambda = \frac{4}{6}\cdot\pi_A + \frac{1}{6}\cdot 1 + \frac{1}{6}\cdot 0,$$

which means that

$$\pi_A = \frac{3}{2}(\lambda - \frac{1}{6}).$$

If x interviewees out of n give a yes-answer, x/n is the only reasonable estimate of λ:

$$\lambda^* = \frac{x}{n}.$$

Thus we have

$$\pi_A^* = \frac{3}{2}(\frac{x}{n} - \frac{1}{6}),$$

which for x = 180, n = 600 gives $\pi_A^* = 0.20$.

So far this more formal approach has told us nothing that intuition did not tell us in Section 2. But when we want to discuss in quantitative terms the precision of the estimator, the present formalism is indispensable. From standard statistical theory it follows that the standard error $d(\lambda^*)$ of the estimator λ^* is given by

$$d(\lambda^*) = \sqrt{\lambda^*(1-\lambda^*)/n}.$$

Hence the standard error of our estimator $\pi_A^* = \frac{3}{2}(\lambda^* - \frac{1}{6})$ is

$$d(\pi_A^*) = \frac{3}{2}\sqrt{\lambda^*(1-\lambda^*)/n}$$

and in the numerical example where $\lambda^* = 180/600 = 0.30$, n = 600, we have

$$d(\pi_A^*) = \frac{3}{2}\sqrt{0.30\cdot 0.70/600} = 0.028.$$

Thus a 95 % confidence interval for π_A is given by

$$0.20 \pm 1.96 \cdot 0.028,$$

i.e.

$$(0.14,\ 0.26).$$

---

[1] This section is written for those who have some prior knowledge of statistical theory.

Just as a comparison, suppose a straightforward interview investigation had been possible and that a sample of 600 persons had given 120 yes-answers resulting in the same numerical value 0.20 of the estimate $\pi_A^* = 120/600$ as we had in the randomized interview. Then the standard error would have been

$$d(\pi_A^*) = \sqrt{\pi_A^*(1-\pi_A^*)/n} = \sqrt{0.20 \cdot 0.80/600} = 0.016$$

giving

$$0.20 \pm 1.96 \cdot 0.016,$$

i.e.

$$(0.17, 0.23),$$

as a 95 % confidence interval for π_A.

Since the RR interval is larger than the interval from the ordinary non-randomized investigation, it is clear that there is no point in using the RR technique in situations when the interviewees are likely to tell the truth in a non-randomized interview. On the other hand, in those cases when an ordinary interview would give an unknown bias, it is clearly of value to obtain an interval which, although larger, does have the confidence coefficient that it claims to have.

4. Some problems that can be treated by RR

So far we have concentrated on the simple problem of estimating one single relative frequency. A natural generalization of this is the problem of estimating simultaneously two or more relative frequencies, all relating to sensitive characteristics; one may e.g. wish to estimate in one single set of interviews both the frequency of people using hard drugs and the frequency of people using drugs but not hard drugs. This type of problem can also be handled by RR methods.

A somewhat more intricate problem is to estimate the correlation between two sensitive qualitative characteristics; as an example may be mentioned the hypothetical research topic "Is illegal abortion commoner among drug users than in the population as a whole?" The RR technique can be adapted to take care of that situation as well.

Let us also mention that RR is by no means restricted to investigations of qualitative properties; some quantitative variables that may be investigated by RR are:

the number of times that the interviewee has used drugs during the last year,

the number of illegal abortions that the interviewee has undergone,

the amount of income that the interviewee suppressed on his last income tax return.

Finally, it should be pointed out that in those cases when the RR technique is utilized, one does usually not randomize the whole interview. In general only a few questions of a very private nature are included in a questionnaire; the remaining part of the interview is performed in the ordinary way.

4.2 APPLICATIONS OF THE RANDOMIZED RESPONSE TECHNIQUE

by

Sven Eriksson

Aims

The randomized response (RR) interview method[1] has often met with scepticism from researchers who have not heard of it before. One reason is, of course, that one cannot expect the respondents to understand how the answers they give can be used to obtain information on the distribution of a sensitive variable in the population.

Another reason for the scepticism may be that it is overlooked that the method is designed only for very restricted use. The application of RR interviews is limited to the following situation.

a. the usual measurement methods work very unsatisfactorily,

b. very experienced and skilful interviewers are available,

c. the RR interviews are used only for one or a few sensitive questions at the end of a relatively long interview.

The RR method serves two purposes:

1. Protection of privacy

The kind of protection is different from that offered by other methods such as anonymous answers and confidential handling of data. Nobody, not even the interviewer, is aware of the respondent's true value of the sensitive variable. In other words, there exists no sensitive information that may be misused. Individual records are meaningless, while data on the group or aggregate level are informative.

It is important to note that this kind of protection of integrity is not achieved at the price of lost identifiability. Therefore the possibility of longitudinal studies remains even if RR interviews are used.

2. Reduction of systematic errors

The original aim of the method was to diminish the rate of partial refusal and the rate of false answers.

As RR interviews induce an additional random error in the responses, the accuracy of the estimates can be improved only if the reduction of systematic errors is larger than the added random error.

As random errors of estimates decrease with increasing sample size while systematic errors are independent of sample size, reduction of the total error (if possible) is obtained only if the sample exceeds a certain size.

It may be questioned whether protection of integrity and improved accuracy are conflicting interests. The answer is probably no for with unsatisfactory protection of privacy (too revealing RR measurement procedure) the systematic errors will also be large. In order to avoid systematic errors as far as possible one will therefore have to provide

---

[1] An introduction to this method is given by Lanke in another paper in this volume (pp. 115-118).

adequate protection of privacy. Certain requirements on the degree of protection may be given in advance in terms of probability. It may be prescribed for instance, that the probability that a person belongs to a certain sensitive group given a certain answer may not for any answer exceed a given level.

Types of analysis possible

The analysis of RR data is not limited to estimation of means and proportions. Cross-classification with one or both variables observed with the RR technique (Eriksson (5)) and regression analysis (see a forthcoming paper by Eriksson) are also possible.

As mentioned above, time-series studies of different population subgroups may also be conducted. All these types of analysis, however, require relatively large samples.

Empirical studies

The development of the statistical theory of randomized response techniques has been very rapid since it was first proposed by Warner in [18]. During the last few years there has also been a relatively rapid growth of the number of empirical studies published in scientific journals and technical papers, and probably several more will appear in the near future.

A list with examples of applications, nearly all from the USA, is given below. The numbers refer to the list of references at the end of this paper.

Induced abortion	(1), (11), (13), (15)
Illegitimate births	(2)
Contraceptive use	(11), (15)
Sexual behavior	(15)
Emotional problems	(11)
Use of narcotic drugs	(3), (10), (14), (17)
Arrests	(7)
Drunken driving	(8), (9), (16)
Contact with organized crime	(14)
Illegal gambling	(14)
Involvement in bankruptcy	(16)
Voting behavior	(16)
Income	(12), (17)

It has been surprisingly easy in all the surveys to explain to the respondents how to interpret the instructions using the randomizing instrument (e.g. the deck of cards, the coin or the die). The method has worked even among illiterate respondents.

The partial refusal to answer sensitive questions has been very low when the randomized response technique has been used.

A problem is whether the respondent trusts the method or suspects that a trick is involved in the measurement process. The extent to which the interview instructions have been followed is difficult to measure. Comparisons of the size of the estimates from randomized response estimates and estimates from control groups given open interviews indicate, however, that the systematic errors of randomized response estimates often are the smaller ones if the investigations

are conveniently designed.

Surveys without control groups, especially two American abortion studies with approximately 3,000 respondents, have also produced estimates considerably higher than estimates from earlier conventional surveys.

In some studies a projective technique has been used to elucidate the views of the respondents. Questions such as "Do you think that your friends would honestly answer a direct question concerning induced abortions?" and "Do you think that your friends would honestly answer a randomized response question concerning induced abortion?" have been used. The answers to such questions, which of course must be interpreted with caution, and also the size of the estimates of the parameters of the sensitive variables, indicate that the majority of the respondents do not distrust the randomized response method.

Matters that are highly sensitive in one country are not necessarily so in another country. Nor does the fact that randomized response interviews have been shown to work well for some sensitive variables imply that they will also do so for others. Therefore there is a great need for further evaluation studies of randomized response interviews concerning various sensitive matters.

A Swedish field study

A small-scale test of the randomized response technique was made in a Stockholm suburb in 1973 by the author and the National Central Bureau of Statistics (SCB). The main purpose was to discover whether or not the measurement method was accepted by the respondents.

Two non-random groups of persons living in a certain area participated in the survey. Both groups were given the same questions concerning their household members, their flat, the service in the area and the economy of the family. The only difference was that the respondents in one of the groups answered two questions on public relief (socialhjälp) using the randomized response technique while the respondents in the other group gave direct answers. The number of completed interviews is given in table 1.

Table 1. Non-response and number of interviews completed

	Group given randomized response interviews	Group given open interviews
Total number planned	100	100
Empty dwellings	3	6
Occupied dwellings	97	94
Interviews	77[1]	74[1]
Refusal[2]	15	14
No contact[3]	5	6

[1] One interview was conducted by mistake as a randomized response interview instead of an open interview. The figures in the tables below will therefore add to 78 and 73 respectively.

[2] Total refusal, no question answered (independent of interview method).

[3] After several trials.

A number of persons participating in the survey were selected from the register on public relief (socialregistret) with permission of the authorities (Swedish Board of Health and Welfare). Therefore it was possible to analyze separately the answers of persons who had received public relief during 1972 and those who had not. (The interviewers were not given this information.) This opportunity was not available in other evaluation studies as no records have been existent. The possibility of separating persons possessing the sensitive attribute from those who do not is very important as the two categories perhaps do not react in the same way to the randomized response technique.

Table 2 shows the composition of the sample of respondents given open interviews and the answers to question 26 which reads as follows.

Question 26: Have you or your family during 1972 received public relief or borrowed money from "socialbyrån"?

Even if the sample size is very small it is obvious that public relief is a sensitive variable which causes large systematic measurement errors. 27 % of the respondents (6 out of 22) deny the receipt of public relief. This means that the ratio between the "estimate"

Table 2. Question 26. Number of persons admitting receipt of public relief in the sample given open interviews

Answer / True state	Have got public relief	Have not got public relief	Total number
Have got public relief	16	6	22
Have not got public relief	-	51	51
Total number	16	57	73

$\hat{p}$ = 16/73 and the true value p = 22/73 of the proportion of the 73 persons receiving public relief is 0.73.

At the end of the interviews the respondents were instructed on the randomized response technique and given the randomizing instrument, a deck of cards with the following composition:

Give a true answer	35 cards
Say: "No"	5 "
Say: "Yes"	15 "

The interview technique was first demonstrated using a question on month of birth. Then it was used to obtain answers to question 26 (see table 3). The expected number of answers is calculated under the assumption of completely truthful reporting. The deviations between outcomes and expectations are accounted for by the combined effect of sampling fluctuations in the selection of cards, untruthful reporting and memory errors. The ratio between the estimate (obtained as described by Lanke in another paper in this volume) of the proportion of persons receiving public relief and the true figure is now 0.75 (slightly larger than for open interviews). But the sample sizes are too small to allow any comparisons with the results of open interviews with

regard to the accuracy of the methods. Either of the methods may be the more accurate one.

Table 3. Question 26. Randomized response interviews. Number of persons asserting receipt of public relief
Expectations (E) and standard deviations (σ) conditional upon the number of answers (34 and 42 respectively)

True state \ Answer	Have got public relief	Have not got public relief	Re-fusal	Total number
Have got public relief	22 (E=30.9) (σ= 2.81)	12 (E=3.1) (σ=2.81)	(2)	34(+2)
Have not got public relief	16 (E=11.5 (σ=2.70)	26 (E=30.5) (σ=2,70)	0	42
Total number	38	38	(2)	76(+2)

The respondents were also asked a question on the amount of public relief received.

The deck of cards for the randomized response procedure now had the following composition:

GIVE THE TRUE ANSWER: that is the true one of the following alternatives: Have not received public relief or Less than 1 500 kr or Between 1 500 and 3 500 kr or More than 3 500 kr	30 cards
GIVE THIS ANSWER: "I have not received public relief"	5 cards
GIVE THIS ANSWER: "Less than 1 500 kr"	5 cards
GIVE THIS ANSWER: "Between 1 500 and 3 500 kr	5 cards
GIVE THIS ANSWER: "More than 3 500 kr"	5 cards

The interviewers were positive about the method. The technique had been easy to explain, even in the quantitative case, and the extra time consumed in explanation was on the average 2 minutes.

The respondents generally accepted the method without objection; only 2 refusals were obtained. When asked about their opinions the majority of the respondents were positive, some characterized the technique as ridiculous or silly and the two persons who refused to answer the questions 26 and 27 were, of course, wholly negative. More detailed reports of the reactions are found in Eriksson (4) and in an unpublished paper from SCB, Utredningsinstitutet (Dertell, H. [1973]. Försöket med randomiserad response).

Data banks

The randomized response technique may be used, as described above and also by J. Lanke, in the interview phase in order to preserve the privacy of the respondent.

Alternatively it may be used with the same aim in later stages of data handling

a. data can be transformed using the same technique before they are stored in registers. In this case it would be possible to transform data without distortion of the population means and other parameters. The advantage of this over disidentification is e.g. that time series studies for different subgroups would still be possible,

b. data stored subject to disclosure restrictions can be transformed using some randomized response technique before they are made available for scientific and other purposes.

References

(1) Abernathy, J.R. & Greenberg, B.G & Horvitz, D.G., 1970, Estimates of Induced Abortion in Urban North Carolina. Demography 7, No. 1, pp. 19-29.

(2) Abul-Ela, A.L.A. & Greenberg, B.G. & Horvitz, D.G., 1967, A Multi-proportions Randomized Response Model. Journal of the American Statistical Association, 62 (September, 1967), pp. 990-1008.

(3) Brown, G.H. & Harding, F.D., 1973, A Comparison of Methods of Studying Illicit Drug Usage. Technical Report 73-9. Human Resources Organization.

(4) Eriksson, S., 1972, Randomiserad response, en intervjuteknik för känsliga frågor. Statistisk Tidskrift 1972, No. 3, pp. 223-234.

(5) Eriksson, S., 1973, A New Model for Randomized Response. International Statistical Review, 41, No. 1, pp. 101-113.

(6) Eriksson, S., 1973, Randomized Interviews for Sensitive Questions. Unpublished Thesis. Department of Statistics, University of Göteborg, Sweden.

(7) Folsom, R. 1974, A Randomized Response Validation Study: Comparison of Direct and Randomized Reporting of DUI Arrests, Final Report, pp. 255-807. Prepared for Research Center for Measurement Methods, U.S. Bureau of the Census, Research Triangle Institute.

(8) Folsom, R. & Greenberg, B.G. & Horvitz, D.G. & Abernathy, J.R., 1973, The Alternate Questions Randomized Response Model for Human Surveys. Journal of the American Statistical Association, 68, pp. 525-530.

(9) Gerstel, E.K. & Moore, P. & Folsom, R.E. & King, D.A., 1970, Mecklenburg County Drinking-Driving Attitude Survey 1970. Report prepared for U.S. Department of Transportation. Research Triangle Institute, Research Triangle Park, N.C.

(10) Goodstadt, M.S. & Gruson, V., 1975, The Randomized Response Technique: A Test on Drug Use. Journal of the American Statistical Association, 70, pp. 814-818.

(11) Greenberg, B.G & Abernathy, J.R. & Horvitz, D.G., 1970, A New Survey Technique and Its Application in the Field of Public Health. Milbank Memorial Fund Quarterly, Part 2, pp. 39-55.

(12) Greenberg, B.G. & Kuebler, Jr., R.R. & Abernathy, J.R. & Horvitz, D.G., 1971, Application of the Randomized Response Technique in Obtaining Quantitative Data. Journal of the American tistical Association, 66, June, 1971.

(13) I-Cheng, Chi, Chow, L.P. & Rider, Rowland V., 1972, The Randomized Response Technique as Used in the Taiwan Outcome of Pregnancy Study. Studies in Family Planning 3, pp. 265-269.

(14) Illinois Research Institute and the Chicago Crime Commission, 1971, A Study of Organized Crime in Illinois. Report Prepared for the Illinois Law Enforcement Commission, Chicago, Illinois.

(15) Krotki, Karol, J. & Fox, B., 1974, Randomized Response Technique. The Interview and the Self-administrated Questionnaire. An Empirical Comparison of Fertility Reports. Proceedings of the Social Statistics Section, American Statistical Association, pp. 367-371.

(16) Locander, W. & Sudman, S. & Bradburn, N., 1974, An Investigation of Interview Method, Threat and Response Distortion. Proceedings of the Social Statistics Section, American Statistical Association, pp. 21-27.

(17) Smith, L. & Federer, W.T. & Raghavarao, D., 1974, A Comparison of Three Techniques for Eliciting Truthful Answers to Sensitive Questions. Proceeding of the Social Statistics Section, American Statistical Association, pp. 447-452.

(18) Warner, S.L., 1965, Randomized Response: A Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association, 60, pp. 63-69.

4.3 COMBINED QUESTIONS: AN ALTERNATIVE DATA-GATHERING DEVICE TO RANDOMIZED RESPONSE FOR SENSITIVE QUESTIONS

by

Bengt Swensson

1. Background

Two measurement methods widely used in surveying human populations are data collection by means of personal interview and data collection by means of mail inquiry. Both methods present considerable difficulties when the questions are of a sensitive or embarrassing nature.

When faced with sensitive questions some respondents will refuse to answer, while some respondents will deliberately give false answers - both cases giving rise to non-sampling errors. This fact has for a long time been, and still is to a considerable degree, a great obstacle to collecting data on sensitive matters in a statistically satisfactory way.

During the last few years researchers, government offices and survey institutes have witnessed a rapidly increasing unwillingness among the public to cooperate in surveys, even those not dealing with sensitive matters. Thus, as is also evidenced by the present symposium, there is an urgent need for methods which make possible the gathering, and protection, of survey data. Such methods have been under development by survey statisticians since 1965.

2. Warner's breakthrough

In Warner [4] a new technique for coping with the problem of evasive answers is suggested. The technique is meant for data collection based on personal interviews with purpose of estimating the proportion of units belonging to a specified group (A) characterized by a stigmatizing attribute, and it is designed on the assumption that cooperation of the respondents will increase with increasing anonymity, thus eliminating or at least largely reducing evasive answers.

The method of assuring the respondent's privacy consists in allowing every respondent to select one of two complementary questions (concerning group A affiliation) using a random device. The respondent is then asked to answer the selected question truthfully - without revealing which of the two questions he is answering. Knowing the probability of selecting either question, the statistician will be able to estimate the unknown proportion unbiasedly.

The Warner technique of eliminating evasive answers to sensitive questions has been extended in several papers. In this paper no account of these extensions will be given, since they will be found in the material gathered by Prof. Dalenius and in the lectures by Dr. Eriksson and Dr. Lanke.

3. The purpose of this paper

The purpose of this paper is to draw attention to an alternative to

randomized response proposed by the present author (Svensson [3]). The technique assures the respondent's anonymity without the use of random devices. It is applicable to data collection by means of personal interview as well as by mail inquiry. This is in contrast to randomized response techniques, which usually demand personal interviews (although some work has been done to adapt the technique to mail inquiries as well).

The technique consists of using combined questions on the sensitive attribute (A) and two insensitive complementary attributes (Z and $\bar{Z}$ = not-Z).

4. The idea

We will study two dichotomies, A, $\bar{A}$ (= not-A) and Z, $\bar{Z}$ (not-Z) and combinations of the two in a well-defined population of N units (persons). It is embarrassing or condemning to have attribute A, while attributes $\bar{A}$, Z and $\bar{Z}$ are neutral.

The purpose of the survey is to estimate P_A - the proportion of the population belonging to group A (consisting of persons with attribute A).

Since attribute A is stigmatizing, direct questions concerning this attribute will lead to evasive answer bias. To avoid this we now propose the use of combined questions.

