Journal of Official Statistics
Vol. 9. No. 2, 1993, pp. 339–344

Discussion
Disclosure Limitation and Data Access

L.W. Cook[1]

1. Introduction

The juxtaposition of the Lambert and Reynolds papers provides a stimulating contrast. Reynolds argues that disclosure in research studies is a low risk activity, because no major reactions have been known to occur. Lambert suggests that disclosure is not only a risk of violation, but a risk of a perception of violation. Those risks vary according to the information known to the individual "intruder." Lambert provides a comprehensive analysis of the disclosure problem faced by any continuing survey organisation.

2. Perceived Versus Actual Disclosure

Both Lambert and Reynolds conclude that the increased capability to access individualised records focuses on the trust held in the researcher. That trust will vary with the other information held by the researcher, as that will affect the risks of disclosure or perceived disclosure. Lambert's paper is analytically based where Reynolds's is presumptive, and the two papers combined are needed to fully recognise the imperative for better access to individual records, and the constraints on how to do this.

Lambert notes on pages 19 and 20 that current awareness among the public of their existing risks of disclosure may be slight. This would suggest that without any change in release of information, heightened public interest in data release practices could still create a significant adverse reaction. This would be a risk from a changed perception of disclosure. Also, because awareness is low, a disclosure from one government agency can be associated with other agencies. It is important that uniform practices are used across all agencies, otherwise a "slip-up" in one agency can affect the confidence in all agencies. This will be especially true in the USA's decentralised system.

The concept of perceived disclosure provides a more effective means for survey organisations to assess the range of risks that data release can generate. It recognises that the influences of phenomena that are unrelated to the specific dataset of interest may be far more significant than the actual risk determined by the specific dataset. For example, in New Zealand the legislation enabling the Department of Social Welfare to access a wide range of datasets for benefit compliance purposes highlights the benefit policing role of the department. The interests of its staff in record linkage for research purposes might be distinguished from those in compliance activities by a few informed officials, but it would severely test the faith of the general public to recognise that distinction in personnel, management structure and

[1] Department of Statistics, PO Box 2922, Wellington, New Zealand.

function. As a consequence, that department is less likely to be provided access to official statistical data, based on trust, despite the high trustworthiness of its research staff, because of its very public compliance activities.

This non-involvement with policy administration is one of the reasons official statistical agencies, with a good reputation for confidentiality, get much higher response rates. It is also a reason why response rates deteriorate when there are "slip-ups" or "perceived slip-ups."

3. Reynolds's Advocacy Thrust

The Reynolds paper provides a strongly argued advocacy for increasing the access to personal records, and their linkage, by researchers and policy analysts. Reynolds's prime proposition is that "Further advances in understanding will be possible only if more diverse and complete information can be assembled."

He then notes that "Existing public datasets are one major, low cost source of such information." The thrust of the paper is focused on this point, with the object of comparing privacy risks with research benefits. Information is provided extensively in this paper, unfortunately to support this line of reasoning, rather than to develop an analytical framework for comparing risks and benefits. Consequently, the analysis of risks and benefits is not complete.

In support of Reynolds's prime proposition, official information sources are usually drawn from administrative records, or statistical surveys and censuses. There are also many research studies that are based on neither. In the case of administrative records, they have limitations as information sources for reasons of coverage and consistency. Their characteristics are inevitably accepted by statisticians, rather than developed from any research design.

4. Research Design of Statistical Surveys

Statistical surveys usually have a research design, and their quality is balanced against the cost. From a policy analyst perspective, the research design of official statistical surveys should ideally be drawn as a whole from a comprehensive framework of statistical needs, which recognises the capability of administrative records.

The absence of an effective research design leads to a second-best proposition, that data linkage after collection can provide additional information. The conditions for this to hold, or to be most effective, are not discussed in the paper. There is an extensive body of literature on imputation which gives a good understanding of the limits to the data structures that can be created by matched datasets designed for unrelated purposes, but which share variables in common, or most usefully, allow exact matching. There should be linked variables designated, to support exact and statistical matching at a micro level, or exact matching at the macro level. This matching should be aimed at increasing the extent of internal analysis possible.

5. Continuing Trust as Basis for Official Data Collection

Reynolds's analysis focuses on existing datasets, and in considering issues of trust, totally ignores the fact that most public databases are continuing collections, whether it be a quarterly tax return or a decennial population census. Information providers must have continuing trust in those who operate systems which produce individual records, if compliance costs

are not to be excessive, or reliability jeopardised. Indeed, for a national statistical office, the level of trust is a significant asset in its continuing operations. There is an assertion by Reynolds on page 308 that "However, as all the data will have already been collected, many of the criteria related to selection of participants, informed consent, and the like would not apply." This is quite untrue for any continuing collection.

