Journal of Privacy and Confidentiality (2011) **3**, Number 2, 91–100

Rejoinder

Gerald W. Gates*

I would like to thank Steve Fienberg who encouraged me to submit my paper for publication in the *Journal of Privacy and Confidentiality*. I am also happy that Steve was able to get such a distinguished group of experts to provide comments on the issues I raise. I have known each of the commenters for many years—some for much of my career. I have learned a great deal from each about the challenges we face in finding the appropriate legal, policy, and technical approaches to foster privacy and promote researcher access to data.

As Prewitt and Fienberg assert, the statistical system is at a turning point in assessing alternative data sources to meet the nation's statistical data needs. Whether these sources come from government or private sector records, issues of privacy and confidentiality have, and will continue to, play a prominent role. For their part, many of the commenters have delved more deeply into how the problem is framed by analyzing the relationship between concepts of privacy, confidentiality, and informed consent and how they apply in government statistics. They also argue for expanding the focus beyond administrative data to private-sector-generated digital data. Finally, commenters have offered additional options that are both technology-driven and people-driven.

Each of these comments is important to the privacy debate and some go beyond my paper's focus on privacy in the context of the statistical use of administrative records. While there is some overlap in comments, each commenter has provided a unique perspective on key issues and helped to highlight choices between different strategies in dealing with the uncertainty we face. Prewitt says that privacy is about *intrusiveness* whereas Madans states that the individual's ability to *control* uses of their personal information is at the heart of the privacy debate. Duncan and Madans argue that *individual consent* or *proxy consent* is likely to be required to address privacy concerns regarding administrative records use. Fienberg, Zaslavsky and Madans say that disclosure cannot be *avoided*, only *limited*. Lane argues that *technology* is superior to *bureaucracy* in addressing confidentiality and privacy concerns. Duncan and Lane say that *electronic data* (e.g., geospacial, interactional, and transactional) are surpassing *administrative data* as an alternative data source. Scheuren says that the playing field between statistical and administrative agencies is not level because of different *perceptions* about confidentiality. Reiter offers options to allow the individual to accept some loss of privacy through *incentives* or *waivers*.

In this rejoinder I will discuss each of these as a pair of choices and explore the consequence of choosing one over the other. At the heart of each choice is a great deal of uncertainty because we don't know if our views are the same as the public. If we

*Formerly Chief Privacy Officer, U.S. Census Bureau. Served as Chief of the Census Bureau's Policy Office from 1998–2005 where he led the establishment of the Census Bureau's Data Stewardship Program. He has worked on privacy, confidentiality, and data access issues and supported statistical uses of administrative records for over 25 years. mailto:gwgates@verizon.net

don't get a handle on it though we won't be prepared for the major changes that Prewitt believes are coming.

1 Intrusiveness or control?

Prewitt and Fienberg argue that privacy is about the individual's desire to be left alone—"don't ask me, it's none of your business." Based on his experience as Census Bureau Director and the research he cites, Prewitt argues that privacy in the context of statistical surveys and censuses is really about a perceived intrusiveness and (I conclude) a questioning of the validity of the request. This approach looks at privacy as entirely focused on the point of collection and considers all that follows to be about confidentiality.

On the other side, statistical agencies have viewed privacy based on what the Privacy Act requires—collect only information necessary and relevant to authorized programs; ensure the adequacy of the information for the intended use; convey information about the purpose, uses, users, and any obligations to provide the information; and provide the opportunity for the individual to make choices about any uses not explicitly allowed in the Act or identified, as in a Privacy Act routine use, by the collecting agency. They also view statistical use limitations as pertaining to privacy since the Privacy Act acknowledges the inherent privacy protections in statistical uses. In this view, confidentiality is only about the obligation to make sure unauthorized persons cannot identify respondents or their answers.

As Madans notes, the agency view of privacy is focused on the individual having control over uses and users of their personal information. This definition compares favorably with the one posted on the Privacy Rights Clearinghouse Web page: "The Right to Privacy refers to having control over [one's] personal information. It is the ability to limit who has this information, how this information is kept and what can be done with it."[1]. This definition seems to acknowledge that people routinely are asked for, and provide, details about their personal lives in interactions with government, organizations, and businesses.