There are several ways of combining questions, different combinations giving more or less anonymity for the respondents. Before giving two examples of how to combine questions we will consider the following four-fold table:

	Z	$\bar{Z}$	
A	P_{AZ}	$P_{A\bar{Z}}$	P_A
$\bar{A}$	$P_{\bar{A}Z}$	$P_{\bar{A}\bar{Z}}$	Q_A
	P_Z	Q_Z	

where

$Q_A = P_{\bar{A}} = 1 - P_A$

$Q_Z = P_{\bar{Z}} = 1 - P_Z$

The table gives the proportion of persons in the population belonging to various sub-groups; e.g., P_{AZ} is the proportion of persons in the population having both attribute A and attribute Z (the proportion belonging to $A \cap Z$). (For example, let the members of group A be characterized by having used narcotics at least once, and let the members of group Z be characterized by preferring pink roses to yellow roses. Then P_A is the proportion of the population having used narcotics at least once, P_Z is the proportion of the population preferring pink roses to yellow roses, and P_{AZ} is the proportion of the population having used narcotics at least once and preferring pink roses to yellow ones.)

---

[1] A more comprehensive account of the technique will be found in a forthcoming Ph.D. thesis.

Now consider the following two combined questions:

(i) "Do you belong to group AUZ?"
(ii) "Do you belong to group $A\bar{U}Z$?"

where members of group AUZ are characterized by belonging to group A or group Z (or both), and members of group $A\bar{U}Z$ are characterized by belonging to group A or group $\bar{Z}$ (or both).

Further, suppose we don't know whether a person is a member of group Z or not.

Then it is obvious that if a person truthfully answers "yes" to question (i) we don't know if he has attribute A, or if he has attribute Z but not A. That is, we don't know whether he belongs to group A or not.

Finally, if another person truthfully answers "yes" to question (ii) we don't know whether he belongs to group A or not.

In the next section we will show how the two questions can be used to estimate P_A unbiasedly while at the same time preserving anonymity for the respondents.

To do this we will first introduce the following notation:

$P_1 = P_A + P_{\bar{A}Z}$ = the true proportion in the population having attribute A or attribute Z (or both)

$P_2 = P_A + P_{\bar{A}\bar{Z}}$ = the true proportion of the population having attribute A or attribute $\bar{Z}$.

Using the fact that $P_{\bar{A}Z} + P_{\bar{A}\bar{Z}} = Q_A$ it is easily verified that

$$P_A = P_1 + P_2 - 1. \tag{1}$$

5. An unbiased maximum likelihood estimator of P_A

To estimate P_A we draw two independent simple random samples with replacement of size n_1 and size n_2 from the population. We will then ask different questions of the persons in the two samples.

The respondents in sample 1 are asked to reply "yes" or "no" to the combined question (i): "Do you belong to group AUZ?"

Let v_1 be the number of "yes" answers.

The respondents in sample 2 are asked to reply "yes" or "no" to the combined question (ii): "Do you belong to group $A\bar{U}Z$?"

Let v_2 be the number of "yes" answers.

Under the assumption that the respondents reply truthfully, it is now possible to prove the following theorem.

THEOREM 1. An unbiased maximum likelihood estimator of P_A is given by

$$\hat{P}_A = \hat{P}_1 + \hat{P}_2 - 1, \tag{2}$$

where

$$\hat{P}_1 = v_1/n_1 \quad \text{and} \quad \hat{P}_2 = v_2/n_2.$$

The variance of $\hat{P}_A$ is given by

$$V(\hat{P}_A) = \frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}, \tag{3}$$

which is estimated unbiasedly by

$$v(\hat{P}_A) = \frac{\hat{P}_1\hat{Q}_1}{n_1-1} + \frac{\hat{P}_2\hat{Q}_2}{n_2-1} \tag{4}$$

where

$$\hat{Q}_1 = 1 - \hat{P}_1 \quad \text{and} \quad \hat{Q}_2 = 1 - \hat{P}_2 \tag{5}$$

Under an optimum choice of n_1 and n_2, for fixed total sample size n $(= n_1+n_2)$, the variance of $\hat{P}_A$ reduces to

$$V_0 = \frac{1}{n}\left(\sqrt{P_1Q_1} + \sqrt{P_2Q_2}\right)^2. \tag{6}$$

The above theorem gives the information necessary for measuring and controlling the degree of uncertainty associated with approximations of P_A based on combined questions and simple random sampling, assuming truthful answers from the respondents. (Truthful answers will unless otherwise explicitly stated, be assumed throughout the paper.)

One pertinent question is now: How does the technique of combined questions perform relative to the techniques of randomized response? The answer to the question must be based on theoretical as well as on empirical considerations. In this paper we will deal predominantly with the theoretical side.

To throw some light on the relative merits of combined questions (CQ) in comparison with randomized response (RR) we will start by discussing the degree of protection given by CQ.

6. On the degree of protection[1]

To measure the protection given the respondents we will follow closely the suggestions for RR-plans by Lanke [2] and Andersson [1].

Let $\Pr(A|\text{yes})$ = the conditional probability that the respondent belongs to group A, given the respondent has given the answer "yes"

and let $\Pr(A|\text{no})$ be analogously defined. These two conditional probabilities will be called risks of suspicion.

Then $\Pr(A|\text{no}) = 0$ for "no"-answer respondents in sample 1 as well as in sample 2.

---

[1] Here and in the rest of this paper we will for simplicity assume that A and Z are statistically independent, that is $P_{AZ} = P_A P_Z$.

For "yes"-answer respondents in sample 1 Pr(A|yes) is given by

$$Pr_1(A|yes) = \frac{P_A}{P_Z + P_A Q_Z} \tag{7}$$

while for "yes"-answer respondents in sample 2 Pr(A|yes) is given by

$$Pr_2(A|yes) = \frac{P_A}{Q_Z + P_A P_Z} \tag{8}$$

Finally, define

$$M = \max[Pr_1(A|yes), Pr_2(A|yes)] \tag{9}$$

as the measure of protection.

For this measure of protection we have: the smaller the value of M, the larger the degree of protection (for respondents with the largest risk of suspicion).

7. Combined questions versus Warner's original RR-plan

In the original RR-plan given by Warner [4] the measurements on n units (persons) of a simple random sample with replacement are obtained as follows.

Every respondent selects one of the two statements

(i) "I belong to group A"
(ii) "I belong to group $\bar{A}$"

in that way the probability of selecting statement (i) is p_W and the probability of selecting statement (ii) is $1-p_W$. Without reporting which statement he has selected the respondent says yes if he has selected the correct statement about his group membership, no otherwise.

With x units reporting a "yes" answer Warner gives the unbiased maximum likelihood estimator

$$\hat{P}_A = \frac{p_W - 1}{2p_W - 1} + \frac{x}{(2p_W - 1)n} \tag{10}$$

and its variance

$$V_W = \frac{P_A Q_A}{n} + \frac{1}{n} \frac{p_W(1-p_W)}{(2p_W-1)^2} \tag{11}$$

For the measure of protection we have

$$M_W = \frac{P_A p_W}{P_A p_W + Q_A q_W} \qquad \text{where } q_W = 1-p_W \tag{12}$$

if

$$p_W \geq \frac{1}{2} \tag{13}$$

Due to the symmetry of the Warner model we can, without loss of generality, assume that (13) is fulfilled.

Since both CQ and RR are built on the assumption that cooperation on part of the respondents will increase with increasing anonymity, it seems fair to compare the two techniques under equal degree of protection. The following theorem gives the basis for such a comparison.

THEOREM 2. For every value of P_Z there is a unique value of p_W, namely

$$p_W = \begin{cases} 1/(1+P_Z) & \text{if } P_Z \leq 1/2 \\ 1/(1+Q_Z) & \text{if } P_Z > 1/2 \end{cases} \tag{14}$$

such that CQ and RR are equally protective.

Now let the cost of survey operations leading to the RR-estimator be k times the cost of survey operations leading to the CQ-estimator. Then it is possible to prove the following theorem.

THEOREM 3. For given P_Z choose p_W according to theorem 2. Then the cost efficiency E of the RR-estimator compared to that of the CQ estimator (using optimum choice of n_1 and n_2), for $P_Z \leq 1/2$, is given by

$$E = \frac{1}{k} \frac{Q_A Q_Z}{1-P_A Q_Z} \left\{ Q_Z + \sqrt{\frac{1-Q_A P_Z}{1-Q_A Q_Z} P_Z Q_Z} \right\}^2 \tag{15}$$

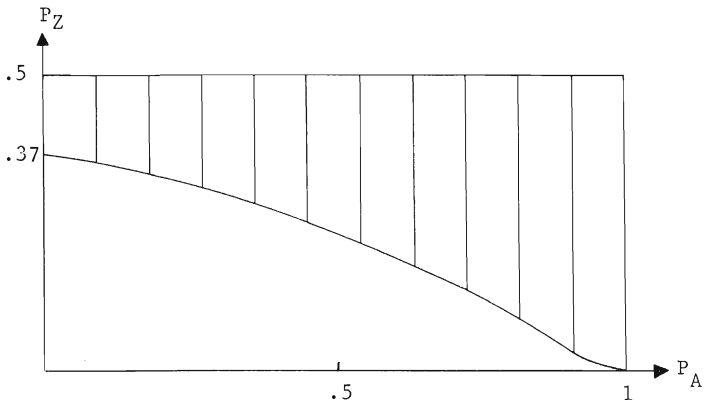
The cost efficiency for the case $P_Z > 1/2$ can be computed from formula (15) by exchanging P_Z for Q_Z.

Assuming truthful answers for RR as well as for CQ and approximate knowledge of k, P_A and P_Z we will, with the help of (15), have guidance for the choice between RR and CQ.

For the case k = 1 (that is, the cost of survey operations leading to the RR estimator is the same as the cost of survey operations leading to the CQ estimator), we give in figure 1 the approximate boundaries in the (P_A, P_Z)-plane where CQ is more efficient than RR.

From figure 1 we see that there are many situations in which CQ is more efficient than RR, even under the assumption of equal costs. For example, we see that CQ is always superior to RR when P_Z must be chosen in the interval (.37, .5), which corresponds to a choice of p_W in the interval (.67, .73). When RR must be based on personal interviews and CQ can be based on mail inquiry, CQ will be superior in many more situations.

Figure 1



Note: Hatched area indicates the region where $V_0 < V_W$, for $P_Z \leq 1/2$.

8. Combined questions versus Simmons' unrelated question RR

From the articles gathered by Prof. Dalenius and from the lectures by Dr. Eriksson and Dr. Lanke it is seen that RR-plans more efficient than the original Warner plan have been developed, some of them based on an idea by Simmons. In a near future a report by the present author in Prof. Dalenius' research project Confidentiality in Surveys will show that there are situations where CQ is superior to many of these plans.

References

[1] Anderson, H., 1975, Efficiency versus protection in randomized response designs. Ph.D. thesis, Lund.

[2] Lanke, J., 1975, On the degree of protection in randomized interviews. The research project Confidentiality in Surveys, report No. 2, Department of Statistics, University of Stockholm.

[3] Swensson, B., 1974, Combined questions: a new survey technique for eliminating evasive answer bias. The research project Errors in Surveys, report No. 70, Department of Statistics, University of Stockholm.

[4] Warner, S.L., 1965, Randomized response: a survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, pp. 63-69.

4.4 PROTECTION OF INFORMATION STORED IN A COMPUTER SYSTEM

by

Ingemar Ingemarsson

1. Introduction

The need for information protection is certainly as old as the need for communication. As soon as we have information to convey to a specified destination there is always the risk that the information may be destroyed, changed or lost in some unpredictable way. In earlier days, when information was processed by human beings, we either trusted the persons involved or hid the information by using sealed envelopes or ciphers.

The security problem was not widely recognized until the advent of electronic data processing (EDP). There are several reasons for this. In my opinion, the most important one is the lack of widespread knowledge about EDP systems. Automated information processing seems more mysterious than manual, and thus we react with suspicion and fear. There are other, more technical reasons for the accentuation of the security problem in EDP systems. They handle more data and have higher processing speed than manual systems. An EDP system is complicated and may be vulnerable to misuse and technical errors. The fact is, however, that an EDP system is not more complicated than many other systems that are in use. The problem is that we have as yet no "theory of computing" which would enable us to make more precise statements about the performance than we are able to make today.

2. Overview of models and methods

Some work has been done towards usable models for data processing. These may be roughly divided into two types of approach. The first, which is used mostly by computer scientists, regards the computer as a logical machine responding to input programs and data. The response is dependent on the previous input to the machine. The goal of the designer is that the system shall respond only as predicted.

The other type of approach, used mainly by communication scientists, is to regard the computer as a part of a communication system. This stresses the difference between data and information. Information is transmitted from a source, stored and processed in the EDP system and conveyed to a destination. The important thing is to protect the information thus circulating in the communication system. Data, in the sense of strings of binary symbols, may be changed or lost without change or loss of the information in the data.

Which type of model we use depends upon the problem to be analyzed. Our goal is to design the system so that information is protected against unpredicted change, loss or destruction. Several features are included in the EDP system to facilitate information protection. The basic protection mechanism is based on identification (of users, machines and programs) and authorization. The methods may be divided into ticket-oriented and list-oriented systems. To describe these methods we may use a model of an EDP system consisting of guards, walls and doors. Different areas (in the abstract sense) are separated by logical walls. This

means that an ongoing process has access to the facilities in one area; to be able to perform processing belonging to another area the process must pass the wall through a "logical door". The door is opened if the process satisfies specified requirements. This is controlled by a "logical guard" who operates the door.

In a ticket-oriented system the user (or user program) has a ticket (a binary word). The ticket is controlled by the guard (i.e. so as to satisfy a defined condition) at the door to an area to which the user is authorized. A drawback to a ticket-oriented system is that the user must have as many tickets as the number of areas to which he is authorized.

In a list-oriented system the guard possesses a list of those users who are allowed to pass the door. The user identifies himself and the guard searches his list to see if he is included. The operation is of course done automatically. A drawback of a list-oriented system is the time delay due to the repeated list searches. Another is the vulnerability of the lists.

Most EDP systems have a combination of ticket and list-orientation The user identifies himself and is then (via a list) given a number of tickets for temporary use during the processing.

The model with walls, doors and guards is appropriate for the basic protection mechanism described above in the sense that it describes how the access routes are controlled rather than the information itself. It is analogous to locking up secret papers in safes and drawers and then equipping the users with a system of keys. In addition to basic access control, it may, moreover, be necessary to protect the information as such. This can be done by the use of encryption. This means that the information from the source is converted (translated, if you will) into a form which is readable only by authorized users who possess keys to decrypt the data. Encryption is described in more detail below.

More about methods can be found in refs. (1) and (2).

3. Identification

Basic to all methods of information protection are reliable identification methods. The information source, whether it is a human being, a computer terminal, a computer or a program, must be properly identified. It is equally important, however, to identify the destination of the information.

Identification may be described as an exchange of common information. The source is emitting some information which it has in common with the destination. In the case of human beings this information may be contained in something the person has (e.g. an identification card) or knows (e.g. a decimal number) or it may be a personal characteristic (e.g. signature or fingerprint). The most reliable method today is a combination of the first two. The person to be identified possesses a card or a similar piece of hardware containing an identification number. He also remembers another number (with fewer digits) that is a secret transformation of the number on the card. During the identification process the card number is transformed and the result is compared to the memorized number. If they coincide the person is surely the legitimate owner of the identification number on the card.

The most effective methods of mutual identification are handshaking procedures. Handshaking means that information is transmitted back and forth between the source and destination. This obviously facilitates the most reliable form of exchange of common information.

A note of warning is appropriate here: Always choose identification numbers randomly! A person selecting his own identification number very often chooses a familiar number such as his birthday, age, street number, 1984, 4711 or anything like that. Though easy to remember, such numbers are also easy to guess.

4. Information transformation

As mentioned before, the ultimate goal is to protect the information contained in the data. To achieve better information protection than is accomplished by using data protection alone we may employ information transformations. These may be divided into two groups: invertible (or information-preserving) and non-invertible (or information-reducing) transformation. The first type of transformation (with secret inverse) is usually called encryption.

Encryption (or enciphering) of information has been used for several hundred years to protect the contents of transmitted messages. When it is used in EDP systems two new problems arise. The first is that EDP systems have more than one information source and more than one information destination. This causes some complications which, however, may be overcome by generalizing the methods used in single-user cryptography. The second problem is that the central processing unit in general cannot process encrypted data. We have to decrypt at least those parts of the data which are essential for processing. This is a potential risk because the data are processed in readable form and because the key (see below) must be stored in or transmitted to the primary memory or registers.

It would take too long to describe encryption methods here. We merely mention that the U.S. National Bureau of Standards is in the process of standardizing a particular method (see ref. (4); the standard will probably be in operation in the fall of 1976). This method uses 64 bits for the input and output words and 56 bits for the key. Thus there are 2^{56} different transformations. Which one is used is known only to those who know which key is used. The transformation is contained in a piece of hardware or software and the particular transformation is chosen simply by "inserting" the key, i.e. the 56-bit binary word. Decryption is done with the same unit.

Non-invertible transformations can be used when some information contained in the original data may be destroyed. This is for example, the case when statistical data are published. In this case we may want to publish data about the persons in a survey without revealing the identity of the persons. Non-invertible transformations may be used directly at the source (e.g. randomized response) or in the EDP system. In the latter case one method is to use stochastic transformation. This means that the data are transformed randomly, so that the information to be preserved is disturbed as little as possible but the unwanted information is heavily corrupted (see ref. (4)).

References

(1) Schroeder and Salzer, The Protection of Information in Computer Systems. Proc. IEEE, Vol. 63, No. 9.

(2) Hoffman (ed.), Security and Privacy in Computer Systems. Melville Publishing Company, Los Angeles 1973.

(3) Feistel, Notz and Smith: Some Cryptographic Techniques for Machine-to-Machine Data Communication. Proc. IEEE, Vol. 63 No. 11, pp. 1545-1553.

(4) Ingemarsson, Stochastic Transformation to Preserve Anonymity in Stored Data. Report No. 4 in Confidentiality in Surveys. Department of Statistics. University of Stockholm, August 1975.

4.5 RECORD LINKAGE IN LONGITUDINAL AND CORRELATIONAL RESEARCH: ITS JUSTIFICATION AND IMPLICATIONS FOR INDIVIDUAL PRIVACY[1]

by

Robert F. Boruch

1. Introduction

In social survey research, the respondent's identification ordinarily serves as an accounting device. Both identification and associated data are maintained under the proviso that they will be used only for research purposes and, in particular, will not be used to make personal judgements about identified individuals. Despite the proviso, the need for identifiers may bring social research into sharp conflict with the law and social custom. This paper deals with two features of the conflict - the products of such research and the way in which privacy of the respondent can be assured regardless of the product.

In particular, Parts 2 and 3 concern longitudinal and correlation research in which identifiers are normally deemed essential. There is a special emphasis on the practical consequences, including loss and distortion of information, engendered by thoughtless abridgement of one's ability to track individuals over time. And because the social benefits of the research will often clearly offset the privacy depreciation effects, there is a special emphasis on benefits of the research product. Our illustrations are taken from medical studies, economics, education, psychology, and sociology.

Part 4 briefly covers a topic which is already familiar to some of you - the privacy problems implied by social research efforts in general, and by the illustrations in particular. In Part 5, some general strategies for resolving problems are laid out, together with a few examples of their application. Here too the discussion is brief but broad in its coverage of procedural, statistical, and law-based solutions to the problems. The main theme here is minimizing degradation of privacy without preventing good research designed to better understand human behavior.

2. Longitudinal Inquiry: Its Definition, Justification, and Bearing on Record Linkage

Longitudinal research refers here to the process of tracking a group of individuals over time to establish how the state of that group varies and, more importantly, to establish the average relation between an individual's state at one point in time and his state at some other time. For example, one may conduct a study of adults to learn not only how health status of the group changes with age, but also to understand how the individual's health at one age is correlated with status at a later age. Obtaining an accurate characterization of this sort is necessary for describing and predicting health status. And it is crucial for the more demanding task of explaining the bio-social mechanisms which underlie health status development.

---

[1] Background research for this paper was supported under a contract (NIE-C-74-0115) with the National Institute of Education.

Usually, this methodology requires that an observation on a person at a particular time be linked with observations made on that person at subsequent times, for each person in a sample. The vehicle for linkage is typically, though not always, the individual's identification. The linkage implies some degradation of privacy, and so it behooves us to ask why such research is justified, to ask what we can learn or have learned from such research.

In the following remarks, some evidence bearing on these questions is presented. Section 2.1 covers some of the logical traps in which we can easily be ensnared if we choose not to do longitudinal research. Sections 2.2 and 2.5 consider a few discrete examples of longitudinal research and their products.

2.1 Traps, Artifacts, and Circularity

One of the simplest ways to illustrate why longitudinal data may be essential for even primitive understanding is to compare it with a (ostensibly equivalent but) less demanding mode of data collection. Cross-sectional studies, for example, have been suggested as a way of learning as much about human behavior as longitudinal investigations. And because they involve observation of a large sample at only one point in time, they are said to degrade privacy to a lesser degree than the longitudinal approach.

