Statistical Policy Working Paper 20 - Seminar on Quality of Federal Data - Part 1 of 3

Statistical Policy

Working Paper 20

Seminar on Quality of Federal Data

Part 1 of 3

Federal Committee on Statistical Methodology

Statistical Policy Office

Office of Information and Regulatory Affairs

Office of Management and Budget

March 1991

MEMBERS OF THE FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY

(February 1991)

Maria E. Gonzalez, Chair
Office of Management and Budget

Yvonne M. Bishop
Energy Information Administration

Warren L. Buckler

Daniel Kasprzyk
Bureau of the Census

Daniel Melnick
National Science Foundation

Social Security Administration

Robert P. Parker

Charles E. Caudill
Bureau of Economic Analysis

National Agricultural

Statistics Service
David A. Pierce

Federal Reserve Board

Cynthia Z.F. Clark

National Agricultural
Thomas J. Plewes

Statistics Service
Bureau of Labor Statistics

Zahava D. Doering
Wesley L. Schaible

Smithsonian Institution
Bureau of Labor Statistics

Robert M. Groves
Fritz J. Scheuren

Bureau of the Census
Internal Revenue Service

Roger A. Herriot
Monroe G. Sirken

National Center for
National Center for

Education Statistics
Health Statistics

C. Terry Ireland
Robert D. Tortora

National Computer Security Center

Bureau of the Census

Charles D. Jones

Bureau of the Census

PREFACE

In 1975, the Office of Management and Budget (OMB) organized the Federal Committee on Statistical Methodology. Comprised of individuals selected by OMB for their expertise and interest in statistical methods, the committee has during the past 15 years determined areas that merit investigation and discussion, and overseen the work of subcommittees organized to study particular issues. Since 1978, 19 Statistical Policy Working Papers have been published under the auspices of the Committee.

On May 23-24, 1990, the Council of Professional Associations on Federal Statistics (COPAFS) hosted a "Seminar on the Quality of Federal Data." Developed to capitalize on work undertaken during

the past dozen years by the Federal Committee on Statistical Methodology and its subcommittees, the seminar focused on a variety of topics that have been explored thus far in the Statistical Policy Working Paper series. The subjects covered at the seminar included:

- Survey Quality profiles
- Paradigm Shifts Using Administrative Records
- Survey Coverage Evaluation
- Telephone Data Collection
- Data Editing
- Computer Assisted Statistical Surveys
- Quality in Business Surveys
- Cognitive Laboratories
- Employer Reporting Unit Match Study
- Approaches to Developing Questionnaires
- Statistical Disclosure-Avoidance
- Federal Longitudinal Surveys

Each of these topics was presented in a two-hour session that

featured formal papers and discussion, followed by informal dialogue among all speakers And attendees.

Statistical Policy Working Paper 20, published in three parts, presents the proceedings of the "Seminar on the Quality of Federal Data." In addition to providing the papers and formal discussions from each of the twelve sessions, this working paper includes Robert M. Groves' keynote address, "Towards Quality in a Working Paper Series on Quality," and comments by Stephen E. Fienberg, Margaret E. Martin, and Hermann Habermann at the closing session, "Towards an Agenda for the Future."

We are indebted to all of our colleagues who assisted in organizing the seminar, and to the many individuals who not only presented papers and discussions but also prepared these materials for publication. A special thanks is due to Terry Ireland and his staff for their work in assembling this working paper.

Table of Contents

Wednesday, May 23, 1990

Part 1

KEYNOTE ADDRESS

TOWARDS QUALITY IN A WORKING PAPER SERIES ON QUALITY. 3

Robert M. Groves, The University of Michigan and U. S.

Bureau of the Census

Session 1 - SURVEY QUALITY PROFILES

THE SIPP QUALITY PROFILE. 19

Thomas B. Jabine, Statistical Consultant

INITIAL REPORT ON THE QUALITY OF AGRICULTURAL SURVEY PROGRAM . .29

George A. Hanuschak, National Agricultural Statistics

Service

DISCUSSION. 40

Barbara A. Bailar, American Statistical Association

DISCUSSION 46

Nancy A: Mathiowetz, U. S. Bureau of the Census

Session 2 - PARADIGM SHIFTS USING ADMINISTRATIVE

RECORDS

PARADIGM SHIFTS: ADMINISTRATIVE RECORDS AND CENSUS-TAKING. . . .53

Fritz Scheuren, Internal Revenue Service

AN ADMINISTRATIVE RECORD PARADIGM: A CANADIAN EXPERIENCE 66

John Leyes, Statistics Canada

DISCUSSION 77

Gerald Gates, U.S. Bureau of the Census

DISCUSSION 83

Edward J. Spar, Market Statistics

Session 3 - SURVEY COVERAGE EVALUATION

CONTROL MEASUREMENT, AND IMPROVEMENT OF SURVEY COVERAGE 87

Gary M. Shapiro,, U. S. Bureau of the Census; Raymond R.

