SHARING AGRICULTURAL DATA FOR RESEARCH:

BENEFITS, COSTS, AND LEGAL AND PROPERTY ISSUES

Issues Paper Task Force
AAEA Economic Statistics Committee

Leroy J. Hushak, Chair
Mary Ahern
Wen S. Chern
Laura Blanciforti
Edward I. Reinsel

Background paper prepared for the Symposium on Sharing Agricultural Data for Research: Two Examples of Federal Data Sharing at the AAEA summer meetings, Reno, Nevada, July, 1986.

SHARING AGRICULTURAL DATA FOR RESEARCH:

BENEFITS, COSTS, AND LEGAL AND PROPERTY ISSUES

Issues Paper Task Force
AAEA Economic Statistics Committee

Leroy J. Hushak, Chair
Mary Ahearn
Wen S. Chern
Laura Blanciforti
Edward I. Reinsel

INTRODUCTION

Many of us have shared data sets with colleagues. However, the public data we use is generally from data collecting organizations such as the National Agricultural Statistics Service (NASS) (formerly the Statistical Reporting Service), the Bureau of the Census and other federal or state statistical organizations. Such data may have much of its usefulness removed through summarization or other changes in its form. If we share personal data sets, it tends to be with colleagues who are close friends.

Concern with data sharing is timely. There is high probability that federal data budgets will be reduced, at least in real terms (Wallman, 1986). University research budgets will also become tighter if the deficit reduction efforts of the President and Congress are implemented. The importance of easier and more complete access to Federal data will increase as the ability of state researchers to collect their own data declines. Also, Federal needs for state and university data sets will likely expand.

Broader sharing of data has considerable potential benefit for the Agricultural Economics profession. Not only can research efficiency be

improved, but new lines of research can be opened and public understanding of economic issues enhanced. Thus, we believe discussion of data sharing in this symposium will benefit our profession and the public.

A distinction can be made between the sharing in electronic form of data bases which are essentially collections of published information and the sharing of previously unpublished microdata. The profession has much to gain from improved sharing of large data bases. An example is data base development and greater sharing by the Economic Research Service (ERS). Because the sharing of collections of published data is less inhibited by legal, bureaucratic, and cost related difficulties, it is likely to proceed rapidly. Sharing of microdata, i.e., the sharing of individual observations, is by nature a more sensitive area.

Further, with tighter data collection budgets, the social value of particular data sets will increase. Given the public goods nature of data, i.e., its use by one person does not diminish its usefulness to other users (except when another researcher would use it to investigate a problem which is already being examined with the data set), the more a data set is shared the greater its total social value and the less is duplication of effort involved.

Another reason for sharing data is the apparent increasing difficulty of verifying the validity of empirical results in professional publications. Dewald et al. (forthcoming) found it impossible to exactly replicate the empirical results from nine papers submitted to the Journal of Money, Credit, and Banking (JMCB), although correction of errors did not affect the conclusions in most studies. Another well-known example is Feldstein (1974), where a major programming error was discovered by Leimer and Lesnoy (1980).

Several well-known data sets in the psychological literature have been discovered to be fabricated.

The proceedings of a recent conference by the Committee on National Statistics (Sharing Research Data, 1985) presents several papers which examine many of the issues, benefits, costs, and restrictions on the sharing of data. In this symposium we propose to bring this debate to agricultural economists. Our focus is on research issues in the sharing of data. Other types of uses in the sharing of data, such as sharing for cooperative extension programs, are covered only to the extent that they overlap with research issues. This paper consists of three parts: benefits and costs of data sharing, data access rights in data sharing, and the current status of various codes which deal with data sharing issues.

BENEFITS AND COSTS OF SHARING

In this section we summarize the benefits and costs of sharing data sets. We restrict our consideration to a summary of the benefits and costs of data sharing from the Issues and Recommendations of The Committee on National Statistics in Sharing Research Data (1985) and in the papers of Clubb et al. (1985) and Hedrick (1985). Hedrick (1985) considers the interests of five parties to any data set. The first of these is the primary researcher originally responsible for collection of the data. Second is the data requester, the researcher(s) requesting release of the data set. The third party is the participant who provides the data, i.e., the person, firm, town, school, etc. The final two groups are the scientific community in total and society. The benefits and costs of data sharing, and the rights and responsibilities to be discussed below, are not distributed evenly across these groups.