So how do we reconcile these differences in definition? Roger Clarke provides insight to the distinctions between the intrusiveness and control aspects of privacy. He notes that the 1890 Warren and Brandeis interpretation that privacy is "the right to be let alone" was influenced by the author's preoccupation with incursions by the media into the lives of the influential and may not have had the broad implications that it seems to imply (Clarke, 2006). He further suggests that there are four dimensions of privacy: 1) privacy of person; 2) privacy of personal behavior; 3) privacy of personal communications; and 4) privacy of personal data. Clarke lumps the later two into what we know as Information Privacy: "the interest an individual has in controlling, or at least significantly influencing, the handling of data about themselves." He further notes: "An important implication of the definition of privacy as an interest is that it has to be

[1] See `https://www.privacyrights.org/why-privacy`

balanced against many other, often competing, interests. At the level of an individual, it may be necessary to sacrifice some privacy, in order to satisfy another interest. The privacy interest of one person may conflict with that of another person, or of a group of people, or of an organisation, or of society as a whole. Hence: Privacy Protection is a process of finding appropriate balances between privacy and multiple competing interests" (Clarke, 2000).

The intrusiveness view of privacy leads one to conclude, as Prewitt does, that information sensitivity is of growing importance to the public and that promising confidentiality and limiting uses to statistical uses will not necessarily be persuasive. In claiming that the request is none of the government's business, the individual is really saying "don't ask me because these are private matters beyond what you need to do your job." This argues for helping people understand the authority for asking these questions, why they are needed, and how they will be used in government statistics. This is something we should be doing under the Privacy Act and the Paperwork Reduction Act but, I argue, we are not doing well enough.

The control view of privacy argues that accepting some level of intrusiveness, whether by government or businesses, is part of life and people routinely provide their personal information in exchange for goods, services, government benefits, or because of civic duty. Individuals relinquish some privacy when giving up their personal information. In exchange, they retain control over how their information is used (limited to statistical uses) and who can see it (persons sworn to protect confidentiality).

This view does not discount Prewitt's observations from 2000 but rather acknowledges that, for some people, it was not acceptable to relinquish privacy for sensitive items in the interests of society even if they retain some measure of control. For these people, persuasive arguments are needed to help them understand why sensitive questions are being asked and how they will be used—beyond the notion of "statistical purposes"—with the goal of helping them see the value to themselves or society. Since privacy protection involves finding appropriate balances, limits on uses and users are also important messages.

A companion issue that Prewitt raises is about the role of political leaders in influencing views of government intrusiveness. As Prewitt reports, research has shown that during the 2000 Census collection cycle, concern that the information requested was an invasion of privacy grew in large part due to negative comments by opinion leaders. Thus, despite a well-financed advertising campaign that highlighted community benefits and included concepts of confidentiality and use-limitation, some people began to doubt the value of the census or did not accept some of the questions as important to the census purpose. This same thing could happen to agencies as a result of record linkage activities which could also be seen as government intrusiveness. To some degree, the case against question sensitivity can be overcome by concise examples of benefits from the census/survey/linkage and how specific personal information contributes.

In their book *The Hard Count* Prewitt and his co-authors argue for an aggressive, bipartisan public education program to help the public understand that data produced by the federal statistical system are a public good that require the public to accept some

government "intrusiveness"(Hillygus et al., 2006). In the past, issues of privacy were not so urgent or complex and agencies could assert their legal authority and track record as grounds for cooperation. We are now compelled to explain how government statistics are different from all other requests for personal information and why this matters to the individual, their families, and their communities. It is a major effort that can typically only be cost-effective for the decennial census but may be accomplished routinely if statistical agencies pool resources. Opinion leaders will continue to challenge these arguments for various reasons and some people (maybe a growing number) will follow along, but that does not absolve us of making our best case.

In summary, I would argue that the definition of privacy is less important than the policies that guide the collection, use, and dissemination of personal information. These policies typically reflect notions of both intrusiveness and control. As Fanning notes, "In (the area of privacy and confidentiality) definitions do not provide much framework for policy or other choices. In many instances a precise term and its meaning are not important in making sound decisions about use and disclosure of personal information. It is more important to protect privacy than to be able to define it" (Fanning, 2007).

2 Consent for statistical uses of administrative records?

Madans states that we do not know whether the public is willing to "cede control" of their personal information despite agency promises to maintain confidentiality and limit uses to statistical uses. If they are not willing, then the challenge is to develop ways to get *meaningful* consent for obtaining and using administrative records. Duncan also addresses consent in the context of whether individuals should be able to opt out of data transfers for statistical uses of their information and if so, how to meaningfully convey all such uses. He argues that, in some contexts, a more realistic approach would be to have an independent body consider the risks and decide if the data can be transferred for specified statistical uses.