Consider, for example, the problem of understanding how intelligence (or certain intellectual achievement) varies with age. One might conduct a survey of a sample of children of age 3, say, and then continue to survey those individuals annually until they reach an advanced age. Or, in the interest of saving time and perhaps on privacy grounds, we might choose to conduct a single survey of a representative sample of (anonymous) 3-year-olds, a sample of 4-year-olds, and so forth at only one point in time, under the assumption that this cross-sectional survey would yield roughly the same results as the longitudinal survey. This last assumption, that a growth curve based on longitudinal data will be roughly equivalent to a growth curve based on cross-sectional data, is critical.

The assumption also happens to be wrong with alarming frequency. In particular, its espousal by some human-development experts has led to some erroneous, not to say embarrassing, folklore about the development of human intelligence. The same assumption has been a trap in some economic welfare research, in some epidemiological work, and in other areas.

To understand one of the logical traps here, consider Figure 1a, a chart commonly used during the 1940's and 1950's to illustrate the gradual increase in IQ from childhood to early adulthood, and the gradual decrease thereafter. The implication of the graph, which is based on actual cross-sectional data, is that at age 30 one's IQ is at its peak, and things go downhill soon after that. What makes the chart much more persuasive is that similar inverted-U patterns show up in other investigations of human ability based on cross-sectional data. This includes the quality of treatises written by eminent philosophers (rated by eminent philosophers) when plotted against the age at which the author wrote the document. And it includes the level of innovative-

Figure 1. Examples illustrating the confounding of age and cohort differences in cross-sectional research

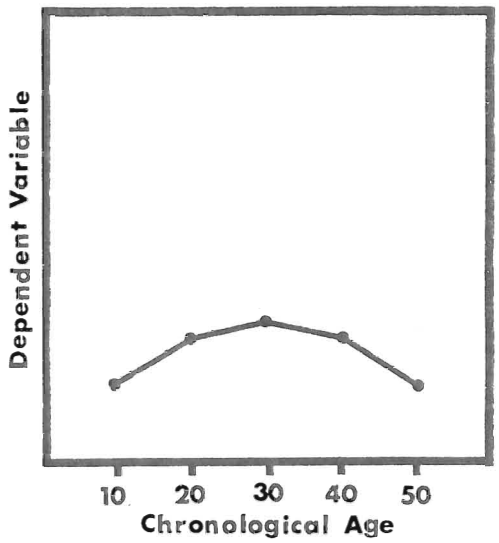


Figure 1a

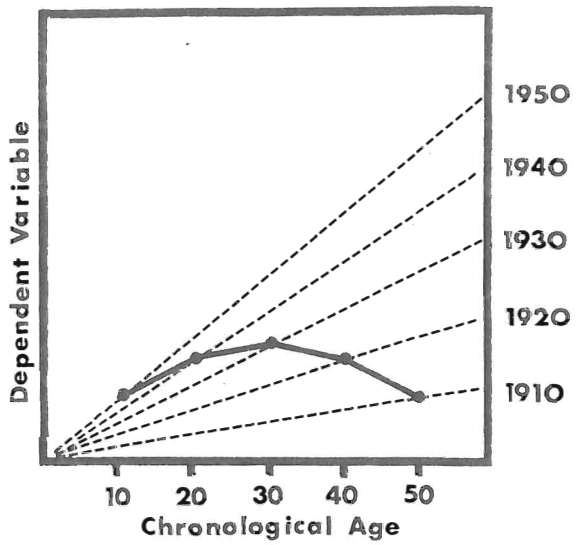


Figure 1b

Source: J.R. Nesselroade and Paul B. Baltes, Adolescent personality development and historical change: 1970-1972. Monographs of the Society for Research in Child Development, 1974, 39, (1, Serial No. 154), p. 4.

ness of theory and invention of chemists when plotted against the chemist's age at the theory's production, and similar data (see, for example, Birren [1964]).

Suppose now that instead of the cross-sectional data, there existed longitudinal data on exactly the same individuals. The dotted lines in Figure 1b illustrate how actual IQ may increase consistently with age without a notable decline, and how the rate of increase can depend on year of birth, i.e., on cohort. The points connected by the solid line correspond exactly to what appears in the plot of cross-sectional data. The chart suggests that individuals born in 1910 increase in intelligence as they grow older. But their rate of increase is lower than the corresponding rate for a younger cohort, e.g. individuals born in 1930. The reasons for differences in development rate, or "cohort effects", are a matter of speculation. They may involve any number of bio-social factors; the differences may even be an artifact of the increasing reliability or culture-relatedness of such tests. Regardless of the reasons, the point is that the longitudinal data offer us a less misleading picture of human development than the cross-sectional data. Moreover, the theory generated by the former will differ markedly from theory generated by the latter. It is clear that relying solely on the cross-sectional data can lead one to a conclusion which is quite contrary to the way nature behaves. In point of fact, there is reliable evidence from studies by Barton et al. [1975], Schaie [1965], and others that Figure 1b is a more realistic portrayal of nature than Figure 1a.

> Exactly the same inferential problems occur for a variety of physical and social measures of individual characteristics. Plots of height, for example, when plotted against age, often show an inverted-U pattern if based on cross-sectional data, simply because rates of growth and upper limits on growth are quite high for children recently born, relative to the growth rate and upper limits for those born 80 years ago. Plots of cross-sectional data on level of extroversion and age of adolescents in certain areas of the United States make it appear that extroversion declines through adolescence when it actually increases on the average and increases most quickly for recently born cohorts. Longitudinal data on adolescent tough mindedness (autonomy, assertiveness) suggests a fair degree of stability over ages 12-15; more recently born cohorts generally exhibit higher levels of trait. But cross-sectional data show a declining trend.[2]

Some of you may regard "soft" social data, like psychological measures, as particularly susceptible to the inferential trap just described. The fact is that even data on "hard" social variables, such as income, are no less immune to the problem. Consider, for example, estimates of lifetime income for individuals. These predictions are important in the commercial arena, e.g. in some credit and loan research in the insurance business. And they are no less important in the government sector, e.g., in planning social security benefits and the like. Often there is a choice between using cross-sectional data or using

---

[2] For empirical data, theory, and policy implications of longitudinal studies in development, see Schaie [1965], Wohlwill [1969,1970], and Magnusson, Dunér, and Zetterblom [1975]. The results of a variety such studies in Britain, the United States, France, and Germany are summarized by Wall and Williams [1970].

longitudinal research, and if both provide equally accurate estimates then one might choose the cross-sectional approach for managerial reasons or on grounds that a cross-sectional survey involves less degradation of the privacy of individuals since one can presumably elicit anonymous responses. Miller and Hornseth's [1970] attempts to estimate lifetime income for certain segments of the population, is interesting in this respect.

> That estimates of lifetime income based on the two kinds of data will not be the same is clear from Tables 1 and 2. Table 1, based entirely on cross-sectional survey, suggests that annual income increases up until age 35, stabilizes during the 35-54 year age interval, then declines. The pattern is similar whether one considers data collected in 1947, or 1948, or 1949. Table 2, on the other hand, is based on longitudinal data and illustrates a much less drastic pattern, notably that increases in income persist over a wider age range, and rates of increase are substantial. The longitudinal data are, of course, affected by inflation and other factors uniquely associated with a given cohort, but similar patterns occur after adjustment for inflation. They are more accurate than the cross-sectional data in the crude sense that they better describe the way observable income behaves as a function of age.

Though the example is recent, the problem of estimating lifetime earnings from cross-sectional data is not a new one for economists. Klevmarken [1972] gives a tidy and brief description of the history of the problem in this context and points out practical needs for better estimates in labor negotiation, actuarial sciences, and elsewhere. More important, he has managed to show, using both longitudinal and cross-sectional data, how one could develop less misleading models of lifetime income curves if one had available only the cross-sectional data. He makes the same point as we do, however, in observing that there is no generally reliable way to establish longitudinal trends from cross-sectional data alone. Any attempt to do so must be based on assumptions which, for the social scientist, may easily fail to be met in reality.[3]

A different but no less important trap is the failure to recognize that longitudinal rather than cross-sectional data may be essential for detecting subtle influences on human behavior. The problem of designing precise investigations is particularly important in estimating the impact of social programs whose effects, we know, are often weak but may nonetheless be politically important. Achieving that objective often depends on the availability of longitudinal data. There is a large array of analytical techniques, for example, which employ the correlation between behaviors at different points in time to expunge irrelevant variation from the data. The use of longitudinal research techniques, especially in conjunction with randomized experiments, usually makes it easier to detect influences which might otherwise be obscured by the normally high variation in human behavior.

---

[3] Cohort effects have been recognized only recently by commercial market researchers as an important variable in predicting and explaining the demand for certain consumer goods. Systematic cohort variation in what is regarded as a luxury item, for example, has some important implications for planning the allocation of an industry's manufacturing resources (see Business Week, January 12, 1976, pp. 74-78).

Table 1. Estimates of Mean Annual Income in Dollars for Age 25 through 64, Based on a Cross-Section of Men Sampled in 1947, a Cross-Section Sampled in 1948, and a Cross-Section Sampled in 1949

Year/Age	25-34	35-44	45-54	55-64
1947	2704	3344	3329	2795
1948	2898	3508	3378	2946
1949	2842	3281	3331	2777

Adapted from data presented by Miller and Hornseth [1970].

Table 2. Estimates of Mean Income in Dollars over 10-Year Intervals for Six Cohorts of Individuals

Year/Age	25-34	35-44	45-54
1. 1947	2704 (1947)	5300 (1957)	8342 (1967)
2. 1948	2898 (1948)	5433 (1958)	8967 (1968)
3. 1949	2842 (1949)	5926 (1959)	9873 (1969)

Year/Age	35-44	45-54	55-64
4. 1947	3344 (1947)	5227 (1957)	7004 (1967)
5. 1948	3508 (1948)	5345 (1958)	7828 (1968)
6. 1949	3281 (1949)	5587 (1959)	8405 (1969)

Note. Each cohort has been surveyed every 10 years. The first cohort, for example, contains individuals who were 25.34 years of age in 1947 and had an average of $2,704; in 1967, when they were 45-54 years of age, their mean income was $8,342. Adapted from Miller and Hornseth [1970].

> Consider, for example, the Cali, Colombia experiments on the impact of nutritional supplements on children's physical growth. Special nutritional supplements were assigned randomly to a sample of malnourished children; supplements, which were in short supply, were unavailable to an otherwise equivalent sample of comparison group of children. The impact of the supplements was not evident from scrutiny of mean changes in treated and untreated groups; the simple natural variation in heights of even malnourished children is sufficiently large to obscure real differences. More sophisticated analyses, using correlations between repeated measures of height of the children, did yield estimates of program effect which differed notably from chance level. As a consequence of the positive finding, the supplements are being improved, put into local production, and tested on a much larger scale in three other less well developed countries. (Bejar [1975]; Sinesterra, McKay, & McKay [1971]).

The same use of a longitudinal approach for the sake of sensitive analysis of program effects is evident in other areas. Heber et al. [1972], for example, have conducted 6-year studies to determine the relative impact of special programs for reducing the risk of functional retardation among infants and young children; based on these Wisconsin pilot tests, similar test programs are being mounted in North Carolina and elsewhere. Beyond the midpoint in Kaiser Permanente's 10-year experiments, Ramcharan et al. [1973] find evidence for the impact of multiphasic screening on prevention of disease, an impact which is bound to be negligible during the first few years of the program. In the economic area, the Housing Allowance Experiments require 8-12-year followups to determine incremental benefits of income subsidy plans on the poor, and to provide information for effective legislation in the area. In these cases and in innumerable others (see Riecken et al. [1974] the effects may be undetectable in the short run, and difficult to detect in the long run, especially if the groups involved are quite small. There is simply no reliable substitute for longitudinal follow-ups in these instances.

The final logical trap of interest here bears on both longitudinal and correlational research; it involves the analysis of data based on aggregate of individuals in order to make judgements about individuals within the groups. To establish the average relation between literacy and race in the United States, for example, one might obtain published census statistics on the proportion of literate persons and the percentage of Negroes for each of 48 states and then compute the correlation between the two variables. Aggregated data might be used here on grounds that the relevant information is easily accessible from published tables. Or, we might justify our action on grounds that the use of published data does not present the privacy-related problems which might be engendered by a special survey.

There are two weaknesses implicit in the argument that aggregate data can be used in lieu of individual data. The more obvious one is that inferences made about groups are not necessarily appropriate to the individual and in fact may be quite inaccurate. The second weakness, more a matter of precision than accuracy, is that analyses based on grouped data are often considerably less likely to reflect changes in individuals than analyses based on data at the individual level.

To be more specific, consider the literacy-race example. At a particular point in time the correlation between literacy rate (percent literate) and color (percent Negro) computed on the basis of the nine census regions of the United States is .95. When individuals are grouped by state rather than region, the correlation is .77. Finally, when individuals are not grouped at all, but the entire disaggregated population is considered, the correlation is .20. (The example is from Robinson's fine paper [1950] on census data prior to 1950.) A similar problem with a different resolution appears if we try to determine the relation between color (white-nonwhite) and occupation (domestic service - other) for female employees in Chicago in 1940. Though a correlation based on percentage data for each of nine areas is .34, the actual correlation based on individuals is .29, not too different from the area-based estimate (see Duncan & Davis [1953]; Goodman [1953]).

In the literacy-race example, the high correlation obtained from the regional data might be interpreted as suggesting that illiteracy is pervasive among blacks, and furthermore, that a massive program of education must be put into effect to counteract the problem. In point of fact, if we look at indiduals' data, rather than at data based on opportunistic groups into which individuals may fall, we reach a considerably less pessimistic and a more accurate conclusion: that the relation between race and literacy was small but notable. Any attempt to resolve the problem of illiteracy by making a massive investment in rehabilitating the reading skills of each individual based on the .95 regional correlation, is bound to be a wasteful allocation of scarce resources.

An obvious problem in these matters is the use of aggregates of individuals as a surrogate for individual persons. Since the aggregates are usually constructed for political or administrative purposes (e.g. census regions, health care service regions), it is unlikely that these "natural" aggregates will constitute valid replicas of real persons. We have only a little theory to guide us in selection of "proper" aggregates. And it is impossible to predict whether an aggregate will be proper without some data at the individual level.

The problem is a chronic one in the social and administrative sciences which must rely heavily on aggregated data - sociology, epidemiology, economics, statistical geography. It is particularly crucial in attempts to evaluate the impact on national social programs on individuals. Many such evaluations, in education for example, rely on data aggregated at the school district level to estimate the impact of a nationally supported compensatory reading program for disadvantaged youth. The inferences made about individuals (based on analysis of aggregates rather than on individuals) are generally biased in an unknown fashion (the individual data not having been analyzed), and are imprecise because the aggregate data are insensitive to changes, even some marked changes, in individuals (see Burstein [1976] for examples). This is not to say that the aggregation problem will always yield biased estimates. It is to say that the problems are crucial and cannot be resolved unequivocally without some evidence based on individual rather than aggregated data.

2.2 Medical Research

There is a fine tradition of longitudinal studies in medical research, dating at least from Hippocrates's efforts to characterize the progressive stages of disease among his own and his colleagues' patients (King [1971]). The systematic tracking of both the healthy and the ill remains a basic weapon in medical research armamentarium. At its best, the approach not only helps to identify the existence and incidence of a disease entity, to determine symptom development and disease consequence, but it is essential in laying out the array of possible progenitors of the disease. Longitudinal methods in this sector have become considerably more efficient over the last 40 years with the development of survey sampling technology. And when coupled to other methods, such as randomized experimental tests, the approach can be dramatic in identifying whether and how well particular treatment programs work.

Examples of the process are not hard to find. But for the sake of detail, suppose we examine a complex research area which, by virtue of gifted science writers (such as Gilmore [1973]) and researchers (such as Kannel et al. [1961]), is among the best documented. Modern work on coronary heart disease appears to have reached a turning point during the 1940s and 1950s with autopsy studies. Those investigations, because of their small size and cross-sectional nature, provided thin support for the linkage among natural development of arteriosclerosis, heart disease, and bio-physical conditions (blood pressure, etc.), and more importantly, provided the evidence necessary to justify longer term longitudinal study of the problem. The Framingham Study (Kannel et al. [1961]) among the largest of subsequent efforts, was designed to better establish relations between prior condition and subsequent death due to heart attack. Spanning 25 years in the lives of 9,000 men, the effort was of sufficient size and duration to permit computation of risk factors operating in the population: Actuarial tables were developed to illustrate the likelihood of heart attack as a function of earlier serum cholesterol level, blood pressure, EKG abnormalities, and so forth. Other studies - animal experiments and comparative investigations of populations with natural differences in these factors - yielded evidence which added to speculation about the role of serum cholesterol level and other factors in heart disease.

Because the ability to describe and predict based on longitudinal research does not necessarily yield unequivocal information on causes of the disease, and because study of human population yields results which are similarly ambiguous, long-term, experimental tests of alternative treatment programs have been mounted. The best of those tests generally involve large samples tracked over long time periods and, moreover, randomized assignment of individuals to one of the competing treatments. As a consequence, they raise problems more serious than those engendered by longitudinal research alone. Nonetheless, pilot efforts, such as the Diet Heart Feasibility Study, have been completed to furnish data on the practical difficulty of field tests and somewhat less equivocal small-scale data on the impact of diet control on heart disease. Such short-term (two-year) studies have paved the way for longer term studies which focus on the more plausible and tractable causal mechanisms, notably reduction of heart disease through diet or drugs which reduce serum-cholesterol levels. The largest of current clinical trials will run five years and involves over 50 institutions and 8,000 patients; it is designed to evaluate the effectiveness of alternative drugs and

drug dosage level for reducing cholesterol level in the bloodstream (Coronary Drug Project Research Group [1973]). Although the primary response variables are mortality rate due to heart disease and related illness, a variety of social, biological, and physiological measures are being obtained. The social measures - smoking habits, lifestyle measures, race, job characteristics, and so on - are expected to add precision to results and to help identify variables which though influential are less amenable to direct control.

The products of earlier longitudinal studies coupled to experimental tests are readily accessible (see Boruch & Riecken [1975]; Riecken et al. [1974] and references cited therein). Long-term followup of released prisoners who have had cosmetic surgery to remedy facial disfigurement has given us evidence for lower recidivism rates among prisoners so treated. Longitudinal experiments on the effectiveness of physician surrogates - nurse practitioners, physician extenders - has yielded information essential for reducing costs of medical service, for planning innovative programs in health care utilization, and the like. A new generation of pharmacy research focuses on both short-term and long-term drug-taking behavior - determining level of patient compliance with medication regimens, determining how special packaging of medication influences compliance especially among the elderly, and so on. In preventive medicine, tests of the impact of multiphasic screening by Kaiser-Permanente, are being run for a 10-year period to assure that long-term effects of annual screening on detection and amelioration of disease are well documented.

In brief, these examples and others like them teach us that there is no way to establish etiology of disease or to evaluate the effectiveness of prevention and treatment programs without longitudinal study. Not that longitudinal research is sufficient. Its natural limitations must usually be broadened by coupling this approach to others, notably experiments, designed to establish cause-effect relations. But the idea is central to medical research and, in principle and in practice, generalizable to other areas.

2.3 Psychology and Psychiatry: Biochemical Bases of Schizophrenia

For the past 100 years, the scientific and lay arguments over the causes of schizophrenia have been supported largely by ambiguous data. The information at its worst has been unreliable and no more than anecdotal in form; at its best it has been based on longitudinal study of very small numbers of individuals and heavily reliant on retrospective reports of unknowable reliability. The debate's focus has changed markedly during the years, however, in part because of longitudinal research which depends heavily on record linkage (Mednick & McNeil [1968]; Mednick, Schulsinger, & Garfinkel [1975]).

One of the basic problems in discovering the origins of schizophrenia, as many of you know, is to disentangle the biochemical causes of the problem from the environmental influences. To resolve the problem, researchers at Denmark's New School for Social Research, at the Psykologisk Institute (Copenhagen), and at the Kommune Hospitalet (also Copenhagen) have conducted longitudinal studies of over 4,000 adopted children to discover how incidence of schizophrenia among them varies with occurrence of schizophrenia in their natural families and in their

adopted families. If, for example, the schizophrenia among children born of schizophrenic parents but raised by adopted nonschizophrenic parents is high, then one has more reason to believe that the malady's origin has a genetic component. Schulsinger's findings, obtained in collaboration with David Rosenthal, Seymour Kety, and Paul Wender of the United States are that:

- The incidence of schizophrenia is substantially higher among adopted children who had schizophrenic natural parents than among adopted children whose foster parents were schizophrenic.
- There is a very low incidence of schizophrenia among children who had schizophrenic parents than among adopted children whose parents were schizophrenic.
- If children whose natural parents are not schizophrenic are later adopted by a schizophrenic foster parent, there is no increase in likelihood that the child will become schizophrenic.