Benefits of Data Sharing

The Issues and Recommendations section of Sharing Research Data lists ten benefits of data sharing which we restate here in more concise form. The major benefit is the reinforcement of scientific inquiry. Science can be more efficiently advanced and more effectively applied in decision making when data sharing is fostered. This scientific advancement can be summarized under the principle of learning from the experiences of others. It extends into methodological and empirical areas.

One of the great benefits of data sharing is the insight we gain from the experiences of others. Through the sharing of ideas, the development of theoretical knowledge and the knowledge of analytical techniques is advanced. Multiple perspectives resulting from contacts with researchers in other disciplines are brought out. Better specification of conceptual models and more complete empirical analysis is promoted. This includes the encouragement of more appropriate use of the data in policy formulation and evaluation. Insight into improving measurement and data collection methods from suggestions by those who share the data can be hastened. In many cases the duplication of effort as well as the possibility of error can be reduced.

Finally the increased provision of data resources enables researchers, those already trained and new students, to learn from existing and known problems in data analysis. The verification, refutation, and refinement of results are promoted through data sharing. Increased data availability aids in reducing errors and counteracting fabrication of the data. The insights and improvements that advance science can be tested. Also, the merger of two or more data sets can aid researchers in addressing new and broader issues and in increasing the understanding of social problems.

Costs of Data Sharing

Four classes of costs are discussed in the Issues and Recommendations of Sharing Research Data. The first is the technical obstacles which result from incompatibility of computer machine hardware and software systems. Further, computer technologies not only affect our ability to share data, but provide avenues to dishonest uses of data as well. The second set of costs covers the upgrading of documentation so that data sets can be properly interpreted by sharing users. Documentation is a particularly serious issue with respect to the sharing of privately held data sets. The user needs to be made fully aware of the nature of the data set, e.g., how the sample was selected, what questions were asked, what edits were made on the data, and how randomness of response was examined. A third set of costs is the costs of documentation, storage and transfer of data imposed on the data holder. Research investigators are generally not rewarded for collection and dissemination of data. Further, they lose control of the data set when it is shared and may face criticism from others who disagree with or find errors in their analysis of the data. Finally, subsequent users of the data will likely face substantial costs in making the data usable in their research problem.

Clubb et al. and Hedrick discuss two further data sharing costs. The proprietary rights of the primary researcher must be protected with the first opportunity to analyze the data and make a contribution to the field. Also the confidentiality and privacy of those who provide the data must be protected. National security issues may be involved. Care must be exercised in preventing disclosure. Compliance with disclosure laws is a responsibility which must be upheld by the original data collection agency or individual. Subsequent users

do not have sufficient incentives to prevent disclosure. The costs of complying with disclosure laws could be shared, however.

In addition the issue of who bears the costs of faulty data needs to be considered. Faulty data can result from misinterpretation of questions by the original researcher, from transposition of the data from respondent to interview form and/or computer format, from incorrect analysis by the original researcher, and for other reasons. The secondary user may also misinterpret the primary information, apply incorrect analytical techniques, or otherwise misuse the data. Pressures to produce may supercede quality and responsible behavior towards data sharing.

Pressure on professional persons, whether economist, other social scientist, or other professional, changes the balance of the benefits and costs of sharing data toward that of sharing. Increased academic pressure to "publish or perish" increases the probability that academic professionals, especially those striving for tenure, may exercise less care in data management and generation of empirical results leading to increased likelihood of error, or may fabricate data or results which support their theoretical premises. The potential promise of large rewards for adoption of the "right" policy by Federal, state or local government in conjunction with the growing use of statistical results to document positions increases the probability that massaging the data or outright fabrication will occur in order to obtain the supporting results. Both circumstances suggest more open access to data in order to offset the increased propensity to error or to cheat with an increase in the probability that someone will catch the erroneous results, i.e., an increased incentive to be honest. Feldstein (1974) is an example of an error in empirical results, while Burt's fabricated data on the heritability of

intelligence (Hedrick, 1985, p. 130) is an outstanding example of dishonesty in science. The rights and responsibilities of the parties to a research data set are addressed in the next section.