This discussion highlights the role of the individual in deciding uses when information is obtained from multiple sources. Typically, "statistical uses" are lumped together when describing the planned uses of personal information in statistical programs. As I mention in the paper, when linkages are planned, statistical agencies often provide the individual an opportunity to opt out, though the consent notices are not uniform and may not be effectively informing individuals of the agency's plans. Administrative agencies are not required to spell out all planned statistical uses, although they must specify the agencies with which the information will be shared. Since statistical uses are routine uses under the Privacy Act, they are generally not spelled out in great detail. Refusing these uses is unlikely since there is no mechanism for opting out of one specified use. Opting out of all uses would mean not getting the benefits for which the person applied—also unlikely.

When Madans says that people cede control of their personal information, the reader might assume that agencies have been given control in deciding uses and protections. This is true only to a degree. All uses must be statistical and confidentiality must be

protected. Agencies decide statistical uses based on program requirements. They also decide on protection measures that will sufficiently limit disclosures. Agencies do not spell out to the individual all statistical uses or all protection measures. In deciding what, if anything, to say about record linkages, agencies are guided by law, ethical obligations, and precedent, as I mention in the paper. The issues Madans raises about appropriateness of consent statements and how changes may impact data access are important and require further discussion and research.

Duncan also discusses the possibility that individuals be able to opt out of data sharing for some statistical uses but recognizes that this would negatively affect the statistical programs. He notes that under Fair Information Practices, information should only be used for the purposes for which it was collected. I argue that statistical uses are compatible and this view has international support. The Council of Europe's Recommendation Concerning the Protection of Personal Data Collected and Processed For Statistical Purposes notes that "Processing for statistical purposes of personal data collected for non-statistical purposes is not incompatible with the purpose(s) for which the data were initially collected if appropriate safeguards are provided for, in particular to prevent the use of data for supporting decisions or measures in respect of the data subject." The report goes on to say that "Processing or communication for statistical purposes of personal data collected for non-statistical purposes shall receive suitable publicity."[2]

A key point of my paper is that much of the privacy uncertainty revolves around people not being aware that certain personal information may be given to a statistical agency to be used for various statistical uses. The Council of Europe recommendation clearly recognizes the importance of public awareness and there are various ways to accomplish this as discussed in my paper. Duncan's argument for an independent board to assess privacy risks for the affected population and determine if a blanket consent is warranted deserves to be part of this discussion.

3 Avoid or limit disclosure?

Fienberg, Madans, and Zaslavsky each affirm that statistical agencies do their best to ensure that statistical data products released to the public do not directly or indirectly identify individual respondents. Nevertheless, methods to protect against disclosure are not foolproof and there is a residual risk of an intruder identifying a respondent based on external data and techniques that the agency did not consider or were not available when the data were published. Fienberg asserts that we cannot "avoid" disclosure but can only "limit" it. I agree.

Given this residual risk, the statistical agency's responsibility is to determine whether it has exercised due diligence in making disclosure an unlikely occurrence. As Madans notes, an important issue that agencies have not addressed is what respondents must be told about this residual risk of disclosure. I would argue that the same could be said

[2] https://wcd.coe.int/wcd/com.instranet.InstraServlet?command=com.instranet.CmdBlobGet&InstranetImage=567406&SecMode=1&DocId=578856&Usage=2

about the risk of a security breach at the agency where personal information is lost or stolen.

Zaslavsky comments about the potential losses from a disclosure and argues that disclosure is only significant if sensitive information is revealed. He states that if the information is less valuable (non-sensitive, has errors, or is not current), the harm to the individual is reduced from a breach. He recommends better understanding of the nature of potential losses in choosing a de-identification policy.

This discussion may benefit from an analysis of the government's response to agency data breaches. OMB (2007) lays out conditions under which breach notification is required. It states that "Agencies should bear in mind that notification when there is little or no risk of harm might create unnecessary concerns and confusion." Five factors are to be considered in assessing likely risk of harm (information sensitivity, number of people affected, likelihood that the information is accessible and useful to others, likelihood the breach may lead to harm, and the ability of the agency to mitigate risk of harm). The same rationale could be applied to determining whether and how to inform individuals of potential disclosure risks in published data and in choosing a de-identification policy. For instance, if the data are quite sensitive, information about disclosure protections may be included in confidentiality pledges or agencies may decide to apply extra restrictions on access. Less sensitive data would require less information about protection measures and fewer restrictions on access.

