Comment

Janet L. Norwood

Duncan and Pearson's article provides an insightful treatment of complex issues ranging from protection of respondent rights to operational methods for protection against disclosure. The authors lay out thoughtful arguments for increasing access, review methods for masking data and provide suggestions for bringing academic analysis and data protection into a more open environment than currently exists. Their discussion is comprehensive and balanced for data collected about people. But the article would have been more useful if it had been broadened to cover data collected from business establishments where disclosure presents more formidable problems.

Duncan and Pearson recognize the tension between academics who wish to access microdata and statistical agencies who wish to protect data privacy and confidentiality. As one of the data stewards for microrecords collected by the Bureau of Labor Statistics with a pledge of confidentiality, I understood fully the concern that statistical agencies have in carrying out their responsibilities. As an economist with a real interest in academic research, I also know that data are needed for research. And I know also that modern research involves microdata, the computer and the ability to match observations. It is easy to sympathize with, indeed, to agree with, the intent of the authors. Their discussion is useful, but, unfortunately, it does not solve the dilemma we face.

It is time that we realized that the laws and customs under which we operate are somewhat contradictory in concept. Preservation of the right of privacy is a basic right in our society. The Privacy Act of 1974 as amended (5 U.S.C. 552a) prevents disclosure of records maintained on individuals while the Freedom of Information Act (5 U.S.C. 552) prevents government agencies from refusing to provide information to the public. Exceptions in the Privacy Act are designed to permit law enforcement; the Freedom of Information Act exemptions protect confidential commercial and financial information that might, at times, be useful in law enforcement. In addition, a series of laws, judicial opinions and administrative orders affect the release of data by the major agencies in the federal statistical system. But we must also remember the statistical roots underlying the preservation of confidentiality. The respondent's belief in the agency's ability to preserve the confidentiality of the data provided tends to ensure cooperation and to enhance quality. Against this background, the bias of the Federal data steward is to withhold microdata rather than to provide it. One can, therefore, sympathize with Duncan and Pearson's intent to work within that set of biases to enhance access to microdata.

The article challenges the statistical data stewards to find new ways to provide microdata access and provides a useful review of a series of approaches that might be considered. Most of the attention in this field thus far has been given to methods for providing users with microdata that has been transformed in some way to mask the identification of respondents. One approach involves combining, deleting or altering the number of records. A second approach alters attributes within each record. And the third method adds random or deterministic noise to the microdata. Of course, all masks reduce the value of data for making statistical inferences, but the most troubling, it seems to me, is the addition of noise. The addition of random noise to microdata could produce problems for research that could be very difficult to overcome even with advanced statistical methods. Because Duncan and Pearson's purpose is to provide an overview, they pay little attention to the issue of how to determine what information is sensitive. What is acceptable risk? And how do the risks vary by the content of the data, the kind of respondent or the user of the data? As a practical matter, it seems that each data file might require a different masking technique.

Several aspects of Duncan and Pearson's vision and proposals for the future relate to the behavior of the researchers who wish to have increased access to the data. The options range from admonition ("take more responsibility"), to a code of conduct, to licensing and even to new legislation which would prescribe penalties for disclosure. Their review of these possibilities is very useful but demonstrates, I think, the difficulty we face in arriving at one set of standards to apply to all data sets and for all purposes. This imposes a considerable burden

Janet L. Norwood is Commissioner of Labor Statistics, U.S. Department of Labor, Washington, D.C. 20212.

on the statistical agency, which has guaranteed the confidentiality of the data, because it must consider policies for a range of unknown risks and uses and then police compliance by users. Establishment of a staff of "gatekeepers" could be useful, but could such a staff built essentially to service academic users be justified in the current tight budget climate?

The authors are correct in suggesting that we need to focus more attention on obtaining the informed consent of respondents. I fully support their suggestion that federal agencies need to conduct pilot studies to assess the effectiveness and respondent understanding of the statements used in data collection. In fact, the Bureau of Labor Statistics currently has underway, with IRS sponsorship, research into respondents' understanding of and reaction to the language used in confidentiality statements. Much more work needs to be done in applying the laboratory techniques that combine the cognitive sciences and survey research to assess these issues.

The focus of the Duncan and Pearson article is on data about individuals. But confidentiality problems with establishment data are much more complex than for those about people. For one thing, there are fewer establishments than there are people. Businesses can much more easily be classified into subgroups, often with a very small number of units in the groups of particular interest. In addition, a good deal of information about business establishments is available in publicly accessible files that can be matched to the federal system's file and then used to help to disclose confidential data. Moreover, the value of such data to Duncan and Pearson's data spy might be much greater than the value of the data collected about individuals—to say nothing of those who wish to use such data in prosecution and enforcement.

The risk of disclosure is also generally greater for establishment data than for data about individuals, and the stakes for the company can be quite high when trade secrets or business practices are involved. On the other hand, some data—for example, the number of employees or the identification of major products—may not be sensitive at all to some firms but of great concern to others. There is no simple formula for determining which items are the most sensitive.

The problems involved in finding methods for improving access to microdata on establishments for research purposes are complex and difficult, but the need to find solutions is becoming increasingly necessary. Academic researchers are becoming more and more interested in the use of longitudinal microdata files on business establishments, and access to such data would clearly improve some of the public policy research. Statistical agencies have only just begun thinking about these issues, however, and much more work needs to be done.

Duncan and Pearson are quite right in pointing out that research interests and computational capabilities have led to new and more varied demand for publicly collected data. They are also quite right in pointing to the slow and somewhat negative responses from the nation's primary statistical agencies. But their suggestions, while useful, do not point the way to a quick and clear solution. We in the statistical system strongly believe that the absolute protection of confidentiality tends to assure the cooperation of respondents in voluntary surveys (and most government surveys are based on voluntary cooperation) and enhances the quality of the responses. It is true, however, that statistical agencies have not done all that they could to find ways to provide researchers with the data they need within the practical and legal constraints under which the agencies operate. The article properly challenges the nation's statistical system to revisit the confidentiality practices now in place. In doing so, it serves a valuable function. But the problems we face are real, they are complex and there is no easy and quick solution to them.

Rejoinder

George T. Duncan and Robert W. Pearson

While generously acknowledging the centrality of the themes we identify, the discussants quite rightly point to wider issues that should command our attention in the future. To give structure to these issues, we cast the discussants' insights into a set of nested frames. The outer frame encompasses the functional effectiveness of a government statistical system in a diverse society with democratic aspirations. The middle frame delineates the nature of the data that society collects and main-