DATA ACCESS RIGHTS

In consideration of the rights and responsibilities of data use, Boruch (1985, p. 101) identifies two proprietary interests which must be balanced. The first is the right of the individual scientist to analyze data; the second is the right of an institution to control who analyzes it. The right of access to research records depends on the type of funding under which the data are collected (Cecil and Griffin, 1985). The access rights for three general data collection funding arrangements are discussed.

Private Research Records

Research records maintained by a private researcher and supported by private funds are the first type of data set considered by Cecil and Griffin. They indicate that there is no case law or legislation which governs rights in privately developed data sets. Since we are dealing with intellectual property, proprietary rights under copyright laws can be deduced. A research data set meets the two necessary conditions for copyright protection: originality and tangible expression. However, these rights can be difficult to protect since a data set is information and not a tangible good.

The rights of those seeking access to a privately held data set and of the participants are very limited. The researcher can benefit from a data set through publication of analyses without distributing the data set. On the other hand, if the researcher does share the data set, copyright protection

will not bar another researcher from creating an identical data set even if the copyrighted data set is used as the basis.

Government Research Records

Research records maintained by government agencies include both Federal and state data sets. The basic policy governing the sharing of Federal records is the Federal Records Act of 1950, which is part of the Administrative Procedures Act. These policies have been modified by the Freedom of Information Act (FOIA) and the Privacy Act of 1974. A statute in the Administrative Procedures Act which allowed withholding of information 'for good cause' was amended by FOIA to assure the "free flow of governmental information 'necessary to an informed electorate'".

The Privacy Act of 1974 addresses restrictions on the ability of agencies to release identifiable information. The act is the first attempt by Congress to provide comprehensive protection of an individual's right to privacy through the regulation of the collection, management, and disclosure of personal information maintained by government agencies. The privacy act requires 1) access by individuals to identifiable records, while 2) ensuring accuracy and timeliness of information, and 3) limiting disclosure of identifiable information to third parties.

The Office of Management and Budget (1985) circular entitled "Management of Federal Information Resources" is a recent document guiding the collection and dissemination of data by Federal agencies. A key provision of that circular states that "agencies shall create or collect only that information necessary for the proper performance of agency functions and that has practical utility, and only after planning for its processing, transmission, dissemination, use, storage, and disposition" (section 8a). The 1985 Farm Bill

includes confidentiality protection of USDA data by limiting publication of data to aggregate form and exempting the data from mandatory disclosure without the consent of persons involved (Glaser, 1986).

Although we have not reviewed state policies regarding access to State data, state policies likely reflect similar interests in data access, maintaining confidentiality, and institutional control over data usage.

Publicly Funded Private Research Records

The third class of data is research records developed with public funds that are maintained by private researchers. These records are not explicitly covered by FOIA or the Privacy Act of 1974. The National Science Foundation (NSF) and the National Institute of Justice (NIJ) both have formal policies which state that data banks and software will be made available to others. The NIJ policy is explicit in requiring the grantee to submit machine readable copies and adequate documentation of all data bases and programs at no additional costs when the grant is terminated, while NSF policy is not explicit. The National Institute of Education and the National Center for Health Services Research both encourage reanalysis of data, but have no formal policies (Boruch and Cordray, 1985).

CURRENT STATUS OF SHARING CODES

In addition to the data access policies which vary by data collection funding source, there are some codes of ethics which deal explicitly with data sharing. In this section we examine, in summary form, the status of several data sharing codes. While the codes have no legal standing, they do serve as a framework for agreement and codes of good practice. The Bellagio principles evolved from a conference of university and government scholars and bureaucrats

from five countries: U.S., United Kingdom, West Germany, Sweden, and Canada. They endorse the idea of the provision of government data to individual researchers or research institutions for legitimate research purposes, but were not designed to apply to the individual researcher sharing his or her data set with others. The Organization for Economic Cooperation and Development (OECD) has also accepted a set of guidelines for data protection submitted by the U.S.