4 Administrative data or electronic data?

I chose to focus my discussion on government administrative records, while recognizing that databases maintained by the private sector are increasingly being used by statistical agencies to support their program needs. Duncan and Lane point to digital data (such as transactional data, social networking data, and geospatial data) as important sources of such information as job availability, product pricing, consumer expenditures, and housing starts. In *The Hard Count*, Prewitt and his co-authors note that commercial data have their own set of privacy issues in that there are few controls over what is collected, how it is collected, or how it is used. They note that it is not so easy to say "leave me alone" when faced with giving up the conveniences of the digital marketplace. Lane counters that transactional data do not need to be centrally housed and controlled by government statistical agencies, thus limiting confidentiality and privacy concerns.

The privacy and confidentiality issues involved in the use of private sector digital data are important to consider. I offer the following as a signal that the public is considering corporate databases as a growing privacy threat:

I just recently received an email from the Consumer's Union (publishers of *Consumer's Report*) asking me to forward a letter to my senators and representative in Congress expressing my concern about companies selling my personal information to third parties. The request drew my attention when it noted that "82 percent of respondents [to a national survey] were concerned about companies selling or sharing their in-

formation without permission." The Congressional form letter prepared by Consumer's Union read:

> - When a company is just selling my information to anyone who will pay for it, I should have the right to have that report deleted.
> - If it is incorrect, I should be able to fix it.
> - And I should not have to pay to find out what's in my own record with that company.
>
> As online databases get more sophisticated, and companies begin to aggregate information about me from both government and business sources, I deserve to have a say in what is being said and shared about me and my family.

I forwarded this message to my senators and representative and was informed of actions each was taking to support legislation to protect consumer privacy.

I agree with the commenters that we should explore ways to incorporate different types of data (including digital data generated by the private sector) into our statistical infrastructure. New issues will arise, however, because companies are subject to sector-specific privacy laws (health, finance, telecommunications, etc.) that impact their ability to share their data. The public's views on sharing these data may also differ from their views on sharing administrative data and need to be studied. Under Lane's view that federal statistics be produced by the private sector with no individual data being shared with government agencies, privacy and confidentiality issues will change but not disappear.

5 Policy (bureaucracy) or technology?

Lane argues that confidentiality and privacy are best protected through technical solutions rather than through "bureaucratic" policies and procedures. Specifically, she dismisses most of what I propose in Section 9 of my paper—a larger role for OMB, changes to the Privacy Act, and programs like Data Stewardship that are designed to strengthen policies and procedures. She favors technology approaches such as the NORC data enclave and new ways to manage data using the cyber infrastructure. I would argue that the technology approaches she suggests are compatible with my recommendations but are not sufficient by themselves, and are most likely to be put in place if we have the structures and changes I recommend.

As long as federal agencies are the primary producers of national statistics—Lane argues they should not be—policy will continue to play a critical role in accessing, using, and disseminating statistics generated from third-party data.[3] This is not to

[3]The Billion Prices Project is an example of how non-government entities can use transactional data to generate national statistics. Although there are considerable advantages to this methodology, there are also shortcomings. The Wall Street Journal notes that "There are some limits to the method.

say that technology is not important. Automated data management systems, like the Census Bureau's Administrative Records Tracking System, establish rules that ensure that agency users get only the data they need and use it only for authorized projects. This has been critical in reestablishing trust between the Census Bureau and the IRS after the 2000 IRS security audit.

I recommended these policy options because I realize that people make decisions about what laws mean, about whether to share data, about the ethics of their actions, and about how to apply technologies to protect confidentiality. Technologies have an important role in facilitating options that work better for all parties after these decisions are made.

6 Perception or reality?

Scheuren writes in his comment that there is an uneven playing field when it comes to the Census Bureau seeking administrative data vs. an administrative agency seeking data products from the Census Bureau. He notes that confidentiality risks exist when data are shared in either direction and suggests that administrative agencies consider approaches employed by statistical agencies to develop synthetic files instead of providing the "real" data to the statistical agency. He supports this notion based on past improper uses of tax data by the Census Bureau (as I reference in Potok (2009)). While it is clear that the Census Bureau did not follow reporting requirements and did not get approval for joint projects with other agencies, I take issue with Scheuren's claim that there were "breaches by Census staff." In fact, Potok reports that "there had not actually been any breaches or leaks of information from Census staff, the RDCs, BLS, or the BLS contractor."

The point Scheuren makes about leveling the playing field is important to this discussion and relates directly to two factors I identify in the paper on negotiating access—incentives and public support. I acknowledge that there is a disconnect between how the Census Bureau and the IRS interact that is influenced by trust and perception. A key reason for this is the enforcement functions of the IRS.