The information is an elementary but important step in establishing the credibility of the idea that the origins of schizophrenia are partly environmental and partly genetic, and it is important in directing attention to more fertile areas of research. The latter include careful studies of the possible genetic mechanisms and of the role played by certain enzymes (for example) which may produce a predisposition toward schizophrenic behavior.

These findings could not have been made without longitudinal data on adopted children and their natural and adopted parents, and without the crucial linkage among existing medical records, social service records, and followup data collected more recently on the basis of the national address registry.

2.4 Longitudinal Study in Manpower Economics

In human resources research, good evidence for the usefulness of longitudinal data has been scanty in part because the relevant data have been in short supply. The recent buildup of longitudinal files has helped greatly to understand the data's benefits and limitations, however. Of particular interest are the National Longitudinal Surveys (NLS) of the U.S. labor market, begun in 1966 by Herbert S. Parnes [1975]. Those data are based on repeated surveys of a national probability sample of 20,000 individuals in four labor market strata: middle-aged men (45-59 years old at the survey's beginning), women (30-44 in 1966), young men and young women (14-24 in 1966). The resultant data are being updated periodically and, stripped of identifiers, are being made available to the community of manpower researchers. Aside from their obvious benefits for temporal description of the labor market, the data can be very informative on account of their longitudinal feature. Parnes [1975] maintains that:

> Perhaps the single most important contribution of longitudinal data is that they facilitate the identification of causal relationships that cannot confidently be identified in any other way. Take, for example, the relationship between attitudes and behavior. In

> cross-sectional data, such relationships are ambiguous, since one cannot be certain whether the attitude produces or reflects the behavior. Does job dissatisfaction lead to turnover, or does an association between the variables simply mean that individuals who quit jobs are likely to rationalize their behavior by reporting (retrospectively) that they were unhappy? When attitudes measured at one point in time can be related to subsequent behavior, such ambiguity disappears. The NLS data for middle-aged men have clearly demonstrated that the degree of job satisfaction predicts the likelyhood of a voluntary job separation and that a commitment to work in general, as well as satisfaction with one's particular job, decreases the likelihood of early retirement.
>
> The usefulness of longitudinal data in clarifying causal relationships is, of course, not confined to instances in which one of the variables is attitudinal. For example, finding that the receipt of training by middle-aged men between 1966 and 1971 was associated with a net earnings advantage in 1971 (controlling for such other factors as education, health, and region of residence) Avril Adams went on to demonstrate that the trainees-to-be had already enjoyed higher earnings in 1966 (again controlling for the same variables). Thus training was found to be a selective process, presumably attracting the more highly motivated or otherwise more productive individuals. To put the matter differently, some part of what would doubtless have been identified by a cross-sectional analysis as training's contribution to earnings was found to have reflected an incompletely specified model - i.e., the failure to control adequately for factors associated both with earning and the probability of receiving training. (Parnes [1975] pp. 246-247.)

Professor Parnes is optimistic about the fruits of research based on his data files. We do not share that optimism, since longitudinal data alone is often insufficient to make unequivocal judgements about the impact of manpower training programs. We do agree that such data are essential for better understanding and prediction of gross labor market behavior, for establishing tentative hypotheses which can be later verified using more controlled studies, and for prediction.

2.5 Longitudinal Study in Child Measurement

The National Child Development Study (NCDS) began in 1958 with a survey of some 5,000 pregnant women. Its main objective (like the United Kingdom's earlier study, the 1946 Population Investigation Committee survey) is to establish the linkages among prenatal conditions, environmental factors, and growth of young children. According to Wall, the results of 1966 followup data on over 9,300 children showed, despite strong suspicion to the contrary, the following variables are not singly predictive of lowered reading ability: maternal hypertension, breech presentation or forceps delivery, Caesarean section. Further, there is an unexpected and strong relation between departure from normal gestation and reading and social adjustment test scores at age 7; gestational maturity is a better predictor than the more commonly accepted birth weight measurement.

The Population Investigation Committee Study yielded other conclusions which could not have been reached without longitudinal data. Taken verbatim from Wall and Williams ([1970], pp. 42-43), we have:

That the effects of social mobility and increasing material prosperity have differential effects according to the educational levels of the parents and the number of children in the family.

That a relatively poor social environment is cumulative in its effect on children's height, girls being more sensitive to this than boys.

That separation from mother, as well as being much more prevalent than had been thought, seems (in the period from birth to five years) to provoke less serious permanent disturbance than might have been expected from clinical studies of a post hoc kind.

That by the age of 5, broken homes apparently do not provoke more than temporary disturbance (bedwetting), and this only in non-manual families.

That the proportions of mothers taking up full or part-time work increased as their children approached the age of 5, but that there was no evidence that their children were less emotionally stable at this age.

That early toilet training leads to earlier bowel control, less bedwetting, and less breakdown later.

That a high proportion of bedwetters bite their nails and have speech defects, difficulties which persist even after they become dry.

That children prematurely born are more vulnerable physically during their first two years but not afterwards; that although rather smaller than normal children in later childhood, this is not the result of prematurity, and that by the age of 8 they tend to be handicapped in mental ability, particularly in reading. (Wall & Williams [1970], pp. 42-43).

2.6 Education and Its Impact

For the sake of better allocation of scarce resources to education, it is reasonable to learn how education affects academic achievement, and subsequently, earnings. We need to know what the most effective elements of the education process are, how they work, and how they affect the individual's intellectual and economic development. The impact question is especially relevant to novel programs designed to overcome disadvantages under which some social groups labor, i.e. designed to introduce more equity into the social system through education.

Most research designed to get at these issues begins with cross-sectional surveys and even a brief incursion into history shows that these have been useful despite the limitations of the cross-sectional approach. Abraham Flexner's 1910 report of his studies of medical schools in the United States relied solely on this methodology and on Flexner's standards of performance to produce a major reformation in medical training. The Thorndike and Ayres studies of school record

systems were similarly useful in moving schools toward higher quality (though still imperfect) record-keeping practices (Goslin & Bordier [1969]). The cross-sectional studies have been, and still are, enormously useful in this context, especially in clarifying the scope of educational problems, especially where standards of quality are fairly clear.

But the difficulty of making inferences about the impact of education, of understanding individual growth, based on cross-sectional data, are no less severe in this sector than they are in the medical arena. It is difficult, often impossible, to discriminate accurately between the influence of background variables and those of the school. It is not generally possible to lay out growth and assay impact without at least some longitudinal data. The scientific and political traps here are exemplified by the current U.S. controversy over busing students from the school district in which they live to another in the interest of fostering equitable, quality education. James Coleman's advocacy position five years ago, based largely on cross-sectional data, is considerably different from his opposition now based on longitudinal data and demonstration projects.

Better interpretation, inference, and prediction are conditional on better theories (models) for simulating social behavior and on data necessary to support that theory. As a consequence of the shortcomings of earlier data, a number of longitudinal studies have been mounted to better understand education impact. We cannot summarize those here - there are far too many to do so reasonably. So we content ourselves with examining a nice study by Fägerlind.

This analysis of longitudinal data was designed, in part, to clarify polar views of the results of public investment in education: that duration of educational experience has a major import on earnings, a view exemplified by Nobel Laureate Paul Samuelson, and Jenck's (and others') view that the impact on earnings of education beyond the post-secondary level is marginal. The Fägerlind research manages to avoid the traps of cross-sectional data and of short-term longitudinal study, by considering individual growth over a 30-year-period. It is an efficient study of a well defined social group insofar as it builds on survey data initially collected in 1930 on a subpopulation of children in Malmö. It obtains both economy and completeness of sampling through the use of population registries for follow-up surveys. Accuracy and temporal relevance of data are enhanced by relying on archival records - military selection test-scores for men, data from tax registries on earnings of the respondent and the respondent's parents, and data from census records on demography, geographic and occupational mobility, etc. Fägerlind supplemented archival records with survey data collected during the 1940's, '50s, '60s, and early '70s.

The product of this particular research is interesting not only in adjudicating polar views: Fägerlind's data, of higher quality than Jenck's, supports Samuelson's theory. It also helps to specify the process, the mechanism underlying education's impact on earnings through an otherwise tangled mass of competing influences such as home, family, and so on. And it has helped to understand shortcomings of competing data and models: quality of education, for example, has been ignored in many such analyses and this one uncovers strong, plausible linkages between quality and earnings from age 30 onwards.

Still, the longitudinal approach used here is only an interim step. It is naturally limited in the extent to which it can be applied in specific, particularly novel settings. More recent research, for example, stresses small longitudinal experiments, mounted alone or in conjunction with larger observational studies, to obtain finer appraisals of innovative educational programs and practices. Some, like the Heber et al. [1972] work is dedicated toward inhibiting intellectual deprivation from infancy. Others, like Middlestart, involve randomized tests of programs designed to improve academic performance of adolescents who are unusually deprived by virtue of their very poor economic condition. Still other experiments, designed to improve medical school education, police training, and manpower training and the like, follow participants through adulthood in the interest of obtaining less ambiguous information about the short-term impact of expensive and specialized education programs (see Boruch & Riecken [1975]; and Riecken et al. [1974]).

3. Correlational Research: Definitions, Justification, and Relevance to Record Linkage

Correlational research refers here to the process of establishing how two characteristics of an individual are related to one another. The average relation, for a large sample of individuals, may be represented in statistical form by a simple correlation coefficient, by the probabilities in an actuarial table, and so on. For example, to identify the relation between level of health status and level of physical activity during work, one might obtain measures of both variables from each member of a suitable sample of individuals, link the two elements of information on each individual, then compute an index of the relation based on that linkage. The correlation may be of descriptive interest alone in that it reflects the existence and strength of a relation between two variables. It may be more important to an individual, in that the correlation helps to predict future health status from current physical exertion levels. Finally, such data make it possible to form tentative ideas about the biochemical mechanism by which exertion influences health status (or vice versa), i.e., to build theory necessary for the development of better control of health status.

In principle, correlational investigation is a general activity of which longitudinal research is an important subclass. Both types of research usually require some form of record linkage to sustain statistical analysis. They are discussed separately here on account of traditional differences in the emphasis of each type of research.

Correlational research often requires that the contents of records which are maintained by independent archives be linked. The special functions of linkage vary considerably, but most can be grouped into one of the following categories:

- To assess and improve the quality of available data from any source;
- To reduce costs, duplication of effort, and respondent burden in surveys;
- To clarify and enrich the data base for applied social research and policy analysis.

The illustrations of the benefits and limitations of linkage are presented below using this taxonomy to organize our experience.

3.1 Assessing the Quality of Data and Improving the Quality of Data Analysis

"Response validity" refers to the association between an individual's response to inquiry under one set of conditions and his response to inquiry under a second set of conditions which are thought to facilitate near-perfect reporting. Most such studies involve one form or another of record linkage. Census data on income of identifiable respondents may be linked to Internal Revenue Service reports, for example, to assay the adequacy of the census interview process. Data from interviews made on one occasion under normal conditions may be linked similarly to later, more intensive, interviews to gauge the adequacy of the "normal" interview conditions. Some mechanism for linkage is critical for computing quantitative indices of average agreement between the two types of reports.

Without some empirical basis for judging the data's credibility, it is impossible to lend any meaning to statistical analysis, unless of course, one is willing to cover the whole matter with a secular act of faith. The absence of validity statistics is especially crucial not only in interpreting descriptive statistics but also in using them to monitor and evaluate social programs. Errors in reporting will usually make it more difficult to detect changes in human status, and in situations where data imperfections go unrecognized, data analysis may result in wildly inaccurate conclusions.

Examples from descriptive survey research. Many of the better validity studies in the U.S. have been conducted by government agencies and by university-based research groups. The studies are frequently designed to furnish sufficient evidence to support an administrative decision about whether or not to continue a particular type of inquiry.

In health survey research, for example, a good deal of the use of record linkage is reported in the proceedings of a recent national conference (Reeder et al. [1975]). The deficiencies in physicians' records, for example, have been examined by matching record content with data from interviews with patients. Distortions in reports made by physicians to their own medical societies have been investigated by linking those reports with intensive interviews subsequently conducted with physicians themselves. Methods of interview designed to minimize embarrassment in health-related surveys have been tested and evaluated using individuals' hospital records as the standard for accuracy. Surveys of health services utilization, necessary for planning such services at the national level have been validated using side studies which link individual responses to records maintained by providers and third-party payers.

Analogous examples appear in manpower research. For example, to appraise the validity of self-reported "occupation five years ago", a question which has appeared in many cross-sectional manpower surveys, the U.S. Census Bureau conducted tests on 2,800 households in 1968, for whom 1963 data on actual occupation were available. Despite the use of a variety of methods to elicit the retrospective report, the differences between retrospective report and actual status were in the range 23-28 % (Jabine & Rothwell [1970]). The linkage here, between 1963 archival records and the 1968 survey, was essential in establishing the validity rate and in the pattern of invalidity. And the statistics themselves influenced

the Census Bureau's decision to drastically reduce the use of the retrospective question in its own surveys, and to routinize the correction of other survey researchers' occupational mobility statistics.

Housing statistics are no less immune from biasing influences and, in some cases, intensive reinterviews are necessary to establish validity of initial interviews. For example, it is not unreasonable to expect that interviewers will vary notably in their ability to rate quality of housing. In testing alternative methods for assuring accuracy of the rating, the U.S. Department of Housing and Urban Development and the Census Bureau found, using reinterviews as a standard, that no particular method of interview classification yielded ratings at a reasonable validity level. And as a consequence, the rating scheme was dropped entirely in the 1960 census. Instead, crude indicators of quality (cooking facility, indoor toilet, etc.) were included in the enumerator's protocol. Again, neither the collection of validity statistics nor the subsequent administrative actions would have been possible without some mechanism for linking initial enumerator reports with more expert reinterviews.

In estimating undercounts in the census of 1960, Marks and Waksberg [1966] report both positive benefits and negligible benefits in using archival records. The use of 1950 Census records, hospital records of birth during 1950-60, records from intermediate census research, and records from the U.S. Immigration Service for special subsamples yield useful and credible evidence for underenumeration of 2.6 to 4.7 % in the 1960 census. Similarly, for special subgroups, undercount estimates were obtained. Lists of college students were obtained from colleges to estimate undercounts in the enumerated count of 2.5 to 2.7 %; Social Security addresses were used in estimating a 5.1 to 5.7 undercount in beneficiaries in the 1960 census. On the other hand, matching of census rolls against lists of relatively inaccessible individuals - list of welfare recipients, postal service listings - "provide no special encouragement for use of matching special lists as a coverage improvement program". Horwitz [1966] conducted similar studies in rural areas which suggested that 20 to 25 % under-reports in death rates and 15 to 20 % under-reports of birth rates are not unusual when hospital and state medical records are used and as a standard.

These examples illustrate how validity statistics, generated through record linkage, can help to delimit the credibility of social survey statistics and can serve as a basis for making decisions about the conduct of a survey effort.

The practice of conducting side studies such as these, based on limited record linkage, is practically nonexistent in commercial survey efforts. It is, however, typical in some governmental surveys and in research conducted by some university-based research groups. That the practice is increasing even in these sectors is evident from the bibliographies published on the topic (notably Scheuren & Alvey [1975]), from new reporting systems such as Studies from Interagency Data Linkage for describing the products of the work, and other evidence.

Examples from program evaluations. Imperfections in either social survey data or administrative records make it difficult to detect and, in the worst cases, can produce statistical artifacts which make pro-

grams appear harmful. Estimates of validity, whether based on record linkage or not, are often essential for refining the design of an evaluation to accommodate the problem.

More specifically, one of the chronic problems encountered in the United States has been the production of biased estimates of program effects under some special but common conditions. Conventional statistical techniques, such as regression analysis, covariance analysis, and matching, when applied to fallible data obtained in some observational evaluations, yield consistently biased estimates of program effects, in part because imperfect measurement goes unrecognized. Consider, for example, the Westinghouse-Ohio evaluations of "Headstart", a preschool program for the economically deprived. The initial evaluation relied on a textbook application of covariance analysis of survey data to explain how children's verbal ability varies as a function of demographic characteristics of the children and of their families, and other variables. The estimates of the impact of Headstart were actually negative, implying that the program had a harmful effect. It is clear from secondary analysis of the same data that if one adjusts the conventional analysis so as to recognize imperfect measurement, the program's effect is negligible and perhaps even slightly positive (Magidson, Campbell, & Barnow [1976]). Similar biases have been discovered in the evaluation of manpower training programs (Direction [1974]), in the estimation of the impact of special medical treatment regimens (James [1973]), and elsewhere (Campbell & Boruch [1975]).

To summarize, we observe that measures of social, psychological, medical, or economic behavior are usually imperfect. If the imperfections go unrecognized, then statistical analysis of the impact of programs designed to ameliorate relevant problems will be insensitive at best, misleading at worst. Statistics bearing on validity and reliability of response are necessary for rational adjustment of conventional statistical analyses so as to reduce bias in estimates of program impact. Record linkage is often, though not always, necessary for production of the necessary information on validity of the observations.

The view that administrative records ought to serve as the standard against which survey records are judged is, at times, clearly unjustified. Administrative records are tied to administrative action, and for that reason, are normally susceptible to a variety of biases and sources of error which do not affect survey data. One of several ways to appraise the credibility of statistics based on those records is through specialized designed surveys.

Prior to 1910, for example, studies by the noted educational researcher E. L. Thorndike on the adequacy of school records led to major reforms in school record-keeping practices. Those studies relied partly on record linkage to furnish evidence concerning deficiencies in existing record systems (Goslin & Bordier [1969]). Later studies, conducted by economists, contributed to what we now know about needs for record accuracy, publicity, and adequacy in preventing abuse of power by public utilities (see Shils [1938]). More recently, Campbell [1975] and others have tried to enumerate more fully the reasons for corruption of administrative records and to develop some crude theory to account for the phenomena. Most of the theory building depends in one way or another on the conduct of surveys to appraise the quality of an archieve's contents. The U.S. Army reporting system for drug abuses, for example, were assessed during the early 1970s using an experimen-

tal interview method which generally yields less distorted information on actual abuse by identified individuals (see Section 5). The debatable quality of criminal records maintained by police has led to Federally funded victimization surveys, conducted by the Census Bureau to determine the nature and incidence of unreported crime, the elasticity in police definitions of crime, and so on. These more recent examples do not depend on record linkage to make their point. But whether a social scientific survey can be mounted to verify the quality of an archival record system depends heavily on administrative endorsement of the idea that multiple indicators of a period that are desirable. As the practice of conducting this kind of study increases, the need for more depth of inquiry and, consequently, linkages between archival record and survey record will undoubtedly increase. It is often possible to eliminate confidentiality-related problems in this context by using the insulated data bank strategy described in Section 5 below.[4]

3.2 Reducing Costs, Duplication of Effort, and Respondent Burden

Partial duplication of a data collection effort by several agencies may be justified on several grounds. Independent archives which maintain some overlapping information, for example, may be warranted by legislation which requires independent collection and maintenance of the data, they may be justified as a device for periodic cross-validation of the contents of files. Nonetheless, exact or nearly exact duplication may be costly to the data collection agencies and to the respondent who must contribute the time required to supply the information to each agency.

Although existing archival records have not often been used as a basis for evaluating the impact of experimental social programs, they do have some promise in this regard. The argument that archival records can be used to mount more economical and more informative evaluations of social programs has been advanced persuasively by the Committee on Federal Program Evaluation of the National Academy of Sciences. We quote verbatim from that report:

> Once the major administrative archives of government, insurance companies, hospitals, etc., are organized and staffed for such research, the amount of interpretable outcome data on ameliorative programs can be increased tenfold. For example, Fisher [1972] reports on the use of income tax data in a followup on the effectiveness of manpower training programs. While these data are not perfect or complete for the evaluation of such a training program, they are highly relevant. Claims on unemployment compensation and welfare payments would also be relevant. Cost is an important advantage. Using a different approach, Heller [1972] reports retrieval costs of $1 per person for a study of several thousand trainees. Even if $10 were more realistic, these costs are to be compared with costs of $100 or more per interview in individual followup interviews with ex-trainees. Rate of retrieval is another potential advantage. Followup interviews in urban manpower training programs have failed to locate as many as 50 % of the population, and 30 % loss rates would be common. Differential loss rates for experimen-

[4] Conducting special social surveys to assess the quality of routinely issued governmental statistics is not a new idea. Neither is government's attempt to suppress the results of special surveys novel. See Boruch [1976] for a review of suppression efforts at the local, regional, and national level.