The American Association for the Advancement of Science (AAAS) has a report of professional ethics in scientific and engineering societies affiliated with AAAS (Chalk el al., 1980). The most frequently appearing statements deal with honesty and balanced reporting. The Joint Committee on Standards for Educational Evaluation (1981) issued professional standards and guidelines for evaluating educational programs, projects and material. The Evaluation Research Society (1980) has a parallel document. Both state that restrictions on access to data need to be negotiated before data is obtained. The American Statistical Association (ASA) in 1980 and 1983 and the American Sociological Association in 1982 also drafted codes of ethics on data sharing. The ASA code recommends that statisticians make data sources available with safeguards for privacy, that transfers are in conformity with pre-established means for the protection of confidentiality, and that data sources be adequately documented (Boruch and Cordray, 1985).

The 1980 code of the Society of American Archivists has several distinctive features (Boruch and Cordray). It encourages quality control and advises members to place materials in repositories where they will be "adequately processed and effectively utilized". Use of holdings is encouraged to the extent that it is consistent with institutional policies, preservation of holdings, legal considerations, individual rights, and donor agreements.

The American Society of Access Professionals was organized in 1980 as a discussion forum for access to information produced by government.

In 1982, the JMCB adopted an editorial policy of requesting programs and data from authors submitting articles for publication, and of making the programs and data available to others (Dewald et al.). In addition to the JMCB, journals or associations which have adopted data sharing policies include ASA, the Journal of Personality and Social Psychology, and the Journal of the American Chemical Society. Dewald et al. argue that an editorial policy which requires submission of programs and data has several significant advantages over no policy. First, authors can supply programs and data sets at least cost when their research is just completed. Second, a journal can provide a cost-effective clearinghouse for data, reducing the cost of a researcher seeking to replicate research results. Third, frequency of errors will be reduced. Fourth, improved evaluation of papers by editors and referees will occur. Finally, the impression of challenging an author's results is avoided when data and programs are obtained from a journal rather than the author.

CONCLUSIONS

The Economic Statistics Committee on behalf of the AAEA has an interest in enhancing the access of its members to high quality data. This paper has described in a qualitative sense the benefits and costs of data sharing, and the current rights, laws, and codes relating to the sharing of data. Sixteen recommendations of the Committee on National Statistics about the sharing of research data are listed in the Appendix. However, many issues remain unresolved.

In the case of benefits and costs, an indisputable dollar value is not possible at this time, and this paper has not attempted to provide one.

However, determining whether or not the net benefits are positive depends on such a quantification. The most significant expected costs in data sharing (the risks of violating the privacy considerations) and the most significant benefits (advancement of knowledge) are also the most difficult to quantify.

The property rights of public data are clear in the case of aggregate data. The Freedom of Information Act, the Management of Federal Information Resources circular, and the latest farm bill all support the right of access to aggregate agricultural statistics. The property rights for access to agricultural micro-data, however, have historically been held by the collecting agency or individual. This is true for two important reasons. First, individual records are more easily identifiable for a larger proportion of farms on agricultural surveys than are individual records for a more general population survey, such as the Consumer Expenditure Survey. Second, the expected costs of disclosure will largely be incurred by the collecting agency. If individual records are shared with researchers outside the collecting agency, then the costs of privacy violation are no longer internalized, and the incentives to avoid them are lessened. Both the relevant language in the 1985 farm bill and a recent informal ruling by the Office of General Counsel in USDA seem to limit the ability by USDA to release microlevel data.

A final issue is in defining the role of AAEA in the data sharing issue. Should the AAEA develop its own data sharing code? Should the AJAE require authors to supply editors with their data sets and computer programs?

The Economics Statistics Committee is bringing these data issues to the attention of agricultural economists because we believe that the ability to do empirical research will be critically affected by current changes in the budgets and priorities of data collection and research funding agencies. These

changes are likely to reduce budgets for data collection by Federal and State data collection agencies as well as the ability to collect primary data within individual research projects, making the sharing of data an important compensating factor. If data sharing is to increase substantially, strategies and procedures must be worked out so that rules and regulations are met and costs are equitably shared.