I personally have been involved in negotiations between the Statistics of Income Division of the IRS and the Census Bureau on a joint project that would use linked data for studies related to tax planning and enforcement by the IRS and the Office of Tax Analysis at the Department of Treasury. The Census Bureau was unwilling to proceed with the project even though the data provided to the IRS and Treasury would undergo disclosure review and confidentiality would be protected. The rationale was that the Census Bureau did not want to appear to be supporting the tax enforcement functions of the IRS. This reaction to public perception is not without merit. A few years ago

They don't reflect haggling for lower prices over items such as cars. And there is no easy way to track prices for services like health care. So while the economists' measure has so far tracked the official price data well, it might stray in the future"Lahart (2010). It is therefore more likely that transactional data will be a valuable supplement, rather than a replacement to, federal statistics—at least in the near term.

the Census Bureau found itself on the defensive when it was reported that the agency produced disclosure-proofed tables focused on Arab Americans for the Department of Homeland Security. Public perception is an important part of this debate and should be part of a more public conversation.

Regarding Scheuren's proposal for synthesizing administrative data before it is provided to statistical agencies, on the surface it sounds like a reasonable approach. The parallels with synthesizing public use microdata, however, are not so apparent to me. Microdata are a public good and can be obtained by both researchers and government officials. Should a administrative agency official decide that public use microdata are valuable to some administrative function such as determining eligibility or enforcing legal compliance, there is no law to prevent this. That is why rules on release of public use microdata take into account the risks that identification may be easy for those who hold files with items of personal information that are identical to information on the public use file. Agencies may promise not to attempt to re-identify individuals but there are no legal penalties if they do.

On the other hand, statistical agencies are legally prohibited from using administrative data for non-statistical purposes. Failure to apply appropriate protections that put the data at risk would also put statistical agency data at risk since the data are linked. In my mind, the only justification for synthesizing on the other end (at the administrative agency) should be based only on whether an extra level of protection is deemed necessary and the synthesized data are acceptable for the statistical purposes.

7 Incentives or waivers?

Reiter offers suggestions on providing individuals additional choices in sacrificing some privacy. He proposes introducing procedures into the data collection process that would allow individuals to identify confidentiality preferences and establish re-use fees for more sensitive questions. In the paper, I cite legal support and precedent for asking individuals to waive confidentiality where there is more than a minimal risk of disclosure.

The idea of asking for confidentiality preferences at the point of collection is intriguing and deserves consideration and testing. Reiter identifies several important methodology questions that need to be considered in such an approach. Another key consideration is that the risks for accepting anything less than full confidentiality protections must be clearly conveyed to the person being interviewed. Also, where one person is responding for others in the household, it is important not to assume the other persons agree to accept this risk.

Re-use fees are also worth exploring, though the operational issues may be overwhelming. Also, requiring consideration are issues common with survey incentives-namely fairness and equity. Accepting a greater confidentiality risk may be more attractive to lower income persons who attach more value to the re-use fees. Is it fair to entice one class of people to give up rights provided to all?

8 Conclusion

In a concluding comment let me acknowledge my failure to report on recent research cited by several of the commenters. I apologize to all whose important contributions I overlooked and thank each of the commenters for expanding the scope of this discussion. I also acknowledge my cursory overview of policy and laws pertaining to privacy and confidentiality. A much more detailed analysis can be found in Fanning (2007).

References

Clarke, R. (2000). Beyond the OECD guidelines: Privacy protection for the 21st century. Xamax Consultancy Pty Ltd. Canberra, January 4, 2000. http://www.rogerclarke.com/DV/PP21C.html.

— (2006). What's privacy. Paper presented for a workshop at the Australian Law Reform Commission on July 28, 2006 (revised August 7, 2006). http://www.rogerclarke.com/DV/Privacy.html.

Fanning, J. (2007). Policy and Best Practices for Ensuring Statistical and Research Confidentiality. PO HHSP 233200500320A, Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services. P. 5.

Hillygus, D. S., Nie, N., Prewitt, K., and Pals, H. (2006). *The Hard Count: The Political and Social Challenges of Census Mobilization*. New York: The Russell Sage Foundation.

Lahart, J. (2010). A way, day by day, of gauging prices. The Wall Street Journal: Economy. http://online.wsj.com/article/SB10001424052748704804504575606801972873866.html.

Office of Management and Budget (2007). Safeguarding Against and Responding to the Breach of Personally Identifiable Information. Memorandum M07-16, OMB. http://www.whitehouse.gov/sites/default/files/omb/memoranda/fy2007/m07-16.pdf.

Potok, N. (2009). Creating Useful Integrated Data Sets to Inform Public Policy. Master's thesis, Columbian College of Arts and Sciences, George Washington University, Washington DC. P. 124. http://gradworks.umi.com/3368742.pdf.