> tal and control groups are also common, with the control groups less motivated to continue. In the New Jersey Negative Income Tax Experiment, over three years, 25.3 % of the controls were lost, compared with a loss of only 6.5 % of those in the most renumerative experimental condition. While retrieval rates overall might be no higher for withholding tax records, the differential bias in cooperation would probably be avoided, and the absence of data could be interpreted, with caution, as the absence of such earnings. (Campbell et al. [1975]).

It takes little imagination to see how relying on existing archival data can reduce the expense of a program evaluation. It is quite another matter to employ such records creatively in difficult research settings. One of the more clever applications of archival data stems from an effort by Robertson and others [1972] to evaluate the impact of TV messages which encourage drivers to wear their seat belts:

> In some recent tests, four different types of TV messages were broadcast over four different TV cables, each cable serving a random set of households within a large region. The research objective was to determine which TV or broadcast fostered the highest rate of seat belt usage. To evaluate usage, the researchers first observed whether or not drivers in the region wore seat belts as they stopped for lights at randomly selected intersections. To link actual usage with area of residence, i.e., with TV message type, some mechanism for identifying each driver's residence was necessary. Rather than question each driver, the researchers merely recorded auto license numbers and employed State Motor Vehicle archives to identify the driver's area of residence. Once each driver's residence and seat belt use were linked, it was an easy matter to compare the crude effects of alternative TV messages on use.

Some examples of the savings engendered by temporary and limited linkage of governmental records have been documented by Hansen and Hargis [1966]. In these cases, a sample of records maintained independently by the U.S. Census Bureau, by the Internal Revenue Service, and by the Social Security Administration were linked to determine how costs of surveys might be reduced.

> Prior to 1954, for example, the Economic Census of manufacturing, retail, and other industries was conducted by field interview survey with some larger firms canvassed by mail. In the interest of reducing costs markedly, mail survey was considered as an alternative to expensive field interview surveys. At that time, the Census had no mechanism for construction and maintenance of up-to-date mailing lists, however. Such mailing lists were maintained by Internal Revenue Service and Social Security Files, based on payroll tax records, and with some modification, the basic lists were checked for validity, then adopted by the Census Bureau as a basis for the mail survey in the economic census. To obtain data on the retail industry, conventional Internal Revenue Service forms were modified slightly, making it possible to eliminate any additional mail or interview surveys of this industry by the Census Bureau. More than $6 million were saved by employing this last strategy.
>
> Similar savings were said to have been realized in the 1967 Economic Census where, for example, modifications to Internal Revenue Service schedules permitted use of these forms to elicit necessary information, and small direct interview samples were adjoined to this effort to obtain necessary data on products, merchandise lines,

and so forth. Finally, "administrative records from the Social Security Administration and from the Census have been used to construct mailing and sampling lists economically for Bureau data collection programs and to avoid duplicating the collection of information."

3.3 Clarifying and Enriching Statistical Data for Policy Analysis and Applied Social Research

By clarification here we mean obtaining a better understanding of the meaning, nature, and limitations of a particular social statistic. "Employment rate", for example, is a deceptively simple label for a characteristic which is complex in origin. Clarification often implies an additional objective, that of enriching the data resource with respect to number and kind of data archived, for the sake of higher quality analysis. Improving the interpretability and analyzability of a data set can be accomplished in a variety of ways. Linking of multiple data sources for statistical purposes is one method of doing so. Note, however, that linkage of all individual records may not be essential; linking a (random) sample of records is often sufficient for this purpose.

To be concrete, consider that in the United States, the Internal Revenue Service, the Social Security Administration, and the Census Bureau each independently collect data on annual income from citizens. The separation of effort is related to differences in the various agency functions. Two of the Social Security Administration's primary missions, for example, are understanding income redistribution at present and estimating the impact of redistribution policy in the future. Most U.S. citizens are required to pay a social security tax based in part on gross income, but Federal employees often do not choose to enroll in the national Social Security plan and so their incomes are not on file in SSA record systems. The Internal Revenue Service directs its attention at a different but overlapping universe, the tax-paying public, it has a different function, taxation, and it defines income differently, notably in terms of "taxable income". The U.S. Census Bureau's definition of income differs from each of the other agencies' definitions because its function is unique - statistical description of the state of the population - and because there are severe limitations on the way in which census data can be collected.

The result of these differences in definition of income, universe, and in function is that the relationships among these various sources of data on "income" have not been well understood. The economist using one source of data to predict the impact of a new health insurance policy might well develop projections which differ notably from projections made by an economist using another source of very similar information. The discrepancy among sources is marked in particular cases, and it is reasonable to use record linkage to bring some order out of this confusion.

To accommodate the problem, a massive Federal effort to reconcile conceptual differences among record contents has been mounted jointly by the U.S. Census Bureau, the Social Security Administration, and the Internal Revenue Service. The relevant data base includes the Bureau's 1973 Current Population Survey and administrative records from IRS and SSA files. The reconciliation has three imme-

diate purposes: to understand the relationships among ostensibly identical categories of information maintained by each agency, to input resultant data into the SSA simulation models of the tax transfer system, and to assess relative biases in Census statistics. The reconciliation involves linking a strategied sample of records on individuals from the various sources, not linkage of the entire data bases. Preliminary results of the study reported by Herriot and Spiers [1975] suggest that census statistics on income are quite reliable for salaried employees and regular wage earners; the overlap between Census reports contents is about 96 %. Income reports of the self-employed show somewhat less accuracy (90 % agreement between Census and IRS); reports of interest and dividends made to census are considerably less reliable (less than 80 % agreement) for most respondent groups.

As a result of such research, the models of economic systems employed by the U.S. Census Bureau and by the Social Security Administration (SSA) can be improved considerably when error rates based on IRS data can be recognized. The differential predictability of male and female incomes becomes more interpretable with evidence on differential accuracy in reporting such income to Census interviewers. The estimates of the impact of training on income become more reliable when corrected for base rate errors in reporting that income. And so on.

Similar benefits accrue from investigations of the differences in count data as a function of archival source.

A study by Cobleigh and Alvey [1975], for example, shows that differences in legally defined coverage of the population by Census and by SSA produce a Census comparable to a universe which is about 94 % of the SSA taxable earner's listings. Given a comparable universe, reports of average annual earnings from the two sources are in remarkable agreement except for very low and very high income groups. In the very low categories, SSA data show about 20 % more wage earners than does the Census data; in the high income categories, however, the Census counts are 10-20 % higher than Social Security reports. These latter differences are attributed by the authors to definitional differences and reporting irregularities including self-employment earnings not reportable to SSA, rounding error in self-reports to Census, late reporting to SSA, and to other factors.

Another type of enrichment involves the use of archival records for specialized research in which the record, though not disclosable by law or social custom to the social scientist, represents a key element in accomplishing applied research goals. Surrogates for the record may be sought, of course, but in the absence of any suitable substitute, it is often possible to capitalize effectively on restricted access records without according special privileges to the social scientist. For example, one of the peculiar and persistent tensions in our society involves the zealous efforts of the U.S. Internal Revenue Service to extract legitimate taxes from citizens and some citizens' equally strenuous efforts to avoid paying them. In an effort to clarify the conditions under which taxpayers will fulfill their responsibility with somewhat less resistance (or at least dissatisfaction), Schwartz and Orleans [1967] mounted some experimental tests of those conditions to compare relative rates of tax payments for a particular category of income.

Taxpayers were assigned randomly to one of three advertising strategies, the strategies differing in respect to their emphasis in justifying payment of taxes. The first condition relied heavily on appeals to moral conscience, the second on threats of punitive legal action, and the third on threats of social embarrassment (tax evasion being a matter for public legal action). The objective of the experiments was to determine which types of appeal led to higher rates of the particular income. To do so credibly required that condition or form of appeal be linked with the individual's subsequent reports of income to the Internal Revenue Service. In order to link the two kinds of records (the researcher's record of condition and the IRS record of income) without breaching IRS rules on disclosure of records (which are confidential by law) and the researchers' rules concerning disclosure of their own records, a mutually insulated file approach, described in Section 5, was used. (The results of the experiment are interesting. Middle-income respondents react most to the threats of legal action; low-income groups respond most to appeals to moral conscience; the high-income groups were most affected by threats of social embarrassment).

The case for merging separate data sets into a permanent consolidated pool of data is based on the assumption that the pooled data will be a more informative basis for social research than separate files. Examples of these are few, however, because the difficulty of match-merging files the differences in terminology, and differences in sample design and data collection procedures have inhibited many researchers from consolidating files. Moreover, it is difficult to anticipate the usefulness of linked files without actually trying the idea out on a small sample of records. Among the large-scale examples, the Wisconsin Assets and Income Studies Archive (Bauman, David, & Miller [1970]) illustrates what can be accomplished, however. Researchers appraise the effects of tax averaging proposals, changing incomes from retirements, capital gains income, and so on by simulating changes in tax laws, using the linked records as the raw material for analysis. Records from the Internal Revenue Service, Wisconsin tax records, the Social Security Administration, are combined in the file, without jeopardizing privacy of individuals on whom records are kept, to permit this research. The products of the research are predictions about the importance of changes in tax laws on individual income, strategy which attenuates the need to rely solely on anecdotal case study, intuition, and fragmented data as a basis for legislation in the tax area.

The more elaborate and more sensitive merged systems are found in the medical arena. Most involve both administrative and research information and, because they are recent systems, the benefits of pooling both kinds of data are not yet clear. Nonetheless, good reviews of the early products of such work are available for social medicine, community health services systems, and the like (e.g. Acheson [1967]). Laska and Bank's [1975] description of the Rockland Institute's psychiatric information system is probably one of the best of its kind. There is a strong emphasis on legislative and technical safeguards for assuring the confidentiality of the records. There is a hard-nosed product orientation: aside from common demographic information, the system facilitates quality control over treatment, time series analyses, and projective studies of the incidence and development of mental illness, and permits some uncontrolled studies of the effectiveness of treatment. Perhaps most importantly, the system can be coupled

neatly to experimental tests of alternative treatments to better understand whether and how well the treatments work (Endicott & Spitzer[1975].

4. Privacy Implications: Private with Respect to Whom?

Any longitudinal research involves linking observations made on an individual (or some other unit of analysis) at one point in time with observations made at a second point. The average statistical relation derived from the constellation of individual observations is, as we've said, useful for description at least, and is often essential for planning and evaluating social programs, for understanding change in human behavior, and for building theory and simulation models. The linkage is usually but not always made on the basis of clear identification of the respondent. Insofar as the identified respondent does share information about himself, the sharing process may be regarded, in principle, as a depreciation of the individual's privacy. That depreciation may be quite innocuous in the sense that information disclosed is innocuous; or it may be controversial, as in longitudinal studies of mental health.

Similarly, correlational data analysis must often be based on linkage of records from different archives. And if that linkage is based on clear identification contained in each record, then privacy may be depreciated in principle here as well. The custodian of an administrative archive may, by permitting linkage, violate law at worst or social customs at best by disclosing records to a researcher for linkage, however worthwhile the purpose of linkage. There may be a similar breach of a promise of confidentiality for a researcher who discloses his own records on identifiable individuals to an administrative archive, for example, in order to verify his records against those maintained by the archive.

These implications are almost useless in developing general strategies for assuring individual privacy. For although disclosure of information may represent a depreciation of privacy in principle, the fact of the matter is that neither government, nor social or adminstrative science, nor the respondent could get on well without some exchange of information about individuals. Admitting this, the focus must change from absolute assurance of confidentiality to balancing social information against the privacy-related needs of the individual. One approach to achieving that balance in a concrete way is to try to minimize depreciation of privacy without notably abridging our ability to collect meaningful data on human behavior. Doing so requires that we first identify the sources of risk in social research, then build mechanisms - procedural, statistical, and legal - to attenuate that risk.

We recognize, for example, that privacy may be reduced directly with respect to the social scientist. In the past, any such depreciation has been innocuous partly because social research itself has been fairly innocuous. But as applications of social research to social problems increase, as social scientists investigate more important or more controversial topics, the attention given to their inquiries are likely to increase. The import attached to relatively minor depreciation of privacy will increase. And so it becomes the social scientist's responsibility to develop mechanisms for minimizing the depreciation of privacy with respect to the researcher. I believe this in spite of the fact that no substantial risks to the respondent are usually engendered by survey research. The lack of risk is traceable to the researcher's

lack of interest in making personal judgments about particular individuals and his interest in statistical analysis of the relevant data. Identifiers serve merely as an accounting device, rather than as a vehicle for administrative action against (or for that matter for) an individual. Nonetheless, if identifiers could somehow be eliminated in the research process, or if the tie between identifier and response could be made useless for making personal judgments about individuals, without damaging research objectives needlessly, then we would do so. Parital solutions to the problem of doing so (Section 5) have been developed partly as a matter of principle, and partly because risks of disclosure may be generated by persons or agencies other than the researcher.

It is clear, too, that faudulent researcher, i.e., individuals posing as social scientists, can and occasionally do deceive citizens. They are motivated by financial gain (e.g., salesmen posing as pollsters), by pathological influences (e.g., rapists posing as survey interviewers, or at times, as policemen), or by other factors. In the interest of preserving the integrity of the profession and public trust in the social scientist, the social scientist must take some responsibility for protecting respondents against these infrequent but important dangers.

Social research records on identifiable individuals are often irrelevant for making administrative judgments about those individuals. We deal in samples rather than populations, and idiosyncratic ones at that. We deal with information which is usually not at the correct level of relevance or detail for administrative use. This partial relevance of research records on individuals usually serves as an inhibition against the appropriation of records for nonresearch purposes. Nonetheless, appropriation can and does occur. It may emerge under legal mandate as it has in the United States where, in a few instances, research records have been subpoenaed for use in judicial investigation of particular survey respondents. Exploitation may occur under legal traditions which are quite arbitrary and at times border on the capricious, as in some Congressional investigating committee activity. Or, the exploitation may be quite illegal, as in the theft and use of research records for personal profit or for the purpose of embarrassing the respondent. The consequences to the respondent can be serious: social embarrassment, legal sanction, personal discomfort. The consequences for research are no less serious: its inhibition and abrogation, now and in the future.

These risks are in principle real, if in practice remote. And so they deserve attention too. In particular, it is reasonable to examine mechanisms which protect the respondent from capricious action by law enforcement agencies, from criminal action based on the information he provides to a researcher, and from other attempts to appropriate research records for nonresearch purposes. This is especially true for those cases in which the benefits of the research are likely to offset greatly the social benefits of legal appropriation of records.

5. Competing and Conjoint Approaches to Assuring Confidentiality of Response in Social Research[5]

The general implication of the preceding section is that we take as an objective reducing depreciation of privacy without severe abridgment of research goals. Accommodating this joint task is difficult but there have been a variety of efforts mounted recently to do so. The major strategic approaches can be grouped into three broad categories - procedural, statistical, and law-related - which we consider next. This examination is brief; details are given in Boruch [1976].

5.1 Procedural Approaches

For longitudinal data collected periodically within the same framework, the simple device of using alias identifiers is obvious if underutilized. The alias may be created by the respondent and used consistently in response to permit intrasystem linkage. It may be created by social scientists, provided to the respondent, then purged from the social scientists' files to achieve the same ends. To decentralize the process, some neutral brokerage agency (a census bureau, a nongovernmental agency) may similarly create an alias for the respondent and destroy its own records of any linkage between clear identification and alias.

The strategy has been field tested with some success in U.S. drug studies, political attitude surveys and the like. Aside from logistical problems, its major shortcomings are the limitations imposed on linking the data elicited under alias with any other existing data on individuals.

To accommodate some logistical problems as well as the limitation on intersystem linkage, procedures such as the link file system have been developed. In this technique, a dictionary of double aliases is created by the social scientist and given over for safekeeping to an independent agency. The decentralization of the process enhances physical security, and if the agency is legally entitled to resist governmental appropriation of files, the procedure is legally secure. The dictionary is used as a basis for linking information which is periodically obtained from individuals. The main benefit of the strategy is that it reduces the social scientist's need to maintain longitudinal records on identified individuals, in general, and it reduces the time during which the social scientist has access to any given wave of data containing identifiers to an arbitrarily short period (see Astin & Boruch [1970]).

For those cases in which records from different archives must be linked, a variety of methods have been developed to permit linkage without violating the customs or law governing linkage. Among the better known systems for doing so is the "mutually insulated file approach", used in the Schwartz-Orleans [1967] study cited earlier. Basically, the system involves two files of records operated under different auspices; all records are identified and there is some overlap between the samples of individuals on which the records are maintained. To accomplish the linkage, the first archive (assume it is the social scientist) crypto-

---

[5] For a detailed examination of the benefits, shortcomings, vulnerability, and legal implications of some of these strategies, see Boruch [1974], and Campbell, Boruch, Schwartz, and Steinberg [1975].

graphically encodes the information portion of each record, producing a new file without meaning to any outsider, which is then transmitted to the record archive. The archive then matches the encoded records with its own records, based on the clear identifiers appearing in each record. Upon completion of the match, identifiers are deleted and the linked records are returned to the social scientist who then decodes relevant portions and the linked records and conducts his statistical analysis of the anonymous records. (See also Boruch [1972], and Campbell et al. [1975]).

These procedural approaches are simple, and in some cases, vulnerable to corruption. Nonetheless, they are useful in some, but not all research settings, to assure confidentiality of data with respect to the researcher and outsiders, and they can be tailored to accommodate longitudinal or correlational studies. Their refinement has been undertaken by both research community and the Federal bureaucracy to enhance the procedures' flexibility and protection level (Boruch [1976]). Some of the refinements depend on statistical approaches considered below.

5.2 Statistical Approaches

The devices just described are most often relevant to more impersonal forms of observation - questionnaires and the like - rather than to direct interview research. And in some instances, the logistical difficulties attached to their use are considerable. Partly for these reasons, it may be more appropriate to capitalize on one of the statistical strategies which have been developed to reduce depreciation in privacy. A variety of these approaches exists and these may be used alone or in conjunction with the procedural devices.

The best known class of approaches is the randomized response tactic currently under test and development by Greenberg in the United States, Dalenius [1975], Lanke, Swensson, Svensson, and Eriksson in Sweden, Warner in Canada, Moors in Holland, and others. In the simplest variation of the approach, the social scientist simultaneously presents a sensitive inquiry to an individual, e.g., "Did you cheat on your income taxes this year?" and an insensitive one, e.g., "Do you prefer potatoes over noodles?" The individual is then instructed to roll a die and to respond to the first question if a one or two shows up, and to the second question if a three, four, five, or six shows. He is also told to refrain from giving the interviewer any indication of which question was answered. When the process is carried out on two large samples of individuals and the instructions are followed by the respondent, it is possible to estimate the proportion of individuals in the sample who have cheated on their income tax forms and the proportion who prefer noodles. In particular, given some simple laws of probability, the odds on answering one or the other question, the odds on answering one or the other question, and the observed proportion of Yes responses, the estimation is a matter of simple algebra.

The technique permits us to establish the statistical character of sensitive properties of groups of individuals. And moreover, it does so without disclosing to the social scientist any information about a particular individual. It has been field tested in drug studies, in fertility control studies and other areas, and those tests continue in the U.S., Canada, Sweden, and elsewhere. The basic method is being refined to make it more efficient in a statistical sense, more acceptable

to the respondent in a social psychological sense, and less vulnerable to corruption in a legal sense.

A separate class of approaches is based on aggregation of response. The individual is asked not to respond individually to each of a set of questions but to respond in aggregated form to the set. In particular variations, for example, the respondent may add up numerical values corresponding to each answer of each question in a set. If "Yes" is assigned a value of 1 and "No" a value of -1, for example, the answer provided to a set of 10 questions each answerable with a Yes or No is a single number whose permissible range is -10 to +10. If numerical assignment is varied from one sample to the next, one needs only a little algebra - notably methods for solving a system of simultaneous equations - to estimate the proportions of individuals in the total sample who have each of the 10 properties.

Again, the technique permits one to elicit even sensitive information in direct interview situations without any deterministic linkage between an identified response to the researcher's question and the actual status of the individual. With some technical improvements, it probably can be applied to some longitudinal studies in which average relations among properties are essential.