Bosecker, National Agricultural Statistics Service

QUALITY OF SURVEY FRAMES 100

Judith T. Lessler, Research Triangle Institute

DISCUSSION 108

Fritz Scheuren, Internal Revenue Service

DISCUSSION 114

Joseph Waksberg, Westat, Inc.

Session 4 - TELEPHONE DATA COLLECTION

QUALITY IMPROVEMENT IN TELEPHONE SURVEYS 123

Leyla Mohadjer, David Morganstein, Westat, Inc.

COMPUTER ASSISTED SURVEY TECHNOLOGIES IN GOVERNMENT:

AN OVERVIEW 137

Marc Tosiano, National Agricultural Statistics Service

DISCUSSION 155

William L. Nicholls II, U. S. Bureau of the Census

DISCUSSION. 161

James T. Massey, National Center for Health Statistics

Part 2

Session 5 - DATA EDITING

OVERVIEW OF DATA EDITING IN FEDERAL STATISTICAL AGENCIES. . . . 167

David A. Pierce, Federal Reserve Board

EDITING SOFTWARE (An excerpt from Chapter IV of Working

Paper 18) 173

Mark Pierzchala, National Agricultural Statistics

Service

RESEARCH ON EDITING. 180

Yahia Ahmed, Internal Revenue Service

DISCUSSION 184

Charles E. Caudill, National Agricultural Statistics

service

DISCUSSION 186

Richard Bolstein, George Mason University

Session 6 - COMPUTER ASSISTED STATISTICAL SURVEYS

OVERVIEW OF COMPUTER ASSISTED SURVEY INFORMATION COLLECTION . . 191

Richard L. Clayton, U. S. Bureau of Labor Statistics

A COMPARISON BETWEEN CATI AND CAPI. 197

Martin Baum, National Center for Health Statistics

COMPUTER ASSISTED SELF INTERVIEWING 202

Ralph Gillmann, Energy Information Administration

COMPUTER ASSISTED SELF INTERVIEWING: RIGS AND PEDRO,

TWO EXAMPLES. 205

Ann M. Ducca, Energy Information Administration

DATA COLLECTION. 209

Cathy Mazur, National Agricultural Statistics Service

DISCUSSION 212

Robert N. Tinari, U. S. Bureau of the Census

DISCUSSION 216

David Morganstein, Westat, Inc.

Thursday, May 24, 1990

Session 7 - QUALITY IN BUSINESS SURVEYS

IMPROVING ESTABLISHMENT SURVEYS AT THE BUREAU OF LABOR

STATISTICS 221

Brian MacDonald, Alan R. Tupek, U. S. Bureau of Labor

Statistics

A REVIEW OF NONSAMPLING ERRORS IN FEDERAL ESTABLISHMENT

SURVEYS WITH SOME AGRIBUSINESS EXAMPLES. 232

Ron Fecso, National Agricultural Statistics Service

DISCUSSION. 243

DISCUSSION. 247

Charles D. Cowan, Opinion Research Corporation

Session 8 - COGNITIVE LABORATORIES

THE BUREAU OF LABOR STATISTICS COLLECTION PROCEDURES RESEARCH LABORATORY: ACCOMPLISHMENTS AND FUTURE DIRECTIONS. . .253

Cathryn S. Dippo, Douglas Herrmann, U. S. Bureau of Labor Statistics

THE ROLE OF A COGNITIVE LABORATORY IN A STATISTICAL AGENCY. . .268

Monroe G. Sirken, National Center for Health Statistics

DISCUSSION 278

Elizabeth Martin, U. S. Bureau of the Census

DISCUSSION 281

Murray Aborn, National Science Foundation (retired)

Session 11 - STATISTICAL DISCLOSURE - AVOIDANCE

DISCLOSURE AVOIDANCE PRACTICES AT THE CENSUS BUREAU. 367

Brian Greenberg, U. S. Bureau of the Census

THE MICRODATA RELEASE PROGRAM OF THE NATIONAL CENTER

FOR HEALTH STATISTICS. 377

Robert H. Mugge, National Center for Health Statistics

(retired)

DISCUSSION 385

George T. Duncan, Carnegie Mellon University

Session 12 - FEDERAL LONGITUDINAL SURVEYS

FEDERAL LONGITUDINAL SURVEYS. 393

Daniel Kasprzyk, U. S.reau of the Census; Curtis

Jacobs, U. S. Bureau of Labor Statistics

THE ADVANTAGES AND DISADVANTAGES OF LONGITUDINAL SURVEYS. . . 407

Robert W., Pearson, Social Science Research Council

LONGITUDINAL ANALYSIS OF FEDERAL SURVEY DATA. 425

Patricia Ruggles, Joint Economic Committee

DISCUSSION ... .438

Michael Brick, Westat, Inc.