6. Increased Conflict of Interest Between Analysis and Compliance

The capability of researchers to match data from different data sources, and carry out complex analytical studies has been significantly increased by advances in information technology. Those same advances have also increased the capacity of agencies to link information for compliance purposes. While Reynolds does not refer to this, in practice it is a major concern to national statistical offices, in that the same matching capability that increases research options heightens public concern with privacy.

In New Zealand, for example, there is a shift away from a welfare system with a mixture of universal and targeted benefits, to one with wholly targeted benefits. Concern with compliance with benefit eligibility requirements has led to increased exchanges of information between a wide range of public agencies, including taxation and welfare authorities. These exchanges have been highly publicised.

One consequence is that the national statistical office, responsible for over 80 percent of official statistics, has specifically distanced itself from information transfers of any sort. The most significant policy change has been to cease the sale of listings from the tax-based directory of businesses.

Thus, the same public policies which need better access to micro datasets for modelling policy design and analysis also have increased the need for exchange of information for compliance purposes. These contradictory uses of an individual's data have led to a proposal in New Zealand that the national statistical office should set up a protected environment, a data laboratory, where modelling studies can be carried out under secure conditions and within the limitations provided by legislation. Statistics Canada has moved this further, and itself has undertaken modelling studies in a wide range of areas including taxation, benefit eligibility, business dynamics, labour market, and life expectancy variability.

In providing the modelling capability, a national statistical office may itself become part of the continuum of policy design analyses, evaluation, and presentation. Thus, the national statistical office could be directly linked to particular public policies, despite having no advocacy position.

7. Standard of Confidentiality

The Reynolds paper asserts that uniform (federal) standards or procedures that are applicable only to research could be devised that take precedence over individual agency guidelines. This is inconsistent with the accountability of chief statisticians – who wears the risk? It notes (page 308) that information providing agencies would not be expected to consider protection of individual privacy or organisational confidentiality in making decisions regarding approval of the research. This underestimates the cost of research organisations complying with security/confidentiality practices that are in place in statistical offices, for example, implementation of

random rounding, topcoding and other masking techniques.

This presumes that public perception of the privacy risk facet from a conglomerate dataset is comparable to that faced by a series of unlinked datasets. Swedish experience in the public perception of the privacy risk faced from conglomerate datasets, partly noted on page 289, not only resulted in the termination of the project but also has had a lasting effect on the public survey response rates in Sweden.

There is a somewhat patronising inference in Section 6 of the paper that this experience and that in Germany and the Netherlands reflects a poor popular appreciation of the benefits of responsible social science research. In all three situations, gaining public confidence relates not only to some of the purposes of the data but to how it is to be obtained. This includes the public perception of the integrity and accountability of the survey takers and researchers.

In fact, in some countries – Australia and New Zealand, for example – response rates in recent years have been above those experienced a decade earlier, yet there has been no reason to indicate that the distrust in government there has not increased, as with the world over. In both these countries, the quinquennial Census of Population has been preceded by a major public relations campaign emphasising the benefits of census statistics, and their link to hospitals, schools, and other community investments. One side benefit is a "halo" effect on the response rates of other surveys, lasting up to two years. While this is not a scientific study, it supports the proposition that where people are effectively informed of the value of their information, and the trust they can have in those obtaining it, then compliance would be higher.

Reynolds measures public perception on page 277 as being determined by the absence of any apparent public controversy in the mass media or courts. This could be because there are already safeguards in place. I would argue that responsiveness to information gathering must be the relevant measure, given that the courts and the mass media must be the stage for but a little of any public questioning of the trust they can have in those they provide information to. Whether response rates are affected by the value placed on research based on the data provided depends on the knowledge available to respondents, and the risks to privacy perceived from other uses of the data. Thus the benefits of research will be placed in the context of the costs of disclosure to other parties. In this regard, as noted earlier, Reynolds's views seem rather crude in comparison with the analysis of Lambert.

That record linkage itself arouses public concern suggests that without increased accountability to the public, then compliance will reduce. Indeed Reynolds has not considered whether in fact the better design of public datasets by designing them at the start to address matters of public importance might not lead to more significant information gains than to record linkage, in many cases.