The third and final class of statistical techniques which has received some attention is aggregation of the sample. The technique requires that one obtain data not on single identified individuals but rather on very small and carefully constructed clusters of individuals. If the cluster's composition remains the same over time, each cluster can, under certain conditions, be regarded as a synthetic person, a composite of all the properties of the small set of individuals it comprises. Some informative data analyses can be conducted on those aggregates and, insofar as aggregation helps to assure anonymity of individual response, there is no depreciation of individual privacy.

The applications of sample microaggregation have so far been limited to economic research on commercial units. Banks, for example, may be reluctant to release information about their operations to any outside economist. They are willing, however, to have the social scientist analyze aggregates of banks in the interest of reconciling bank privacy with future research. And indeed, a major system of data maintenance and dissemination has been built up on this theme by the University of Wisconsin (see Bauman, David, & Miller [1970]).

5.3 Approaches Based on Law and Government Practice

The final class of approaches to facilitating the privacy of the respondent in social research concerns formal legal action by legislators, the courts, or governmental executive agencies. Such action is taken to assure that when identifiable data must be collected for research purposes, the data will not be used for purposes other than research. As a practical matter, this means not only strengthening legal sanctions against criminal appropriation of research records, but also defining bounds on governmental appropriation of records. The actions are taken to reduce the likelihood that research records on identifiable individuals will be used to depreciate privacy any more than is normally required by research and to isolate that research against temporary threats, legal or otherwise, when the potential benefits of research justifies this course of action. The forms which such protection may

take vary considerably, and so we describe only a few stereotypes here.

In some of the United States, public officials such as the governor are empowered by the state constitution or by legislative act to offer testimonial privilege to a social researcher. That privilege entitles the recipient to legally resist any legal effort to appropriate his records on identifiable individuals. The threat of appropriation may stem from a prosecutor's idea that he may use even an unwilling researcher as a criminal investigator. It may stem from arbitrary exercise of subpoena power by legislatures or the courts. In order to legally assure that data will not be so appropriated, and consequently to increase the likelihood that individuals will cooperate in the research, a governor may then provide testimonial privilege on an ad hoc basis. To take a specific example, the governor of Vermont gave such privilege to researchers and respondents who participated in roadside surveys of drivers. The survey objectives were to estimate the proportion of drinking drivers (bloodtests were given to drivers) and the privilege was essential in getting high cooperation rate. Drivers who were legally intoxicated were driven home by a policeman. No record of any identified individual's condition was ledged with any law enforcement agency or other government archive, though drivers would normally be prosecuted under the law.

This sort of privilege can be applied in special cases where potential benefits of the survey are high and the relevant government executive is well enough informed to recognize the fact. However, we cannot always rely on expected benefits of research, for although some research may be important, it may also be risky with respect to its payoff. Nor can we always rely on the good offices of the public official, for the awarding of such privilege is discretionary and political factors may argue against it. In any event, discretionary privilege may be as susceptible to abuse from the naive researcher, just as it has been abused occasionally by some government executives.

Judicial discretion is another potential source of support for social scientists who, having collected identifiable data and having established a need for its maintenance, wish to secure it against non-research uses. In some cases, it has been possible for the scientist to legally resist a court-issued subpoena on grounds that the disclosure of identified records to the court would badly disable a major research effort. Evidence that breaches of confidentiality can be harmful to research efforts is readily available and can be used effectively to show cause why the records should not be used except in anonymous form. In fact, a similar line of argument has been used in a case involving the Negative Income Tax Experiments in New Jersey: The suspicion of fraud among people who happened to participate in the research led to a grand jury investigation and subpoena of research records on identified individuals.

Judicial discretion, like executive discretion, is by definition a bit arbitrary at best, and wildly unpredictable at worst. So its usefulness in protecting the confidentiality of data is not especially promising.

Legislative action in the form of concrete law is both feasible and, from the point of view of uniformity, very desirable. In particular, it is possible to build law to grant testimonial privilege to le-

gitimate social scientists under well defined conditions and uniformly applied criteria. It is also possible to build into such law sanctions against the fraudulent researcher or the corrupt social scientist or the public official who might attempt to appropriate research data for research purposes.

The 1970 Drug Abuse Act and the 1970 Alcohol Abuse Acts, for example, each carry a statute which permits the Attorney General to accord privilege to social scientists who are funded by the government to conduct research on those topics. Under the Public Health Act, persons engaged in research on mental health, including the use of alcohol and other proactive drugs, can be accorded privilege by the Secretary of Health, Education, and Welfare to protect the privacy of individuals who are subjects of such research.

These are new laws, enacted specifically to assure the confidentiality of social research records on identifiable individuals. They represent a delimitation of power on governmental access to social research records, and a delimitation of the conditions under which the researcher may act. They represent a spirit of support for the social sciences as well as an appreciation for the negative impact which even legal appropriation of research records may exert on policy-relevant research. At least one such law has been tested by the courts, and it's intent has been reaffirmed in that arena as well.

References

Acheson, E.D. Medical record linkage. London: Oxford University Press, 1967.

Astin, A.W., & Boruch, R.F. A "link system for assuring confidentiality of research data in longitudinal studies. American Educational Research Journal, 1970, 7(4), 615-624.

Barton, E.M., Plemons, J.K., Willis, S.L., & Baltes, P.B. Recent findings on adult and gerontological intelligence: Changing a stereotype of decline. American Behavioral Scientist, 1975, 19(2), 224-236.

Bauman, R.A., David, M.H., & Miller, R.F. Working with complex data files: The Wisconsin assets and income studies archive. In R.L. Biscoe (Ed.), Data bases, computers, and the social sciences. New York: Wiley-Interscience, 1970.

Bejar, I.I. Substudy 1a: Secondary analysis of the Cali tests of the nutrition and cultural enrichment program (Evaluation Research Memo #1IIB). Evanston, Illinois: Nortwestern University, Department of Psychology, February 1976.

Birren, J.E. The psychology of aging. Englewood Cliffs, New Jersey: Prentice-Hall, 1962.

Boruch, R.F. Costs, benefits, and legal implications of methods for assuring confidentiality in social research (Evaluation Research Report NIE-020). Evanston, Illinois: Psychology Department, Northwestern University, 1974. Reprinted as: Statische und methodische Proceduren zur Sicherung der Vertraulichkeit bei Forschung. In A. Eser and K.F. Schumann (Eds.), Forschung im Konflikt mit Recht und Ethik. Stuttgart: Ferdinand Enke Verlag 1976.

Boruch, R.F. Strategies for eliciting and merging confidential social research data. Policy Sciences, 1972, 3(3), 275-297.

Boruch, R.F., & Creager, J.A. Measurement error in social and educational research (ACE Research Reports, Vol. 7, No. 2). Washington, D.C.: American Council on Education (One DuPont Circle), 1972.

Bureau of the Census. Some preliminary results from the 1973 CPS-IRS-SSA exact match study: Invited papers on the reconciliation of survey and administrative income distribution statistics through data linkage. Reproduced report. Washington, D.C.: U.S. Department of Commerce, Bureau of the Census, September 30, 1975.

Burstein, L. The unit of analysis in educational research. Presented at the Annual Meetings of the American Educational Research Association, 1975, Washington, D.C.

Campbell, D.T. Administrative experimentation, institutional records, and noncreative measures. In J.C. Stanley (Ed.), Improving experimental design and statistical analysis. Chicago: Rand McNally, 1967, 257-291.

Campbell, D.T. Assessing the impact of planned social change. In G.M. Lyons (Ed.), Social research and public policies: The Dartmouth-OECD Conference. Hanover, New Hampshire: University of New Hampshire Press, 1975, 3-45.

Campbell, D.T., & Boruch, R.F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In A. Lumsdaine and C.A. Bennett (Eds.), Evaluation and measurement: Some critical issues in assessing social programs. New York: Academic Press, 1975, 195-296.

Campbell, D.T., Boruch, R.F., Schwartz, R.D., & Steinberg, J. Confidentiality preserving modes of access to files and to interfile exchange for useful statistical analysis. In A. Rivlin (Ed.), Report of the National Academy of Sciences, the Committee on Federal Agency Evaluation Policy: Protecting individual privacy in evaluation research. Washington, D.C.: NAS, 1975.

Campbell, D.T., & Erlebacher. A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), Compensatory education: A national debate. New York: Brunner/Mazel, 1970.

Cobleigh, C., & Alvey, W. Validating reported social security numbers. In Proceedings of the American Statistical Association: Social Statistics Section. Washington, D.C.: ASA, 1975, 145-151.

Coronary Drug Project Research Group. The Coronary Drug Project: Design, Methods, and Baseline Results. American Heart Association Monograph, 1973, 38.

Dalenius, T. The invasion of privacy problem and statistics production. Forskningsprojektet, Del I Undersökningar, Rapport Nr 61. Stockholm, Sweden: Statistiska Institutionen, Stockholms Universitet, December 1973.

Director, S. Evaluating the impact of manpower training programs. Ph.D. dissertation, Northwestern University (Evanston, Illinois), 1974.

Duncan, O.D. The 1970 Census: National uses--challenge and opportunity. In Proceedings of the American Statistical Association: Social Statistics Section, 1969. Washington, D.C.: ASA, 1966, 1-6.

Endicott, J., & Spitzer, R.L. Patient assessment and monitoring. In E.M. Laska and R. Bank (Eds.), Safeguarding psychiatric privacy. New York: John Wiley, 1975, 285-298.

Fägerlind, I. Formal education and adult earnings: A longitudinal study on the economic benefits of education. Stockholm: Almqvist & Wiksell, 1975.

Fisher, J.L. The uses of Internal Revenue Service data. In M.E. Borus (Ed.), Evaluating the impact of manpower programs. Lexington, Mass.: D.C. Heath, 1972, 177-180.

Gilmore, C.P. The real villain in heart disease. New York Times Magazine, March 25, 1973.

Goodman, L. Some alternatives to ecological correlation. American Journal of Sociology, 1959, 64, 610-625.

Goslin, D.A., & Bordier, N. Record keeping in elementary and secondary schools. In S. Wheeler (Ed.), On record: Files and dossiers in everyday life. New York: Russel Sage, 1969, 29-61.

Hansen, M.H., & Hargis, B.J. Census Bureau uses of tax data. In Proceedings of the Business and Economic Statistics Section: American Statistical Association. Washington, D.C.: ASA, 1966, 160-164.

Heber, R., Garber, H., Harrington, S., Hoffman, C., & Falender, C. Rehabilitation of families at risk for mental retardation. Madison, Wisconsin: University of Wisconsin, Rehabilitation Research and Training Center, 1972.

Heller, R.N. The uses of social security administration data. In M.E. Borus (Ed.), Evaluating the impact of manpower programs. Lexington, Mass.: D.C. Heath, 1972, 197-201.

Herriot, R.A., & Spiers, E.F. Measuring the impact on income statistics of reporting differences between the Current Population Survey and administrative sources. Proceedings of the Social Statistics Section, American Statistical Association. Washington, D.C.: ASA, 1975.

Horvitz, D.G. Problems in designing interview surveys to measure population growth. Proceedings of the Social Statistics Section, American Statistical Association. Washington, D.C.: ASA, 1966. 245-249.

Jabine, T.B., & Rothwell, N.D. Split-panel tests of census and survey questionnaires. 1970 Proceedings of the Social Statistics Section, American Statistical Association. Washington, D.C.: 1970, 4-13.

James, K.E. Regression toward the mean in uncontrolled clinical studies Biometrics, 1973, 29, 121-130.

Janson, C.-G. Project Metropolitan: A longitudinal study of a Stockholm cohort (Research Report No. 1). Stockholm, Sweden: Department of Sociology, Stockholm University, 1975.

Kannel, W.B., Dawber, T.R., Kagan, A., Revotskie, N., & Stokes, J. Factors of risk in the development of coronary heart disease--Six-year-followup experience: The Framingham Study. Annals of Internal Medicine, 1961, 55(33).

King, L.S. (Ed). A history of medicine. Middlesex: Penguin, 1971.

Klevmarken, A. Statistical methods for the analysis of earnings data. Stockholm: Almqvist & Wiksell, 1972.

Laska, E.M., & Bank, R. (Eds.). Safeguarding psychiatric privacy. New York: John Wiley, 1975.

Magidson, J., Barnow, B.S., & Campbell, D.T. Correcting the under-adjustment bias in the original Headstart evaluation (Evaluation Research Report No. 2JM). Evanston, Illinois: Psychology Department, Northwestern University, 1976.

Magnusson, D., Dunér, A., & Zetterblom, G. Adjustment: A longitudinal study. Stockholm, Sweden: Almqvist & Wiksell, 1975.

Marks, E.S., & Waksberg, J. Evaluation of coverage in the 1960 Census of the Population through case by case checking. Proceedings of the American Statistical Association: Social Section. Washington, D.C.: ASA, 1966. 62-70.

Mednick, S.A., & McNeil, T.F. Current methodology in research on the etiology of schizophrenia. Psychological Bulletin, 1968, 70, 681-693.

Mednick, S.A., Schulsinger, F., & Garfinkle, R. Children at risk for schizophrenia: Predisposing factors and intervention. In M.L. Keitzman, S. Sutton, and J. Zubin (Eds.), Experimental approaches to psychopathology. New York: Academic Press, 1975. 451-464.

Miller, H.P., & Hornseth, R.A. Cross-sectional versus cohort estimates of lifetime income. Proceedings of the American Statistical Association: Social Statistics Section. Washington, D.C.: ASA, 1970, 339-341.

Milles, M.T., & Kilss, B. Exploration of differences between linked Social Security and Internal Revenue Service wage data for 1974. Proceedings of the Social Statistics Section, American Statistical Association. Washington, D.C.: American Statistical Association, 1975.

Nesselroade, J.R., & Baltes, P.D. Adolescent personality development and historical change: 1970-1972. Monographs of the Society for Research in Child Development, 1974, 39(1).

Parnes, H.S. The National Longitudinal Surveys: New vistas for labor market research. Journal of the American Economic Association, 1975, 65(2), 244-249.

Potter, R.G., Chow, L.P., Jain, A.K., & Lee, C.H. Social and demographic correlates of IUCD effectiveness: The Taichung IUCD medical follow-up study. Proceedings of the American Statistical Association: Social Statistics Section. Washington, D.C.: ASA, 1966, 272-277.

Ramcharan, S., Cutler, J.L., Feldman, R., Siegelaub, A.B., Campbell, B., Friedman, G.D., Dales, L.G., & Collen, M.F. Multiphasic checkup evaluation study: 2. Disability and chronic disease after seven years of multiphasic health checkups. Preventive Medicine, 1973, 2, 207-220.

Reeder, L.G., et al. Advances in health survey research: Proceedings of a national invitational conference. Rockville, Maryland: National Center for Health Statistics, 1975.

Robertson, L.S., Kelley, A.B., O'Neill, B., Wixom, C.W., Eiswirth, R.S., & Haddon, W. A controlled study of the effect of television messages on safety belt use. Washington, D.C.: Insurance Institute for Highway Safety (Watergate Six Hundred), 1972.

Robinson, W.S. Ecological correlations and behavior of individuals. American Sociological Review, 1950, 15, 351-357.

Schaie, K.W. A general model for the study of developmental problems. Psychological Bulletin, 1965, 64(2), 92-107.

Scheuren, F., & Alvey, W. Selected bibliography on matching. In F. Scheuren et al., Exact match research using the 1973 Current Population Survey. Studies from Interagency Data Linkages (Report No. 4). Washington, D.C.: Office of Research and Statistics, Social Security Administration, 1975.

Scheuren, F., Herriot, R., Vogel, L., Vaughan, D., Kilss, B., Tyler, B., Cobleigh, C., & Alvey, W. Exact match research using the March 1973 Current Population Survey--Initial states. Studies from Interagency Matching (Report No. 4). Washington, D.C.: Office of Research and Statistics, Social Security Administration, 1975.

Schwartz, R.D., & Orleans, S. On legal sanctions. University of Chicago Law Reviews, 1967, 34, 282-300.

Sinesterra, L., McKay, H., & McKay, A. Stimulation of intellectual and social competence in Colombian preschool children affected by multiple deprivations of depressed urban environments (Progress Report, mimeo). Cali, Colombia: University Center for Child Development, Human Ecology Research Station, Universidad del Valle, 1971.

Steinberg, J., & Pritzker, L. Some experiences with and reflections on data linkage in the United States. Bulletin of the International Statistical Institute, 1969, 42 (Book 2), 786-805.

Wall, W.D., & Williams, H.L. Longitudinal studies and the social sciences. London: Heinemman, 1970.

Wohlwill, J.F. Methodology and research strategy in the study of developmental change (ETS Research Memorandum RM-69-24). Princeton: Educational Testing Service, December 1969.

Wohlwill, J.F. The age variable in psychological research. Psychological Bulletin, 1970, 77(1), 49-64.

Young, A.F., Selove, J.M., & Koons, D.A. Measuring quality of housing. Proceedings of the American Statistical Association: Social Statistics Section, 1966. Washington, D.C.: ASA, 1966, 33-42.

4.6 INDIVIDUAL DISCLOSURES FROM FREQUENCY TABLES

by

Ove Frank

Abstract

By "individual disclosure from a frequency table" we mean the possibility of using the information in a table to deduce some information about an individual. Determination of the extent of disclosure can be extremely complicated if we take into account comparisons between several frequency tables which refer to some common variables and overlapping populations. An analysis of the disclosures is presented which is based on a specification of what (information about whom) is disclosed by whom. The disclosure relations that occur between the individuals in a population can be illustrated by using a directed graph in which the nodes are individuals and the arcs indicate disclosures. The structure of the disclosure graph is studied and various kinds of prior knowledge are discussed. In particular, a coalition is introduced which consists of individuals who know individual data concerning each other. Possibilities and risks of disclosure are given for such coalitions.

Key words

Disclosure, invasion of privacy, individual integrity, graph, coalition.

1. Introduction

The computerization of society and the creation of large data banks have made many people feel anxious about the potential risks of misuse of data and consequent invasion of privacy. In the production of official statistics it has become increasingly important to study various problems related to statistical confidentiality. Fellegi [4] discussed various aspects of the problem of protecting the individual against invasion of privacy. Dalenius [3] gives an overview of the problems and discusses various approaches.

In this paper we will consider the problem of checking or estimating the extent to which it is possible to use the information in a frequency table in order to disclose individual items of data. The information in a frequency table can alternatively be given as a list of anonymous individual items of data, i.e. a data record without identifying labels. It is difficult to determine which individual disclosures can be made since information from several different sources can be combined, e.g. from tables and records with some common variables and overlapping populations. With the increasing flow of data in society it therefore becomes more and more difficult to guarantee that official statistics do not disclose individual information in any way. It would be desirable to find methods of estimating the degree of protection (measured in some specified and suitable way) which can be maintained against disclosures.

Even if direct disclosure of sensitive information is never possible from a single frequency table, a particular individual may consider too great the risk of disclosure or misinterpretation of some information about him. The risk concept which is used here is subjective

and lacks a clear interpretation. Maybe the very fact that he cannot judge certainly whether or not disclosure is possible can in itself be considered as an invasion of privacy.

Frank [5, 6] and Cassel [2] have defined the risk of disclosure by introducing random guessing in order to reconstruct the individual data from the table frequencies. Another way of defining the risk of disclosure will be introduced in this paper. It is based on a stochastic model of the available prior or supplementary information which can be used together with the table in order to make disclosures.

Section 2 is an introductory discussion of the disclosure problem. Section 3 briefly discusses disclosure by guessing, and Section 4 disclosure by using prior information. Section 5 takes up the particular kind of prior information given by a coalition of individuals who know the data about each other. The smallest coalition needed to make a disclosure can be looked upon as a measure of the level of protection against disclosures. Section 6 introduces a model of uncertain prior information, and defines the risk of disclosure as the probability that the prior information available is sufficient to make a disclosure. Randomly composed coalitions are considered in Section 7. The risk of disclosure and the expected number of disclosures per person are determined according to various models of forming coalitions.

2. The disclosure problem

In multidimensional tables it is generally cells of small frequencies which are considered to imply high potential risks of disclosure. We will analyze the meaning of this assumption in some detail.

Among the variables used to define the cells in a multidimensional table we will distinguish between those variables which are generally accessible and those which are not. Such variables for which the individual values are publicly accessible in official registers in Sweden are, for instance, age, sex, place of birth, income, occupation, etc. An arbitrary multidimensional frequency table can be represented by a two-dimensional table with r rows corresponding to the combinations of values of the generally accessible variables and s columns corresponding to the combinations of values of the other variables. In this two-dimensional table it is known (or can be found out) which individuals belong to each row. If it is possible to locate the column of an individual, this is a disclosure. If the column classification is based on sensitive data, a disclosure might be an invasion of privacy.