DISCUSSION 447

Marilyn E. Manser, U. S. Bureau of Labor Statistics

TOWARDS AN AGENDA FOR THE FUTURE

Stephen E. Fienberg, Carnegie Mellon University. 455

Margaret E. Martin. 462

Hermann Habermann, Office of Management and Budget. 465

Part 3

Session 9 - EMPLOYER REPORTING UNIT MATCH

STUDY

INTERAGENCY AGREEMENTS FOR MICRODATA ACCESS:

THE ERUMS EXPERIENCE. 291

Thomas B. Petska, Internal Revenue Service; Lois

Alexander, Social Security Administration

SAMPLE SELECTION AND MATCHING PROCEDURES USED IN ERUMS 301

John Pinkos, Kenneth LeVasseur, Marlene Einstein,

U. S. Bureau of Labor Statistics; Joel Packman, Social

Security Administration

RESULTS, FINDINGS, AND RECOMMENDATIONS OF THE ERUMS PROJECT. . 309

Vern Renshaw, Bureau of Economic Analysis; Tom Jabine,

Statistical Consultant

DISCUSSION.. 318

W. Joel Richardson Charles A. Waite, U. S. Bureau of the

Census

DISCUSSION 324

Thomas J. Plewes, U. S. Bureau of Labor Statistics

Session 10 - APPROACHES TO DEVELOPING

QUESTIONAIRES

TOOLS FOR USE IN DEVELOPING QUESTIONS AND TESTING

QUESTIONNAIRES 331

Theresa J. DeMaio, U. S. Bureau of the Census

TECHNIQUES FOR EVALUATING THE QUESTIONNAIRE DRAFT. 340

Deborah H. Bercini, National Center for Health Statistics

DESIGNING QUESTIONNAIRES FOR CATI IN A MIXED MODE

ENVIRONMENT 349

Gemma Furno, U. S. Bureau of the Census

DISCUSSION 360

Carol C. House, National Agricultural Statistics Service

Part 1

Keynote Address

TOWARDS QUALITY IN A WORKING PAPER

SERIES ON QUALITY

TOWARDS QUALITY IN A WORKING PAPER SERIES ON QUALITY

Robert M. Groves

The University of Michigan and

U.S. Bureau of the Census

1. Introduction.

Although this meeting has the title of the "Seminar on the Quality of Federal Data," its structure follows quite closely the topics covered in the multi-paper series of Statistical Policy Working Papers sponsored by the Office of Statistical Policy and, Standards. There are as of this date, 19 Statistical Policy Working Papers written since the first in 1978. That is about 1.6 per year over the 12 years of the series, (see Figure 1). They range over a wide terrain, involving issues of the topical focus of surveys to a set of methodological and statistical issues affecting survey quality.

I am unaware of the processes that led to my being asked to

give the keynote address at this meeting. I must admit that I speak to you today as someone who has a very biased opinion about the OMB Statistical Policy Working Papers - I love almost all of them; I like the idea that they exist and only recently, because of my change of job sectors, have I appreciated their worth from another perspective. I have used them in graduate courses for students in survey methods (they are fine introductions to important design topics). I have used them in my research work (they are unique sources of documentation about what goes on in the Federal Statistical System). I recommend them to others calling for consulting assistance.

Although I speak as a friend, 45 minutes of praise from me wouldn't act to improve this series and runs the risk of "head inflation" for those who developed the papers. Instead, I want to be a constructive critic and will divide my remarks into several categories:

a. alternative goals of the OMB series

b. the need for a structure to their topics

I note that what follows are my personal views as a close observer from afar of the system and a rookie member of the system.

2. Alternative Perspectives on Goals of the Working Paper Series

2.1. OMB Series as Review of the State of Practice

Some of the papers in the series address a topic that spans many surveys of different populations (see Figure 2). The papers on coverage error and telephone data collection are examples of this. These kind of papers are compact summaries of the state of the art on a current issue facing all surveys. They often describe activities in both household surveys and those in economic surveys. Many times they end with case studies of different surveys across the Federal system and how they handle the particular issue at hand.

Figure 2

Alternative Perspectives on Goals of the Working Paper Series

1. OMB series as a review of the state of practice

2. OMB series as agency cross-fertilization

3. OMB series as a prod to new developments

These kind of papers are valuable to the extent that they have deep depth and wide breadth. By that I mean, they cover all the sources of data quality and cover them in sufficient depth that real learning is likely on the part of most readers.