Reynolds's conclusion on the extent to which public privacy can be breached is summarised on page 280: "As professional snoops, social scientists must live with the notion that they are not always 'nice people' when they study important topics." This seems to leave unanswered questions of accountability. Contrary to Reynolds's thrust, official statisticians rely on public trust as one of their most vital assets. To breach this in some way by not respecting the privacy of respondents may initiate a reaction that could have a major effect on non-compliance. Consequently, the risk

that Reynolds can accept ex poste is not one that can be borne where continuing surveys are being operated. The importance of this is again underestimated by Reynolds when he considers measures taken to ensure privacy in the United States, and he has underrated the conventions and practices of its statistical bureaux, whether founded in law or not.

In fact, on page 306 Reynolds is quite scathing of the concern of statistical offices in maintaining voluntary compliance, on the grounds that the concerns only exist because it "would complicate their work" to do otherwise. That it has also led to the cessation of regular population censuses in Germany and the Netherlands has not gone unnoticed by Reynolds (page 289), but he places no value on data quality issues from poor response.

In noting: "The insular concerns of agencies focusing on their current responsibility" (page 306), on one hand Reynolds could have been articulating the long held concerns of statisticians with the second rate status their work is given by administrative agencies whose records also provide statistics. By his preceding comment, Reynolds has widened the "insularity" to include official statisticians, and perhaps anyone who disagrees with him.

There are in some countries data archives that are managed outside of the legislative protection provided to statistical offices. The Social Research Data Archive at the University of Essex is perhaps the most well known. Its very success indicates that public trust can be extended to such institutions, which can effectively act as agents in conferring access rights to researchers.

Unfortunately Reynolds's paper provides us with little basis to span these two positions, and to influence or satisfy public confidence.

8. Informed Consent

Reynolds continues his narrow assessment of public trust on page 285, when considering the issue of whether informed consent can be given for access to data. He asserts: "It is assumed that individuals have representatives to make judgements on their behalf. If elected officials agree that it is in the best interests of their constituents that research be conducted with administrative databases, they can be assumed to provide 'informed consent' on behalf of their constituents, who are also the source of the data."

This assertion assumes a trust in elected officials that is contradicted later on page 289. It also conflicts with the function of elected officials noted by Reynolds on page 305, to support . . . "the use of data for routine administrative, monitoring, and compliance objectives." The features and advantages of informed consent outlined by Reynolds on page 284 would not be met, without challenge, if elected officials were to provide the authority for unintended linkage of individual records.

Consent and confidentiality is complicated by the uncertainty in the confidentiality expected of ethnic, religious, regional and other community structures to which people belong. In New Zealand, the indigenous people are some 14 percent of the total population, and they are made up of some 80 tribes (iwi). The issue of privacy is now a matter of debate, in that tribal statistics could be seen as belonging to the tribe.

9. Conclusion

A major contemporary issue for national statistical offices is how to increase the capability of users to carry out secondary analysis on individual records, yet not to jeopardise the trust of the public

in institutions which hold that information in custody. The paper describes a selection of situations and experiences relating to data access and public confidence.

Reynolds has articulated a (Federal) Data Base Research Certificate. Such a certificate would be accompanied with the privilege of unlimited access to any data maintained in any agency (with their approval).

This proposition is a consequence of the arguments presented by Reynolds, and the lack of a comprehensive analytical framework in the paper makes its value and limits difficult to assess directly.

Some of the assertions in Reynolds's paper are contradictory, and in general the thrust is to access now what has already been collected rather than a wider view of improving the nature of data collection, and enhancing the public confidence in better, more relevant data collection.

The partial analysis of Reynolds does not provide a path for willing official statisticians to tread, in expanding the research access to large-scale datasets on individuals and organisations. It does not lead to datasets which more significantly relate to the policy questions faced by government and the public. This can be done by using existing knowledge about improving the integration of datasets, but most effectively by establishing a nationwide framework for relating data collected to public questions.

10. The Direction of the Future

The direction for the future should have the following characteristics:

a. Major initiatives to develop information frameworks which relate diverse public datasets to public and political decision-making on social and economic well-being, the social and economic structures relevant to this, their variability, and likely advancement factors in society.
b. Increase integration of datasets by common classifications, and enumeration units and sampling methodology.
c. Micro datasets, where there has been sufficient masking to ensure the risk of disclosure is low, supported by legislation which specifies the obligations of both the provider and the researcher.
d. Setting up data laboratory arrangements (possibly with remote access capabilities) to allow researchers to analyse datasets without having direct access to the data. (This access would be arranged by the statistical offices that would use their normal procedures for ensuring the confidentiality of output.)
e. Maintaining a log of all access to the datasets.
f. Full support of acknowledged guardians of the public interest, in obtaining information about data, matching and research access to datasets.