We will consider a finite population U of N individuals labeled by the integers $1,\ldots,N$. The various combinations of values of the generally accessible variables will be labeled by the integers $1,\ldots,r$, which are considered as the values of a row index x. The various combinations of values of the other variables will be labeled by the integers $1,\ldots,s$, which are considered as the values of a column index y.

Individual u belongs to row x_u and column y_u. The number of individuals u with $x_u = i$ and $y_u = j$ is denoted N_{ij}. The table with r rows, s columns and cell frequencies N_{ij} is denoted by $\underline{N}$. The set of those individuals which belong to cell (i,j) is denoted by

$$U_{ij} = \{u \in U : (x_u, y_u) = (i,j)\},$$

and the individuals who belong to row i and column j are denoted by $U_{i.}$ and $U_{.j}$ respectively. The row populations $U_{i.}$ are known, but the cell populations U_{ij} are generally unknown.

The number of rows r is equal to the product of the numbers of levels of the generally accessible variables. It follows that with an increasing number of such "background" variables the number of rows grows rapidly. In practice it may occur that many row frequencies are equal to unity even for a relatively small number of background variables. In this case the background variables can be used to identify the individuals and disclose their columns. Block and Olsson [1] have investigated a record of individual data items which the Swedish Central Bureau of Statistics has been required to unlabel, since some of the information is sensitive. They found that most of the individuals can be identified by use of the background variables, and consequently that unlabeling is not sufficient to protect the individuals against disclosures.

Direct disclosure in the table $\underline{N}$ is possible if and only if there is a row in which only one cell has a positive frequency. In such a case anyone can disclose the columns for all the individuals in that row.

When a direct disclosure is impossible but there is a row with only two positive cell frequencies both equal to unity, then each one of the two persons in the row can disclose the column of the other, provided that they know their own columns.

When a direct disclosure is impossible but there is a row with only two positive cell frequencies only one of which is equal to unity, then that person can disclose the column of all the others in the row, provided that he knows his own column.

From these simple observations it follows that the possible disclosures ought to be specified by stating who can be disclosed by whom and what supplementary information is needed for this to be possible.

In a row where none of the above reasoning is applicable, for instance a row with only two occupied cells each one of which contains two persons who know their own columns, there is no possibility of disclosing with certainty information about anyone in the row without further information. This observation indicates the need of a concept of risk of disclosure.

After discussing briefly the risk of disclosure based on random guessing we will turn to a new approach based on stochastic models of prior information.

3. Simple random guessing

Assume that direct disclosure is impossible in the table $\underline{N}$. If a simple random procedure is used to guess the column of each person in the table, then the probability

$$\prod_{i=1}^{r} \left(\prod_{j=1}^{s} N_{ij}!/N_{i.}! \right)$$

is assigned to each possible distribution of the N persons in the cells, subject to the cell frequency restrictions. The risk of disclosure of

an arbitrary individual can be defined as the probability that the column of that particular individual is guessed correctly. It follows that the risk of disclosure of an arbitrary person in cell (i,j) is $N_{ij}/N_{i.}$. Consequently the risk of disclosure for the person in row i who is most easily disclosed is equal to $\max_j N_{ij}/N_{i.}$. Moreover, the expected risk of disclosure of a randomly chosen individual in row i is $\sum_j (N_{ij}/N_{i.})^2$.

Simultaneous disclosure of several persons can be considered, and for instance it is found that the probability that at least one person in cell (i,j) will be disclosed is equal to

$$1 - \binom{N_{i.} - N_{ij}}{N_{ij}} \Big/ \binom{N_{i.}}{N_{ij}},$$

the probability that all the persons in cell (i,j) will be disclosed is equal to

$$1 \Big/ \binom{N_{i.}}{N_{ij}},$$

and the probability that all the persons in row i will be disclosed is equal to

$$\prod_j N_{ij}! \; / \; N_{i.}!.$$

Frank [5, 6] and Cassel [2] have used guessing procedures and probability calculations of this kind, and have discussed various measures of protection against invasion of privacy in multidimensional tables.

It is, of course, easy to object to the interpretation that a correct guess is a disclosure. A disclosure in the strict sense does not exist when the person who is guessing does not know whether the guess is correct or not. It is partly for this reason that it may be interesting to consider the risk of disclosure in an alternative way, one which is not based on guessing but on the supplementary information required to make a disclosure in the table $\underline{N}$.

4. Disclosure by using prior information

Let $\underline{\alpha}$ be an N×N matrix where the element α_{uv} equals 1 or 0 according to whether person u has sufficient prior information to be able to find the y-value for person v without recourse to the table $\underline{N}$.

For a fixed prior information matrix $\underline{\alpha}$ a disclosure matrix $\underline{\delta}$ is defined by giving the element δ_{uv} the value 1 or 0 according to whether person u can disclose the column of person v by combining his prior information with the table $\underline{N}$. Prior information in itself will not be considered as a disclosure, i.e. $\delta_{uv} = 0$ if $\alpha_{uv} = 1$.

The posterior information, i.e. the prior information extended by the disclosures, is given by a matrix $\underline{\beta} = \underline{\alpha} + \underline{\delta}$.

The prior information, the disclosures and the posterior information can be represented by directed graphs. The nodes are the individuals and an arc from node u to node v means that person u has prior knowledge, can disclose, or has posterior knowledge about the y-value for person v. The matrices $\underline{\alpha}$, $\underline{\delta}$ and $\underline{\beta}$ are the adjacency matrices of the prior information graph, the disclosure graph and the posterior information graph.

The matrix $\underline{\delta}$ gives a complete description of the disclosures based on $\underline{\alpha}$ and $\underline{N}$. Some summary statistics which may be of interest are

$$\frac{1}{N}\sum_{u=1}^{N}\max_{v}\delta_{uv} \quad \text{and} \quad \frac{1}{N}\sum_{v=1}^{N}\max_{u}\delta_{uv}$$

which are the proportions of disclosing and disclosed persons respectively (the proportions of persons in the population who can disclose or can be disclosed by at least one person),

$$\frac{1}{N}\sum_{u=1}^{N}\sum_{v=1}^{N}\delta_{uv} = \frac{1}{N}\sum_{u=1}^{N}\delta_{u.} = \frac{1}{N}\sum_{v=1}^{N}\delta_{.v}$$

which is the mean number of disclosures (disclosed or disclosing persons) per person,

$$\max_{u}\max_{v}\delta_{uv}$$

which is an indicator of at least one disclosure, and

$$d(m,n) = \sum_{u=1}^{N} I(\delta_{u.} = m,\ \delta_{.u} = n)$$

which is the number of persons who can disclose m and can be disclosed by n persons.

Let us first consider the general case of an arbitrary prior information matrix $\underline{\alpha}$. When person u uses his prior information he can reduce the table $\underline{N}$ to a residual table $\underline{N}(u)$ with frequencies

$$N_{ij}(u) = N_{ij} - \sum_{v\in U_{ij}}\alpha_{uv}.$$

The frequency $N_{ij}(u)$ is the number of persons in cell (i,j) for whom person u does not know the y-value. It is possible to write

$$\underline{N}(u) = \underline{N} - \underline{X}\,\underline{\alpha}(u)\,\underline{Y}'$$

where the r×N matrix $\underline{X}$ and the s×N matrix $\underline{Y}$ have the elements

$$X_{iu} = \begin{cases} 1 & \text{if } x_u = i \\ 0 & \text{otherwise} \end{cases} \qquad Y_{ju} = \begin{cases} 1 & \text{if } y_u = j \\ 0 & \text{otherwise} \end{cases}$$

and the N×N matrix $\underline{\alpha}(u)$ is a diagonal matrix with elements $\alpha_{u1},\ldots,\alpha_{uN}$.

Now, person u can obviously disclose those and only those persons who belong to the rows with only one occupied cell in the residual table $\underline{N}(u)$. If no row in $\underline{N}(u)$ has only one occupied cell, then person u cannot disclose anyone.

In the simple case of no prior information, i.e. all $\alpha_{uv} = 0$, the trivial result is that

$$\delta_{uv} = \begin{cases} 1 & \text{if } u \in U \text{ and } v \in U' \\ 0 & \text{otherwise,} \end{cases}$$

where U' is the set of persons belonging to some cell (i,j) with $N_{ij}=N_{i.}$ for i=1,...,r and j=1,...,s. The disclosure graph thus consists of a complete subgraph with node set U', and all the other nodes have arcs to each node in U'.

Another simple case is when each person only has prior knowledge of his own y-value, i.e. α_{uv} is 1 or 0 according to whether u = v or u ≠ v. Let us denote by U''_{ij} the set which is equal to U_{ij} or the empty set according to whether there exists a k ≠ j such that $N_{ij} + N_{ik} = N_{i.}$ and $\min(N_{ij},N_{ik}) = 1$. Let N''_{ij} be the number of persons in U''_{ij}. Then it follows that

$$\delta_{uv} = \begin{cases} 1 & \text{if } u \neq v,\ u \in U \text{ and } v \in U' \\ 1 & \text{if } u \neq v,\ u \in U''_{ij} \text{ and } v \in U_{i.} \text{ for some } (i,j) \text{ where } N''_{ij} = 1 \\ 0 & \text{otherwise.} \end{cases}$$

This result implies a more explicit summary description, which is given in the following theorem. We need the following terminology and notations. Rows with only one occupied cell are called primary rows. Rows with only two occupied cells, at least one of which has unit frequency, are called secondary rows. The numbers of persons in the primary and secondary rows are N' and N" respectively. The number of cells of unit frequency in the secondary rows is r". r"(n) of these cells belong to the rows in which the occupied cells have the frequency 1 and n for n=1,2....

Theorem 4.1. If each person has prior knowledge only of his own y-value, then the number of persons who can disclose m and be disclosed by n persons is given by

$$d(m,n) = \begin{cases} N' & \text{if } m = N'-1 \text{ and } n = N-1 \\ N-N'-N'' & \text{if } m = N' \text{ and } n = 0 \\ N''-r'' & \text{if } m = N' \text{ and } n = 1 \\ r''(1) & \text{if } m = N'+1 \text{ and } n = 1 \\ r''(n) & \text{if } m = N'+n \text{ and } n = 2,3,... \\ 0 & \text{otherwise.} \end{cases}$$

In particular, N' persons can be disclosed by everyone, N"-r"+r"(1) persons can be disclosed by only one person, and N-N'-N"+r"-r"(1) persons cannot be disclosed by anyone.

Proof. If U' and U'' are the sets of persons in the primary and secondary rows respectively, it follows by a systematic examination of $\underline{\delta}$ that

$$(\delta_{u.}, \delta_{.u}) = \begin{cases} (N'-1, N-1) & \text{if } u \in U' \\ (N', 0) & \text{if } u \in U - U' - U'' \\ (N', 1) & \text{if } u \in U''_{ij} \text{ and } N''_{ij} > 1 \\ (N'+1, 1) & \text{if } u \in U''_{ij}, \ N''_{ij} = 1 \text{ and } N_{i.} = 2 \\ (N'+N_{i.}-1, 0) & \text{if } u \in U''_{ij}, \ N''_{ij} = 1 \text{ and } N_{i.} > 2. \end{cases}$$

The result then easily follows.

In this case the disclosure graph can be described in the following way. There is a complete subgraph with N' nodes, each one of which also has arcs from every other node. There are r''(1)/2 subgraphs with two nodes and mutual arcs. There are r''(n) bipartite subgraphs with a single node having arcs to n nodes for n = 2,3,... .

5. Coalitions

A subset $C \subset U$ is called a coalition if each person in C knows the y-values of all the members of C. A coalition $C \subset U_{i.}$ is called a coalition within row i. A coalition with m members is called an m-coalition. We will consider prior information given by a coalition C, i.e.

$$\alpha_{uv} = \begin{cases} 1 & \text{if } u \in C \text{ and } v \in C \\ 0 & \text{otherwise.} \end{cases}$$

We will also consider prior information given by coalitions $C_1, \ldots, C_r$ within the rows, i.e.

$$\alpha_{uv} = \begin{cases} 1 & \text{if } u \in C_i \text{ and } v \in C_i \text{ for some } i=1,\ldots,r \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 5.1. The smallest coalition which can disclose at least one person has $m = \min_i (N_{i.} - \max_j N_{ij})$ members.

Proof. If there is a row with only one occupied cell, then

$$\min_i (N_{i.} - \max_j N_{ij}) = 0$$

and direct disclosure is possible without any coalition. Otherwise there is a row i with $N_{i.} > \max_j N_{ij} > 0$, and the $\max_j N_{ij}$ persons who belong to a cell of maximal frequency can all be disclosed by the other $N_{i.} - \max_j N_{ij}$ persons in the row, provided that these persons constitute a coalition. No smaller coalition within the row can disclose anyone. Hence it follows that the smallest of these row coalitions is the smallest coalition which can disclose at least one person.

Theorem 5.2. The smallest coalition which can disclose all persons outside the coalition has $m = \Sigma(N_{i.} - \max_j N_{ij})$ members.

Proof. A coalition which can disclose all its non-members must contain at least $N_{i.} - \max_j N_{ij}$ members in row i for each i=1,...,r. Hence the result follows.

Theorem 5.3. Let m be a fixed number less than $N_{i.}$. Each m-coalition within row i can disclose $N_{i.} - m$ persons if $m \geq N_{i.} - \min_j^+ N_{ij}$. No m-coalition within row i can make a disclosure if $m < N_{i.} - \max_j N_{ij}$. There is an m-coalition within row i which can disclose $N_{i.} - m$ persons if $m \geq N_{i.} - \max_j N_{ij}$. Here $\min_j^+ N_{ij}$ denotes the smallest positive cell frequency in row i.

The proof is similar to those given above and is therefore omitted.

6. Risk of disclosure in stochastic prior information models

If the prior information is built up from many different sources and is difficult to manage or get access to, it may seem appropriate to use a stochastic model as a first approximation. If such a view is accepted and the prior information is given by a stochastic matrix $\underline{\alpha}$, then the disclosure matrix $\underline{\delta}$ also becomes stochastic. It may be of interest to study the risk

$$P = E \max_u \max_v \delta_{uv}$$

of at least one disclosure, and the expected mean number of disclosures per person, i.e.

$$\mu = \frac{1}{N} \sum\sum_{uv} E\delta_{uv}.$$

Consider the residual table $\underline{N}(u)$ and define the event

$$A_{ij}(u) = (N_{ij}(u)>0)$$

that cell (i,j) contains some person whose y-value is not previously known to person u. Let

$$B_{ij}(u) = (N_{i.}(u) = N_{ij}(u)>0)$$

be the event that cell (i,j) is the only cell in row i where there is some person whose y-value is not previously known to person u. Now

$$B_{ij}(u) = A_{ij}(u) \cap_{k\neq j} \overline{A_{ik}(u)},$$

and for fixed u and i the events $B_{ij}(u)$ are mutually exclusive for different j. We have that

$$P = P\Big(\bigcup_u \bigcup_i \bigcup_j B_{ij}(u)\Big)$$

Moreover

$$E\delta_{uv} = P(\alpha_{uv} = 0 \text{ and } B_{ij}(u)) = P(\alpha_{uv} = 0 \text{ and } \bigcap_{k\neq j} \overline{A_{ik}(u)})$$

for $v\in U_{ij}$. These formulae can be used to find the disclosure risk P and the expected disclosure mean μ. We will illustrate this by a simple Bernoulli model for $\underline{\alpha}$.

Theorem 6.1. Let α_{uv} be independent Bernoulli variables with $E\alpha_{uv}=p_{ij}(u)$ for $v\in U_{ij}$. Assume that no direct disclosure can be made in $\underline{N}$. Then the risk of at least one disclosure is

$$P = 1 - \prod_u \prod_i \Big[1 - \sum_j \Big(1-p_{ij}(u)^{N_{ij}}\Big) \sum_{k\neq j} p_{ik}(u)^{N_{ik}}\Big],$$

and the expected number of disclosures per person is

$$\mu = \frac{1}{N} \sum_u \sum_i \sum_j N_{ij}[1-p_{ij}(u)] \prod_{k\neq j} p_{ik}(u)^{N_{ik}}.$$

In particular $E\alpha_{uv} = p$ for all u and v implies that

$$P = 1 - \prod_i \Big[1-p^{N_{i\cdot}} \sum_j \Big(p^{-N_{ij}} - 1\Big)\Big]^N$$

and

$$\mu = (1-p) \sum_i \sum_j N_{ij} p^{N_{i\cdot} - N_{ij}}.$$

Proof. From the Bernoulli assumption it follows that $N_{ij}(u)$ is binomially distributed with the parameters N_{ij} and $1-p_{ij}(u)$. Moreover, $N_{ij}(u)$ are independent for different u and different (i,j). Consequently

$$E\delta_{uv} = [1-p_{ij}(u)] \prod_{k\neq j} p_{ik}(u)^{N_{ik}}$$

for $v\in U_{ij}$, and the formula for μ easily follows. The formula for P follows from the expansion

$$P = P\Big(\bigcup_u \bigcup_i \bigcup_j B_{ij}(u)\Big) = 1 - \prod_u \prod_i [1 - \sum_j P(B_{ij}(u))]$$

and

$$P(B_{ij}(u)) = [1-p_{ij}(u)^{N_{ij}}] \prod_{k\neq j} p_{ik}(u)^{N_{ik}}.$$

7. Risk of disclosure with random coalitions

We will now let the prior information be given by one or more coalitions formed according to various stochastic models. In all the cases we will assume that no direct disclosure is possible in $\underline{N}$. The disclosure risk P and the expected mean disclosure μ will be given.

Theorem 7.1. Assume that an m-coalition is formed by simple random sampling without replacement from U. The risk of at least one disclosure is

$$P = 1 - \sum_{m_1} \ldots \sum_{m_r} \prod_i \left[\binom{N_{i.}}{m_i} - \sum_j \binom{N_{ij}}{N_{i.} - m_i} \right] \Big/ \binom{N}{m}$$

where the sum is over all $m_1,\ldots,m_r$ satisfying $\sum_i m_i = m$. The expected mean number of disclosures per person is

$$\mu = \frac{m}{N\binom{N}{m}} \sum_i \sum_j N_{ij} \binom{N-1-N_{i.}+N_{ij}}{N-1-m}.$$

Proof. Let C be the coalition and let m_{ij} be the number of coalition members in U_{ij}. Put $n_{ij} = N_{ij} - m_{ij}$. We have

$$N_{ij}(u) = \begin{cases} n_{ij} & \text{if } u \in C \\ N_{ij} & \text{otherwise.} \end{cases}$$

As in the previous section we define the events

$$A_{ij} = (n_{ij} > 0) \text{ and } B_{ij} = (n_{i.} = n_{ij} > 0)$$

and obtain

$$P = P\Big(\bigcup_i \bigcup_j B_{ij}\Big) = 1 - \sum_{m_1} \ldots \sum_{m_r} \frac{\prod_i \binom{N_{i.}}{m_i}}{\binom{N}{m}} P\Big(\bigcap_i \overline{\bigcup_j B_{ij}} \,\Big|\, m_1,\ldots,m_r\Big)$$

where the probability is conditioned by the numbers of coalition members in the rows. Now

$$P\Big(\bigcap_i \overline{\bigcup_j B_{ij}} \,\Big|\, m_1,\ldots,m_r\Big) = \prod_i P\Big(\overline{\bigcup_j B_{ij}} \,\Big|\, m_i\Big) = \prod_i \Big[1 - \sum_j P\big(B_{ij} \,\big|\, m_i\big)\Big]$$

where

$$P\big(B_{ij} \,\big|\, m_i\big) = \binom{N_{ij}}{N_{i.} - m_i} \Big/ \binom{N_{i.}}{m_i},$$

and the formula for P follows. Furthermore for $v \in U_{ij}$ and $u \neq v$

$$E\,\delta_{uv} = P(u \in C,\ v \notin C \text{ and } B_{ij}) = \begin{cases} \binom{N-1-N_{i.}+N_{ij}}{N-1-m} \Big/ \binom{N}{m} & \text{if } u \in U_{i.} - U_{ij} \\ \binom{N-2-N_{i.}+N_{ij}}{N-1-m} \Big/ \binom{N}{m} & \text{otherwise.} \end{cases}$$

It follows that

$$E\,\delta_{.v} = \left[(N_{i.}-N_{ij}) \binom{N-1-N_{i.}+N_{ij}}{N-1-m} + (N-1-N_{i.}+N_{ij}) \binom{N-2-N_{i.}+N_{ij}}{N-1-m} \right] \Big/ \binom{N}{m} =$$

$$= m \binom{N-1-N_{i.}+N_{ij}}{N-1-m} \Big/ \binom{N}{m}$$

for $v \in U_{ij}$, which yields the formula for μ.