Let me first speak of breadth of topics. I find it most simple to array the topics of the papers along the components of total survey error (see Figure 3). It is unfair for me to present this chart without some clarifying remarks about the missing cells. First, missingness does not imply absence of any treatment of the topics. Indeed, on sampling error, for example, many of the reports comment on the impact of design options on sampling variance. Second, this structure is only one which could be

applied to classify the xx reports. Considering the label of this seminar "quality of Federal Data", however, I find it attractive to use it here.

Despite the weakness of any one classification scheme, let me point out what I believe are weaknesses with the current status of the series. There is a distinct bias toward the household survey domain to the detriment of the economic domain. There is one paper with the overarching title of "Quality in Establishment Surveys", but the fact that it along exists underscores the problem. This is a reflection of the smaller literature in the methodology and evaluation of quality of economic surveys, but it is a status that

I hope will change in the future. Why? We have in the past too quickly assumed the following premises about economic survey measurement:

a. establishment surveys are too diverse to yield themselves

to common methodologies or standards.

b. establishment surveys do not face questionnaire design issues like those of household surveys because the information gathered is factual in nature

c. establishment surveys have nonresponse properties that do not resemble those of household surveys.

Each of these can be refuted with some observation of the various establishment surveys now ongoing. It is true that establishment populations have large variation in size; that their organizational structures are diverse; that their recordkeeping practices are not standardized; that the ideal respondent for different issues may vary across establishments. All of this is true, but should not lead to the extreme that there Are no common problems either across different establishment surveys or between household and economic surveys.

As the Boskin report has observed, economic survey data needs

improvement and the working paper series could be one vehicle of focusing attention on specific needs in this area.

The next most important omission, in my opinion, concerns the issue of nonresponse. I must admit here that the work of the National Academy of Sciences Panel on Missing and Incomplete Data offers a comprehensive review of current theory and practice. Conversely, the issue is vital to the unique inferential power of probability samples and therefore cannot receive too much attention. Even the most basic issues remain unresolved: relationships between response rates and nonresponse error; relationships between likelihood of coverage and likelihood of participation; cost/error evaluations of alternative methods of improving response rates. Mean square errors of survey estimators stem from thousands of individual decisions to cooperate with the survey request. It behooves us to devote more energy to this and the working paper series should do this.

Third, the interviewer has largely been ignored. It has been ignored despite that fact that many Federal surveys use interviewers to assist in the data collection, despite the fact

that evaluative procedures desperately need review and reconceptualization, despite the fact that it is an area where both statistics and social science perspectives work. The attention to the interviewer is even more important given the likely future in which the traditional labor force of underemployed/overskilled part

time homemakers will decline and computer technologies are likely to transform the job.

Fourth, although large portions of data collection in the Federal Statistical System is by mail and self-administered questionnaire there is no focused treatment of the methodology in the series.

Fifth, a few comments specifically on error profiles. When I first read the CPS error profile 12 years ago, I had two reactions. I was attracted to the literary form -- a compilation of quality

measures for the survey, combined with documentation of design features. I then felt and still believe that the structure of an error profile is a valuable way to document leading components of error in survey statistics (we should be grateful to Brooks and Bailar as the mothers (or midwives) of the invention). My second reaction came after digesting the full report. How little we as a community seemed to know about the error properties of the CPS, the largest ongoing and one of the most important ongoing Federal household surveys. Of the 80 pages of the report, for example, only about 25 are devoted to the data collection operations, a source of most of the errors in the process! That combination of reactions led me to the belief that I still have -- the error profile, in the hands of intelligent program directors, can act as an agenda setting document for quality improvement programs.

Finally, there are no serious treatments of costs of data collection - a topic I'll revisit in a few minutes.

Let me now turn to issues of depth. At their worst the reports are catalogues -- they make great reading for someone

interested in buying an idea from those presented, but they don't make thrilling reading for the uninitiated. At the same time, they often assume knowledge of various data series that is not Possessed by many outside experienced statistical system staff. As a corollary, some fail to cite relevant research literature outside that produced within the statistical system.

Part of these features may be a matter of choice of audience. I have assumed that the desired audience consists of both Federal Statistical System staff and researchers in related fields from academia and commercial domains. The government, academic, and commercial research sectors have much to gain from learning about each others methods. The paper series could be enhanced by seeking input from the two other sectors. At the very least, this might entail a forced literature review within each paper; at a higher intensity this might involve the subcommittee membership of those outside the Federal system. Even the input from outsiders may not sufficient.