APPENDIX

The Committee on National Statistics (Sharing Research Data) lists 16 recommendations with respect to the sharing of research data. We list these in order:

1) Sharing data should be a regular practice.

2) Investigators should share their data by the time of publication of initial major results of analyses of the data except in compelling circumstances.

3) Data relevant to public policy should be shared as quickly and widely as possible.

4) Plans for data sharing should be an integral part of a research plan whenever data sharing is feasible.

5) Investigators should keep data available for a reasonable period after publication of results from analyses of the data. 6) Subsequent analysts who request data from others should bear the associated incremental costs.

7) Subsequent analysts should endeavor to keep the burdens of data sharing on initial investigators to a minimum and explicitly acknowledge the contribution of the initial investigator.

8) Funding organizations should encourage data sharing by careful consideration and review of plans to do so in applications for research funds.

9) Organizations funding large-scale, general-purpose data sets should be alert to the need for data archives and consider encouraging such archives where a significant need is not now being met.

10) Journal editors should require authors to provide access to data during the peer review process.

11) Journals should give more emphasis to reports of secondary analyses and to replications.

12) Journals should require full credit and appropriate citations to original data collections in reports based on secondary analyses.

13) Journals should strongly encourage authors to make detailed data accessible to other researchers.

14) Opportunities to provide training on data sharing principles and practices should be pursued and expanded.

15) A comprehensive reference service for computer-readable social science data should be developed.

16) Institutions and organizations through which scientists are rewarded should recognize the contributions of appropriate data-sharing practices.

REFERENCES

Boruch, Robert F. "Definitions, Products, Distinction in Data Sharing." in Sharing Research Data, Stephen E. Fienberg, Margaret E. Martin and Miron L. Straf, eds., National Academy Press, 1985, pp. 89-122.

Boruch, Robert F. and David S. Cordray. "Professional Codes and Guidelines in Data Sharing." in Sharing Research Data, Stephen E. Fienberg, Margaret E. Martin and Miron L. Straf, eds., National Academy Press, 1985, pp. 199-225.

Cecil, Joe Shelby and Eugene Griffin. "The Role of Legal Policies in Data Sharing." in Sharing Research Data, Stephen E. Fienberg, Margaret E. Martin and Miron L. Straf, eds., National Academy Press, 1985, pp. 148-98.

Chalk, R., M. S. Frankel, and S. B. Chafer. AAAS Professional Ethics Project: Professional Ethics Activities in the Scientific and Engineering Societies.: American Association for the Advancement of Science, Washington, D.C., 1980.

Clubb, Jerome M., Erik W. Austin, Carolyn L. Geda, and Michael W. Traugott. "Sharing Research Data in the Social Sciences." in Sharing Research Data, Stephen E. Fienberg, Margaret E. Martin and Miron L. Straf, eds., National Academy Press, 1985, pp. 39-88.

Dewald, William G., Jerry G. Thursby and Richard G. Anderson. "Replication and Scientific Standards in Empirical Economics: Evidence from the JMCB Project." American Economic Review, forthcoming.

Feldstein, Martin. "Social Security, Induced Retirement, and Aggregate Capital Accumulation." Journal of Political Economy 82(1974):905-26.

Glaser, Lewrene K. "Provisions of the Food Security Act of 1985." Economic Research Service, U.S. Department of Agriculture, Agriculture Information Bulletin No. 498, April, 1986.

Leimer, D. R. and S. D. Lesnoy. "Social Security and Private Savings: A Reexamination of the Time Series Evidence Using Alternative Social Security Wealth Variables." Paper presented at the 93rd meeting of the American Economic Association, Denver, Colorado, September 6, 1980.

Office of Management and Budget. "Management of Federal Information Resources." Federal Register, Circular No. A-130, December, 1985.
Sharing Research Data. Stephen E. Fienberg, Margaret E. Martin and Miron L. Straf, eds., Committee on National Statistics, National Academy Press, 1985.

Wallman, Katherine K. "Federal Budget Statistics in the FY 1987 Budget." in AAAS Report XI: Research and Development, FY 1987,
American Association for the Advancement of Science, Washington, D.C., 1986, pp. 257-68.