Discussion

Robert E. Fay

Westat, Inc.
Westat, Rm RE 404, 1600 Research Blvd., Rockville, MD 20850

I'd like to thank John Eltinge for inviting me to discuss this excellent session on the statistical uses of administrative records. Much of the progress in this important area of application has been built on the basis of case studies, but the authors seek to identify generalizations that have emerged thus far and to define a research course for the future. In other words, the interest here is in "the big picture" of how administrative records can be used to improve statistic products.

The complementary coverage of the three papers exemplifies the adage that "the whole is greater than the sum of its parts." Fulton, Martinez, and Eltinge view the issue from a general governmental perspective within the federal statistical system. Their paper frames issues in using administrative records according to a lifecycle model and the general quality perspective of Brackstone (1999) and others, proposing a schematic modeling of quality. The focus of their paper is on the question of how to pursue useful results within the statistical standards applied more routinely to sample surveys and censuses.

Gates focuses his discussion on the vital issues of privacy and confidentiality. For good reason, these issues are often regarded as obstacles to overcome by other researchers hoping to innovate in this area. Instead, his paper provides a thoughtful account of these issues from the perspectives of both the statistical and the administrative agencies. Particularly, administrative agencies are constrained by their mission and often law in their sharing of administrative data. The author first summarizes the legal and policy issues. He then provides an illuminating account of how risks are assessed, both for privacy and confidentiality. His paper is particularly clear on the distinction between the two. For example, he illuminates an important circumstance: When statistical agencies consider linking of data that they already protect under confidentiality safeguards, it is the issue of privacy, rather than confidentiality, that becomes a primary concern in the linking of such data.

Davern, Roemer, and Thomas provide a perspective on these issues from the viewpoint of the research community. The authors compare the strengths and weaknesses of survey data and of administrative data, in part to appreciate the potential for drawing on the strengths of each. They then describe a broad research agenda on the quality of linked data, framed in the familiar concepts of survey error.

Taken together, the three papers provide an excellent overview of many of the challenges faced by government and external researchers in this area. In addition, the individual papers bring out specific points particularly worth highlighting. For example, in their review of quality, Fulton, Martinez, and Eltinge remark on the legislative directive stemming from the Information Quality Act of 2001. The legislation frames quality in general terms, but I concur with the decision of the authors to use the more familiar structure proposed by Brackstone (1999):

- Relevance
- Accuracy
- Timeliness
- Accessibility
- Interpretability
- Coherence

The last two are particularly important challenges in the decentralized statistical system of the U.S.

Gates included his own reflections on both the attraction and the obstacles to using administrative records in the U.S. decennial censuses. He and I share a belief based on general impressions rather than specific evidence, namely, that limited use of administrative record data in the census may be possible, but that extensive use would risk public cooperation and trust in the census. His example provides an interesting test case for the field as a whole.

Davern, Roemer, and Thomas make the excellent point that external researchers, who typically pay close attention to statistical patterns and anomalies in the data, are in a unique position to contribute the quality of the data. Because external researchers typically face additional challenges to obtaining access to linked statistical files than agency employees, they point out the possibility of a quid-pro-quo: access in exchange for information of data quality issues that the external researchers uncover. Their discussion of missing data issues is also especially notable.

The papers leave a few opportunities for a discussant to make suggestions. The paper of Fulton, Martinez, and Eltinge goes to considerable length to frame the issues of the statistical use of administrative data in terms, it would seem, of statistical decision theory. Generally, the approach attempts to structure decision problems in a rational manner, but many researchers are likely to find that working with the theory, supplying the required inputs, and communicating the reasoning and results to others are significant obstacles. I would like to encourage other researchers not to be overwhelmed by the formalism the paper proposes, and that much valuable research proceeds without it.

After summarizing the issues surrounding confidentiality and privacy, Gates proposes a program of research on these issues, specifying both the goals and methods to pursue them. Addressing my comment both to the author and to other researchers in this area, I would encourage consideration of additional research methods, drawn, in part, from social psychology and other behavioral sciences. Admittedly, my suggestion is general, rather than specific, but further discussion by the research community of how to address the research goals he proposes might be fruitful.

Finally, I was pleased that Davern, Roemer, and Thomas offered a rich set of suggestions for future research, particularly on linked data sets. In practical terms, however, the research program will have to recognize resource constraints. Even a simplified "cost-benefit" analysis might help to filter out research efforts with low expected yield relative to cost. For example, the topic of paradata has attracted recent interest, and it appears in their paper. To the extent an administrative agency already produces paradata, analysis of such "long-hanging fruit" does make sense. However, an effort to restructure the capture of administrative data to obtain paradata that did not previously exist may represent a substantial cost to the administrative agency. These efforts are likely to succeed only if supported by a particularly strong base of prior research or a compelling case that the administrative agency stands to benefit from the investment.

In conclusion, the three papers succeed at their goal of providing a high-level overview of important issues in the field, and their insights should guide future research.

References

Brackstone, G. (1999), "Managing Data Quality in a Statistical Agency," *Survey Methodology, 25,* 139-149.