Figure 3

Topics of Statistical Policy Working Papers

Multiple Error Sources	3	- CPS Error Profile
	4	- Nonsampling Error Terms
	13	- Federal Longitudinal Surveys
	15	- Quality in Establishment Surveys
Coverage Error	17	- Coverage Error
Nonresponse Error		
Sampling Error		
Measurement Error: Interviewer		

Measurement Error: Questionnaire	10	- Developing Questionnaires
Measurement Error: Respondents		
Measurement Error: Mode of Data Collection	6	- Uses of Administrative Records
	12	- Telephone Data Collection
	19	- Computer Assisted Surveys
Processing	2	- Statistical Disclosure
	5	- Statistical Matching
	11	- Industry Coding Systems
	18	- Data Editing
Estimation	7	- Time Series Revision

Topics Not Classifiable Easily in Error/Quality Terms

Topical focus	1 - Statistics for Allocation of Funds
	16 - Reporting in Employer Data Systems
Administration	8 - Statistical Interagency Agreements
	9 - Contracting for Surveys

Other 14 - Uses of Microcomputers

Missing Topics of Statistical Policy Working Papers

Coverage Error	Problems using households as sampling frame elements
Nonresponse Error	Combining social science and statistical models of participation
Sampling Error	Statistical software for estimation; generalized variance models; alternative estimators for public use files
Measurement Error:	Training; variance models;

Interviewer	reinterview programs; monitoring of telephone interviewers
Measurement Error. Questionnaire	Developmental methods in cognitive laboratories; pretesting regimens; imbedding experiments in surveys
Measurement Error: Mode of Data Collection	Mail and self-administered surveys; mixed mode surveys
Processing	Statistical quality control; automated coding
Estimation	Model-based Estimation

2.2. OMB Series as Cross-Fertilization Among Federal Statistical Agencies

In my fifteen years of working with Federal statistical agencies from my academic base, I was consistently reminded of the relative isolation of individual agencies from each other. As most people in this room know, it is not uncommon for very similar lines of research and development to be pursued without much coordination across agencies. The arguments for this are that different problems faced by the agencies demand different solutions. The arguments against are that functionally equivalent solutions are often created by two different agencies at twice the cost.

The working paper series has had, I believe, a beneficial unanticipated effect at reduction on interagency duplication. First, the subcommittees consist of members from several different agencies. Second, the tasks of the subcommittees often involve collecting information from many statistical agencies. The members thereby learn of work going on in agencies they normally don't

visit. Third, recommendations of the papers often seek to apply standards across agencies, and the committees are forced to face the difficulty of system wide standards.

This is laudable and necessary. Is it sufficient? Clearly not. That is, working subcommittees of the Federal Committee on Statistical Methodology are temporary, normally have an agenda limited to the report, and do not generally follow up on logical conclusions of the report. Our dispersed statistical system, with all the benefits that specialization offers, misses opportunities to implement recommendations of these working papers.

2.3. OMB Series as a Prod to New Developments

Several of the papers treat topics where only one or two agencies are making major contributions and most others fall behind. For example, the Time Series Revision paper, the industry coding paper, the paper on computer assisted surveys, all fall into this category.

If I can temporarily put on the hat of an OMB staff member, this perspective seems to be the most central to the goals of the group. If reports like this can serve to improve the quality of work ongoing in several agencies, investments by one agency might quickly reap benefits in many agencies.

Some of the reports are poised for such effects, but the statistical system seems to miss more opportunities than necessary. Interagency agreements can be forged to promote such technology transfer. That is, consultation or subcontracting can be obtained within existing regulations. However,, this requires the target agency to acknowledge the need for such upgrading. Could OMB

facilitate this process? I am too naive to know, but the existence of a pool of funds at the OMB staff level to assure the spread of innovation across agencies through detail of staff and other mechanisms would be productive.

Are there areas of innovation that can profit from coordination? Certainly. The use of CATI/CAPI is one that comes to mind quickly. It is now an area in which separate expenditures are being made by several agencies, where no standards have been well-defined, where different solutions, with essentially that same cost/benefit structure, may evolve across different agencies.

The prod to new developments, however, demands that the papers end with a series of recommendations. The authors should stimulate the readers, dare I say, challenge the readers, toward improving current practice. After the detailed investigation needed for these reports, they are uniquely qualified to offer such recommendations. Only a minority of the reports end with such recommendations. This should be part of the charge to each committee.

3. The Need for a Structure of the Working Paper Series

As I age, I must admit that I find more appeal in structures that guide our research and development in survey design and

implementation, as opposed to reacting to each new idea without an explicit framework. In the academic world major theories provide that structure; they help to identify what are the important questions; they guide the development of new ideas. The application of the word "theory" to social and economic data production is rare. We do work that is guided by statistical theories, social science theories, organizational theories, and computer science theories. We are, however, basically on the applied side of research and development. We have a data collection and estimation vehicle (e.g., a survey) which is used for many substantive purposes. We are interested in knowledge that improves the vehicle and less interested in anything else.

As I understand the Federal Committee on Statistical Methodology, the topics for papers are essentially the fruit of discussions of the committee members. This is fine for assuring interest in the paper series among subcommittee members, but fails to assure coverage of important topics. I have suggested a total survey error structure above. The reports should have both measurement and reduction of error in mind. The widely perceived

worth of sampling error as a criterion of evaluation of data owes its existence largely to well accepted estimators of the error. We currently lack comparably well accepted measures for nonsampling errors, but the report series could be used as a vehicle to stimulate such measures.