Theorem 7.2. Assume that an m_i-coalition is formed by simple random sampling without replacement from $U_{i.}$ for each $i=1,\ldots,r$. The risk of at least one disclosure is

$$P = 1 - \prod_i \left[1 - \sum_j \binom{N_{ij}}{N_{i.}-m_i} \Big/ \binom{N_{i.}}{m_i} \right]$$

and the expected mean number of disclosures per person is

$$\mu = \frac{1}{N} \sum_i \frac{m_i}{\binom{N_{i.}}{m_i}} \sum_j N_{ij} \binom{N_{ij}-1}{N_{i.}-1-m_i}.$$

Proof. With the previous notation we have

$$P = 1 - \prod_i [1 - \sum_j P(B_{ij})]$$

where

$$P(B_{ij}) = \binom{N_{ij}}{N_{i.}-m_i} \Big/ \binom{N_{i.}}{m_i}.$$

Let C_i be the coalition within row i. We have for $v \in U_{ij}$ and $u \in U_{i.}$ that

$$E\,\delta_{uv} = P(u\in C_i,\ v\notin C_i \text{ and } B_{ij}) = \begin{cases} \binom{N_{ij}-1}{N_{i.}-1-m_i}\Big/\binom{N_{i.}}{m_i} & \text{if } u\in U_{i.} - U_{ij} \\ \binom{N_{ij}-2}{N_{i.}-1-m_i}\Big/\binom{N_{i.}}{m_i} & \text{otherwise.} \end{cases}$$

It follows that

$$E\,\delta_{.v} = \left[(N_{i.}-N_{ij})\binom{N_{ij}-1}{N_{i.}-1-m_i} + (N_{ij}-1)\binom{N_{ij}-2}{N_{i.}-1-m_i}\right]\Big/\binom{N_{i.}}{m_i} =$$

$$= m_i\binom{N_{ij}-1}{N_{i.}-1-m_i}\Big/\binom{N_{i.}}{m_i}$$

for $v\in U_{ij}$, which yields the formula for μ.

Theorem 7.3. Assume that a coalition is formed by applying independent Bernoulli sampling with selection probability p_{ij} in cell (i,j) for $i=1,\ldots,r$ and $j=1,\ldots,s$. The risk of at least one disclosure is

$$P = 1 - \prod_i\left[1 - \sum_j \left(1-p_{ij}\right)^{N_{ij}} \prod_{k\neq j} p_{ik}^{N_{ik}}\right]$$

and the expected mean number of disclosures per person is

$$\mu = \frac{1}{N}\sum_i\sum_j N_{ij}[(N_{ij}-1)p_{ij} + N_{i.} - N_{ij} + Em - \sum_k N_{ik}p_{ik}](1-p_{ij}) \prod_{k\neq j} p_{ik}^{N_{ik}}$$

where $Em = \sum_i\sum_j N_{ij}p_{ij}$ is the expected coalition size. In particular, $p_{ij} = p$ implies that

$$P = 1 - \prod_i[1-p^{N_{i.}}\sum_j(p^{-N_{ij}} - 1)]$$

and

$$\mu = \frac{(1-p)}{N}\sum_i\sum_j N_{ij}[N_{i.}-N_{ij}+p(N-1-N_{i.}+N_{ij})]\,p^{N_{i.}-N_{ij}}.$$

Note. 0^0 is to be interpreted as 1. In Theorem 6.1 the α_{uv} were independent Bernoulli variables, but here the $\alpha_{uv} = \alpha_{uu}\alpha_{vv}$ are not independent.

Proof. We have

$$P(B_{ij}) = P(A_{ij}) \prod_{k \neq j} P(\overline{A_{ik}})$$

where

$$P(A_{ij}) = 1 - p_{ij}^{N_{ij}},$$

and the formula for P follows in a manner similar to that in the proof of Theorem 7.2. Furthermore we have for $v \in U_{ij}$ and $u \neq v$ that

$$E\,\delta_{uv} = \begin{cases} \lambda_{ij} & \text{if } u \in U_{i.} - U_{ij} \\ p_{ij}\lambda_{ij} & \text{if } u \in U_{ij} \\ p_{kl}\lambda_{ij} & \text{if } u \in U_{kl} \text{ and } k \neq i \end{cases}$$

where

$$\lambda_{ij} = (1-p_{ij}) \prod_{k \neq j} p_{ik}^{N_{ik}} .$$

Consequently

$$E\,\delta_{.v} = \lambda_{ij}[N_{i.} - N_{ij} + (N_{ij}-1)p_{ij} + \sum_{k \neq i} \sum_{l} N_{kl}p_{kl}]$$

for $v \in U_{ij}$, and the formula for μ follows.

References

(1) Block, H. and Olsson, L., 1975, Bakvägsidentifiering.

(2) Cassel, C.M., 1975, On probability based disclosures in frequency tables.

(3) Dalenius, T., 1974, The invasion of privacy problem and statistics production - an overview. Statistisk Tidskrift, pp.213-225.

(4) Fellegi, I.P., 1972, On the question of statistical confidentiality. Journal of the American Statistical Association 67, pp. 7-18.

(5) Frank, O., 1973, Reconstruction of individual data from classification frequency distributions. Research Report 73-13, Department of Statistics, University of Uppsala.

(6) Frank, O., 1974, Individens integritetsskydd i flerdimensionella frekvenstabeller. In a report from a conference in Bergendal. (Ed. Swedish Central Bureau of Statistics, Stockholm.)

4.7 PROBABILITY BASED DISCLOSURES

by

Claes-Magnus Cassel

1. Introduction

What is a disclosure? The statement "Per Svensson's driver's licence was withdrawn" might be a disclosure. It is a disclosure if two conditions are fulfilled:

i) The statement is true.

ii) The individual Per Svensson would not approve if the statement were published.

In a disclosure a value of a sensitive variable is typically associated with an identified individual. In the example above the pair (Per Svensson, licence withdrawn) could thus be a disclosure. (There are other ways to define disclosures.) A disclosure with probability P occurs if the statement (Per Svensson, licence withdrawn) is true with probability P. What does "true with probability P" mean?

To get a probability P we need a probability model. The model should reflect the circumstances of the situation. If we were interested in finding out whether or not our neighbor Per Svensson (P.S.) had his licence withdrawn (L.W.) we would need a model for the quantification of our belief in the statements

i) (P.S.; L.W.)

ii) (P.S.; L. not W.)

If we have no information we could use model I below.

Model I

$P(\text{P.S.; L.W.}) = 1/2$

$P(\text{P.S.; L. not W.}) = 1/2$

This can be motivated by observing that for the individual P.S. there are two possible values of the sensitive variable, (L.W.) and (L. not W.). In the absence of any other information one could select one value at random, this gives model I. However, one may intuitively feel that this model is not satisfying. In general, people's licences are not withdrawn very often. Frequently one would have information of the kind used to set up model II belows.

Model II

Suppose the Table T is given

L.W.	L. not W.
n_1	$n-n_1$

This information could be used to refine model I. We know now that there are n_1 individuals out of n whose licences were withdrawn. Thus the set of n values contains n_1 (L.W.) values. If we were to make a guess under the given restrictions T, we could simply choose a value at random from the set of n values. This would give

$P(L.W.) = n_1/n$

$P(L.\ not\ W.) = (n-n_1)/n$

That is, we would have a chance of n_1/n of getting a (L.W.)-value. We are then able to form the table

		State of nature	
		(P.S.; L.W.)	(P.S.; L. not W.)
Statement	(P.S.; L.W.)	(true)	false
	(P.S.; L.not W.)	false	(true)

which indicates that:

- the statement (P.S.; L.W.) is true with $P = n_1/n$ if the state of nature is (P.S.; L.W.)
- the statement (P.S.; L. not W.) is true with $P = (n-n_1)/n$ if the state of nature is (P.S.; L. not W.)

Thus there could be a disclosure with probability n_1/n that P.S.'s licence actually had been withdrawn.

2. Notation

Consider a set of n identified individuals $\{u_1,\ldots,u_n\}$. The value of the sensitive variable x for individual u_i is $x(u_i)$. The set of individuals then gives rise to the set $\{x(u_i),\ldots,x(u_n)\}$ of values. (Note that "identification" and "sensitiveness" are key words which are used in different senses in different papers. What is to be meant by identification and sensitiveness does not seem to be quite clear. The Swedish Central Bureau of Statistics (SCB) has done some research on this.)

For the purpose of this paper we define

Definition 1: A disclosure for u_i occurs if $(u_i, x(u_i))$ can be formed.

Set U = an individual selected at random, and X = a value selected at random. Then

Definition 2. A disclosure with probability P occurs if ($U = u_i$, $X = x(u_i)$) is true with probability P according to some probability model.

The probability model we will use is model II described in the Introduction. There are two types of disclosure associated with definition 2:

i) if the individual is given and the value of the sensitive variable is wanted. For example, let the individual be P.S. and let the problem be to find out if his licence was withdrawn.

ii) if the value is given and the individuals having this value are wanted. For example, one could want to know whose licences were withdrawn.

We will only discuss case i). (There is research going on in SCB concerning case ii).)

3. The statistician's problem

Like other official statistical authorities SCB produces mainly

i) frequency tables, a typical one is shown below

T

x_1		x_j		x_k
n_1		n_j		n_k

$$n = \sum_{j=1}^{k} n_j,$$

where x is a classification variable, here assumed to be sensitive, and n_j are frequencies.

ii) quantity tables and measure tables are shown below

T

x_1		x_j		x_k
m_1		m_j		m_k

where m_j is a statistical measure based on the units contributing to cell no. j.

The problem is to decide whether the table T can be published or not. Connected to that decision is the calculation of the maximum risk of disclosure if T were published. If T is a frequency table one can use model II for that calculation. Thus, if "guessing under restrictions" is a realistic model, than P is

n_1/n for individuals in cell no. 1
n_2/n " " " " " 2
.
.
n_j/n " " " " " j
.
.
n_k/n " " " " " k

The maximum chance of guessing correctly occurs for individuals belonging to the class for which n_j is as large as possible. Set

$$P_0 = \max_j (n_j/n).$$

If the producer considers every individual equally important, he must base his decisions on P_0. Then T is characterized by P_0.

If T is a quantity table then m_j is the quantity accumulated by the individuals in cell no. j. (We assume that the possibility of disclosure is to be investigated with regard to x and not to the quantities. This can be discussed.) The probability model needs frequencies and since T provides us only with quantities we will have to estimate frequencies. This can be done in several ways depending on what other information is available. We will here use a basic assumption: that all individuals contribute equally ($\bar{m} = \sum_{j=1}^{k} m_j/n$) to the quantities. This gives an estimate of the frequencies as

$$\hat{n}_1,\ldots,\hat{n}_j,\ldots,\hat{n}_k,$$

where $\hat{n}_j = m_j/\bar{m}$, and "guessing under restriction" can be used again. Of course the quality of the estimates depends strongly on what other information is available. If T is a measure table the same principles can be applied. One would, however, have to use more assumptions and other information to get estimates of the frequencies - how much, depends on what measure is tabulated. The basic problem: "can T be published?" still remains. The producer has to balance two types of demands:

i) from consumers who demand the publication of tables

ii) from individuals (or other institutions) who demand protection against disclosure.

These demands cause the producer's dilemma: How should the demand for publication be balanced against the risk of disclosure? How large can P_0 be allowed to be and T still be published? Intuitively one has a feeling that it should depend upon the sensitiveness of x, which in turn depends on the possible losses an individual could suffer if x were disclosed. This suggests an decision-theoretic approach to the problem.

4. A hypothetical decision-theoretic example

The producer's possible actions are

1. publish T
2. do not publish T.

The possible states of nature in T are

1. disclosure occurs
2. disclosure does not occur.

Let us assume that the following loss tables are given:

for individuals

	disclosure	no disclosure
publish	a	0
do not publish	0	0

for consumers

	disclosure	no disclosure
publish	0	0
do not publish	b	b

(The accuracy of the loss tables can of course be discussed. However, in some aspects they seem to be realistic.) The maximum probability of disclosure is P_0. We thus find the expected loss to be

for individuals

$$\begin{cases} aP_0 & \text{if T is published} \\ 0 & \text{if T is not published} \end{cases}$$

for consumers

$$\begin{cases} 0 & \text{if T is published} \\ b & \text{if T is not published} \end{cases}$$

If we accept the rule of choosing the action which has the smallest total expected loss (for individuals and consumers), we must calculate the total expected losses, which are:

total expected loss

$$\begin{cases} aP_0 & \text{if T is published} \\ b & \text{if T is not published} \end{cases}$$

and thus we see that

if $P_0 < b/a$ then T should be published
if $P_0 > b/a$ then T should not be published

The decision to publish or not is thus guided by considerations regarding both the consumer and the individual, as well as the risk of disclosure. Finally, we observe that in the example above the consumer's loss (b) and the individual's loss (a) are regarded as equally important. This is not necessary - it would be possible and desirable to use some kind of weighting.

5. LIST OF PARTICIPANTS

Aase, Asbjörn — professor
Geografisk institutt, Trondheims Universitet, Trondheim, Norge

Anér, Kerstin — riksdagsledamot
Riksdagen, Fack, 100 12 Stockholm 36

Back, Pär-Erik — professor
Statsvetenskapliga inst., Umeå Universitet, 901 87 Umeå

Barabba, Vincent P. Mr.
Director, Bureau of the Census, Washington D.C. 20233, USA

Boruch, Robert F. — professor
Department of Psychology, North Western University, Evanston, IL 60201, USA

Bruhn-Möller, Åke — kanslichef
Statens råd för samhällsforskning, Sveavägen 166, 113 46 Stockholm

Cassel, Claes-Magnus — fil.lic.
Statistiska centralbyrån, Fack, 102 50 Stockholm 27

Dahl, Sven — professor
Inst. för kulturgeografi, Göteborgs Universitet, Fack, 400 10 Göteborg 3

Dahlgren, Hans — fil.lic.
Pedagogiska inst., Lärarhögskolan, Fack, 431 20 Mölndal

Dalenius, Tore — professor
109 Benevolent St., Providence, R.I. 02 906, USA

Danielsson, Jens — avdelningsdirektör
Datainspektionen, Fack, 103 60 Stockholm

Eklund, Gunnar — docent
Statistiska inst., Stockholms Universitet, Box 6701, 113 85 Stockholm 23

Engberg, Ole — ingeniör
Dansk data arkiv, H.C. Andersens Boulv. 38, 1553 Köpenhamn K, Danmark

Eriksson, Sven — docent
Statistiska inst., Göteborgs Universitet, Viktoriagatan 13, 411 25 Göteborg

Faxén, Karl-Olof — docent
Svenska Arbetsgivareföreningen, Box 161 05, 103 23 Stockholm

Frank, Ove — professor
Statistiska inst., Lunds Universitet, Fack, 220 05 Lund 5

Fägerlind, Ingemar docent
Internationell pedagogik, Stockholms Universitet,
Fack, 104 05 Stockholm

Gastwirth, Joseph L. professor
Department of Statistics, George Washington University,
2201 G-street N.W., Washington, D.C. 20052, USA

Grip, Arne universitetslektor
Ekonomiska inst., Linköpings Universitet,
581 83 Linköping

Gullberg, Elsa Maria civilekonom
Inst. för informationsbehandling, Stockholms
Universitet, Box 6706, 113 85 Stockholm

Hammar, Tomas docent
Inst. för statsvetenskap, Stockholms Universitet,
Fack, 104 05 Stockholm 50

Helmfrid, Staffan professor
Kulturgeografiska inst., Stockholms Universitet,
Box 6801, 113 85 Stockholm

Himmelstrand, Ulf professor
Inst. för sociologi, Uppsala Universitet,
Sturegatan 2B, 752 23 Uppsala

Ingemarsson, Ingemar professor
Inst. för systemteknik, Linköpings Universitet,
581 83 Linköping

Janson, Carl-Gunnar professor
Tegnérlunden 6, 113 59 Stockholm

Johannesson, Ingvar professor
Pedagogiska inst., Lunds Universitet,
Fack, 220 07 Lund 7

Karlsson, Georg professor
Sociologiska inst., Umeå Universitet, 901 87 Umeå

Karlsson, Gösta fil. stud.
Statistiska inst., Lunds Universitet, Fack
220 05 Lund

Klevmarken, Anders professor
Statistiska inst., Göteborgs Universitet,
Viktoriagatan 13, 411 25 Göteborg

Källner, Claes-Göran generaldirektör
Datainspektionen, Fack, 103 60 Stockholm 3

Lambe, Bengt byråchef
JK:s kansli, Fack, 103 10 Stockholm

Langefors, Börje professor
Inst. för informationsbehandling, Stockholms
Universitet, Fack, 104 05 Stockholm 50

Lanke, Jan fil.dr
Avd. för matematisk statistik, Lunds Universitet,
Box 725, 220 07 Lund 7

Larsson, Inger avdelningsdirektör
Byrå L3, Skolöverstyrelsen, 106 42 Stockholm

Levin, Lennart byråchef
Universitetskanslersämbetet, Box 163 34,
103 26 Stockholm

Ljung, Bengt-Olov professor
Lärarhögskolan, Fack, 100 26 Stockholm 34

Magnusson, David professor
Psykologiska inst., Stockholms Universitet, Box 6706,
113 85 Stockholm

Malmquist, Sten professor
Statistiska inst., Stockholms Universitet, Box 6701,
113 85 Stockholm

Mattsson, Ingrid fil.lic
Internationell pedagogik, Stockholms Universitet,
Fack, 104 05 Stockholm

Nyström, Britt Marit fil.kand.
Rättssociologiska seminariet, Bredagatan 4,
222 21 Lund

Ohlsson, Ingvar generaldirektör
Statistiska centralbyrån, Fack, 102 50 Stockholm 27

Osborn, Jack L. Mr.
Grannert Build., Purdue University, West Lafayette,
Indiana 47907, USA

Peterson, Olof forskningsassistent
Inst. för statsvetenskap, Uppsala Universitet,
Box 514, 751 20 Uppsala

Pernelid, Åke direktör
Statskonsult AB, Box 4040, 171 04 Solna

Rapaport, Edmund avdelningschef
Statistiska centralbyrån, Fack, 102 50 Stockholm 27

Rydén, Nils avdelningsdirektör
Datainspektionen, Fack, 103 60 Stockholm

Samuelsson, Kjell universitetslektor
Inst. för informationsbehandling, Stockholms
Universitet, 104 05 Stockholm

Sjöström, Åke byråchef
Socialstyrelsen, Fack, 106 30 Stockholm

Sundström, Arne avdelningsdirektör
Lunds datacentral, Sölvegatan 18, 223 62 Lund

Svensson, Allan docent
Pedagogiska inst., Lärarhögskolan, Fack,
431 20 Mölndal

Swensson, Bengt fil.lic
Inst. för statistik, Universitetsfilialen i Örebro,
Box 164, 701 03 Örebro 1

Svensson, Nils-Eric docent
Riksbankens Jubileumsfond, Box 1649,
111 86 Stockholm

Thedéen, Torbjörn docent
Statistiska inst., Stockholms Universitet, Box 6701,
113 85 Stockholm

Thorngren, Bertil tf. professor
Ekonomiska inst., Umeå Universitet, 901 87 Umeå

Thorson, Tore Direktör
Stockholms stads statistiska kontor, Fack,
103 10 Stockholm

Wallberg, Klas avdelningschef
Statistiska centralbyrån, Fack, 102 50 Stockholm 27

Westerholm, Barbro docent
Apoteksbolaget AB, 105 14 Stockholm

Westlund, Anders forskningsassistent
Inst. för statistik, Umeå Universitet, 901 87 Umeå

Wijkman, Anders riksdagsman
Riksdagen, Fack, 100 12 Stockholm 36

Winberg, Christer fil. dr
Landsarkivet, Göteborg

Vinge, Per-Gunnar direktör
Sveriges Industriförbund, Box 5501, 114 85 Stockholm

Wohlin, Lars ekon.dr
Industriens Utredningsinstitut, Box 5037
102 41 Stockholm 5

Wärneryd, Karl-Erik professor
Handelshögskolan, Box 6501, 113 83 Stockholm

Zetterberg, Lars professor
Tekniska Högskolan, 100 44 Stockholm

Åkerman, Sune docent
Historiska inst., Uppsala Universitet, Box 514,
751 20 Uppsala

Öyen, Örjar professor
Sosiologisk Inst., Bergens Universitet, Bergen, Norge