Finally, another way to structure the report series is around major problems facing the Federal statistical system in the near and far term (see Figure 4). These, in my view, should form the core attention of the working paper series. The first I mention may be the most controversial. The statistical literature on survey design is schizophrenic on costs. On one hand, there exist models which demonstrate that only through knowing cost components can design optimization be achieved. On the other hand, there is little serious treatment of survey costs by statisticians or those from other disciplines.

Figure 4

Likely Problems Facing Federal Data in the Near/Far Term

1. Identification of cost components associated with error-related design features

2. Integration of question changes motivated by cognitive research into ongoing surveys

3. Public cooperation with data collection requests and coverage of subpopulations on sampling frames

4. Development of mixed strategy designs, tailored to diverse subpopulations

5. Development of nonsampling error indicators; implementation of statistical quality control procedures

6 Training of statisticians and social scientists in survey research; recruitment/retention of trained staff

The second issue has both a restrictive and more global

meaning. First, the work ongoing in so-called cognitive laboratories is seeking to identify principles influential of measurement error in question-answer sequences. The Federal statistical system at the current time has no good mechanism for the orderly introduction of change in questionnaires. For the vast majority of ongoing surveys, questionnaires remain static despite evidence of improved alterative measures. The value of unbroken time series and the assumptions of canceling biases in over-time comparisons are used to justify inactivity. Americans have very interesting reactions when they visit Cuba or see scenes of the country. They marvel at the maintenance of U.S. manufactured cars in their original state from the 1950's. They are at once proud of the ongoing use of older vehicles and humored by the lack of progress. A U.S. auto manufacturer would quickly go out of business if he were continuing to market 1950's designs. Indeed,

the watchword in that industry in continued investment in change, designing systems to permit ongoing change, making change part of the design. Survey researchers are driving 1950's vehicles in the

1990's. What we dearly lack is the will to mount ongoing programs of ongoing improvement in data series.

The third likely issue of import is the role of voluntary participation in surveys over the coming years. Some countries in Western Europe have experienced political shocks to response rates (e.g., Sweden, West Germany) . Public debate about surveys in these countries has led to lower cooperation with survey requests. In some cases documented effects on survey statistics exist. That is, the nonresponse error becomes visible to even the most naive reader of statistics. At this point, there was little the researchers were prepared to do in terms of reaction of field interviewers or construction of adjustment schemes. We must acknowledge that public cooperation is a fragile base on which the scaffolding of inference lies. To improve participation or to adjust inference in the presence of lower participation, understanding of the decision to participate must be obtained. This is an issue that faces the entire statistical system, indeed, the entire industry of information collection.

The fourth issue is not unrelated to the problems of participation. As the diversity of the U.S. population increases, survey designs that tailor procedures to different subpopulations grow. Large portions of the population remain covered by traditional frames, cooperative and competent to provide information using cheap data collection methods. Others fail to be covered on traditional frames, have difficulty providing information, and fear harmful consequences from their participation. The coming years are likely to find greater appeal in mixed design strategies -- multiple frames, multiple data collection modes, tailored questionnaires to subpopulations. The models exist in the survey design literature, but they need careful attention.

The final problem listed above concerns a crisis looming ahead for the social measurement industry in this country. Like all endeavors that require quantitative literacy social and economic statistics are currently facing a shortage of qualified personnel. If this were not bad enough, we also suffer from a worse problem -- the absence of ongoing training programs. It's not merely that

students aren't entering the field; it's not clear how they can within traditional academic programs. Let's examine the problem. Sampling statistics was well developed by the early 1950's; it is not a "hot" area of development, attracting the best and brightest of students. Instead, a variety of analytic statistical developments are more emergent. Young Ph.D.'s labelling themselves as sampling statisticians are unlikely to have an easy route to tenure in an academic department. Within the social sciences the difficulties might be greater, with great pressure on students to

develop areas of expertise which are central to the dominant paradigms in the discipline. Survey methodology is not one of them in any discipline. There are two results of this: 1) a gross inadequacy of training of new staff coming into the statistical system in topics relevant to survey quality. (This is not a comment on their training as statisticians, psychologists, or economists.) and 2) a reduction in the number, of academic researchers devoted to the craft of social measurement. There is

a clear conclusion here: the statistical system has to get serious about training of staff it needs for the future. This means support of specialized graduate programs, focused continuing education, onsite training and other similar mechanisms.

The two types of structure - quality/cost components of data series and problems facing the system - suggest two paper series, one devoted to technical issues, another to administrative and professional issues.

4. Other Comments, Not Elsewhere Classified

I must admit confusion about the term, "working paper series." In an academic setting this term is used to describe papers in the process of being refined or papers not worthy of being refined. People are sometimes "working" on them. The better ones change over time, they evolve to a better state. This doesn't seem to fit well with the OMB Working Paper Series. Most all remain in their original state.

I don't want to change the name of the series; I'd rather see the series periodically updated. Several of the papers were valuable only for a short period of time (e.g., microcomputers; telephone data collection). Having a well-defined structure to the series might define a set of ongoing updates of papers devoted to individual topics.

There in another connotation of "working" when attached to paper series. That is, they are "working" toward quality improvements in the statistical system. I like this connotation. But it implies two burdens not uniformly accepted: a) a set of recommendations at the end of reports, b) follow through by OMB or individual agencies to implement change. On this definition, I think, the paper series has not achieved full success.

Another problem with the series are the costs and benefits assigned to authors of the reports. Contrary to my colleagues in academia, statistical system staff rarely experience career-enhancing effects of writing such papers. There is the value of education about other agencies, of "networking" with other members

of the statistical system, and of learning more about important issues facing the system. On the other hand, I've learned that this is work essentially performed at nights and weekends by people

already very busy. Now, night and weekend work is commonly very productive and I have no problem with such a plan. What I do regret (and think it bad for the health of the system) is that such work is given so little value by many of the home agencies. OMB might consider remedying this with some more formal recognition of the writers of these reports. At the very least, the authors of the report might be given a more prominent position on the covers of the papers.

It strikes me that this seminar is an ideal forum for generating discussion on the future of this series. I recommend several questions:

Have the basic issues changed since the report?

- because of the paper?

- in spite of the paper?

Is it time to redo the paper, to update it?

Are there subtopics now of sufficient importance that they deserve separate treatment?

5. Personal note

This working paper series consistently contains the name of one person, from the first to the last - Maria Gonzalez. The Federal Statistical System often focuses its attention on data series structures and organizations, not people, but the success of any endeavor that spans decades depends on key people. In this paper series the key person is unambiguously Maria. As those of you who know her well can attest, she has been a rock of rationality, courtesy, integrity, and absolute honesty in her work on the Federal Committee on Statistical Methodology. She alone can

succeed in pressing overworked federal statisticians to take on projects for the benefit of the whole system. Her near unique ability to suggest ideas in a manner that allows the hearers to believe they are their own ideas is a marvel. Her perseverance toward important goals of quality improvement and coordination have made the working paper series and this conference possible.

Session 1

SURVEY QUALITY PROFILES

THE SIPP QUALITY PROFILE

Thomas B. Jabine

Statistical Consultant

A. Introduction

The Survey of Income and Program Participation (SIPP) is a longitudinal national household survey which has been conducted by

the U.S. Bureau of the Census since 1983, following several years of developmental research. The goal of the survey, which uses a rotating panel design, is to provide policy makers with comprehensive and accurate data about the levels and determinants of the income of U.S. persons and households and about their participation in a broad range of income transfer and welfare programs.

The SIPP quality profile summarizes current knowledge about the sources and magnitude of errors based on SIPP. An initial version of a SIPP quality profile was issued in 1987 (U.S. Bureau of the Census, 1987) and an updated and expanded version was prepared in 1989 (U.S. Bureau of the Census, 1990).

This paper describes the purposes of developing a quality profile for a survey or other statistical program and the process of preparing and updating a quality profile, using the SIPP Quality Profile as an illustration. The contents of the updated version will be discussed briefly. Those who wish to evaluate the quality of SIPP data on specific topics or to develop an overall judgement

about the quality of SIPP data are referred to the latest version of the SIPP Quality Profile and the other sources of information that it identifies.

Section B outlines the development of the quality profile concept and identifies some publications of the last 4 decades that could be regarded as forerunners of the current model. Section C explains the origin of the SIPP Quality Profile. Section D provides an overview of the updated version: its intended audiences, purposes sources of information and structure. The contents are discussed briefly in section E. In the concluding section, I discuss the role of a quality profile in the broad context of survey quality control and improvement.

B. Some Forerunners of the Quality Profile

The theoretical foundation for a quality profile rests on various models that have been developed for the measurement and analysis of errors in surveys, especially the Census Bureau model, which integrates components of sampling and nonsampling error and the interactions between them (Hansen, Hurwitz and Bershad, 1959). Dalenius (1974) formalized the concept of total survey design, using the Census Bureau model to guide the allocation of resources to minimize total error in a survey.

Based on this foundation, there have been several broad qualitative and quantitative reviews of the quality of data from censuses and surveys, featuring direct and indirect data about the various components of error. Zarkovich (1966) published what was perhaps the first systematic treatment of nonsampling errors in surveys, with emphasis on procedures for their measurement and control, and including numerous examples of specific information about nonsampling errors from surveys and censuses in many countries. Bailar and Lanphier (1978), in a pilot test of methodology for the evaluation of survey practices, reviewed the quality-related design features of 36 U.S. surveys. Their review

was not based on direct measures of errors, but the frequency with which they found indirect evidence of low quality was high enough to be disturbing and to suggest a need for greater attention to the quality of survey designs and practices.

A United Nations (1982) manual on Nonsampling Errors in Household Surveys, prepared for use in developing countries, systematically explores the different sources and types of nonsampling error and provides illustrative data from numerous household surveys throughout the world. Statistical Policy Working Paper 15 (Office of Management and Budget, 1988) performs a similar function for Federally sponsored establishment surveys in this country.

Compilations of information about the quality of surveys have two main audiences: survey designers/managers and users of survey data. To ensure that the latter have access to such information, standards have been developed for the dissemination, in survey publications, of information about errors. An early example of such standards was Census Bureau Technical Paper 32 (1974). Today,

several Federal statistical agencies apply similar standards in their publication programs.

There have been some publications devoted entirely to the quality of data on a specific topic in a census or survey. An early example was a detailed appraisal of the income data from the 1950 Census of Population (Conference on Research in Income and Wealth, 1958). The most immediate forerunner of the SIPP Quality Profile was Statistical Policy Working Paper 3 (Brooks and Bailar, 1978), which provided an error profile for estimates of

unemployment from the Current Population Survey (CPS) Jabine (1987) provided a detailed analysis of the quality of data on chronic conditions reported in the National Health Interview Survey.

There are two fairly evident differences between the CPS error profile and the SIPP quality profile. The most obvious is the

switch from "error" to "quality" as the defining adjective for the profile's content. While this may seem to be only a semantic change, it reflects a feeling, undoubtedly shared by the authors of the CPS error profile, that the goals of such a publication are constructive. The use of the term quality seems more in keeping with today's emphasis on quality control and improvement in all kinds of endeavors, including surveys. The other basic difference is that the SIPP quality profile covers the quality of estimates for all of the topics included in SIPP, whereas the CPS error profile covered only one of the many topics included in that survey.

Other U.S. statistical agencies are undertaking similar although not identical efforts. The Energy Information Administration, for example, periodically publishes reports in a series called An Assessment of the Quality of Selected EIA Data series. These reports rely largely on the technique of comparing data from EIA surveys with more or less comparable data from other sources and analyzing the differences that are observed. Janet Norwood, in a paper presented at the Census Bureau's Third Annual Research

Conference, stated that the Bureau of Labor Statistics was planning to develop a comprehensive error profile for each of its surveys (Norwood, 1987, pp. 217-218).

C. Origin of the SIPP Quality Profile

The SIPP is a major longitudinal survey. The start of the survey was preceded by several years of research and development, an effort known as the Income Survey Development Program. The evolution of SIPP's complex survey design did not end when the survey became operational late in 1983. Methodological research and evaluation studies have continued at a substantial pace and the results of these studies, along with accumulated performance statistics, feedback from users and adjustments made necessary by reductions in funding, have led to significant changes in the survey design and procedures. Thus, SIPP is still in the early stages of its evolution, in contrast to the Current Population Survey which, although not immune to evaluation and improvement, has reached a more mature and stable phase.

In l984 the Social Science Research Council and the Survey

Research Methods Section of the American Statistical Association, with the encouragement and support of the Census Bureau, established a Working Group on the Technical Aspects of SIPP to provide

advice to the Census Bureau on research priorities and the translation of research findings into changes in the survey design and procedures. (The Social Science Research Council later relinquished its sponsorship role.) An early recommendation of the Working Group was that the Census Bureau prepare a compendium of research results and other information about the quality of SIPP data. Members of the Working Group believed that a systematic account of information about the different kinds of errors that affect estimates from SIPP would be invaluable as a guide in setting research priorities and applying the principles of total survey design to SIPP. Given the substantial amount of ongoing research, they recommended that such a quality profile be updated periodically, perhaps every two years.

The Census Bureau accepted the Working Group recommendation and produced the Quality Profile for the Survey of Income and Participation (King, Petroni and Singh, 1987), early drafts of which were reviewed by several members of the Working Group. New information continued to flow in at a rapid rate and toward the end of 1988, Census decided that it was time to start work on an update. The updated version, published in mid-1990, was prepared by the author of this paper with substantial assistance from Karen King and Rita Petroni of the Census Bureau's Statistical Methods Division. Although the general structure of the two versions is similar, the update contains much new material and some of the earlier sections were significantly revised. It also includes an index. The new version benefitted from reviews by several members of the SIPP Working Group and Census staff. Special thanks are due to Daniel Kasprzyk and Rajendra Singh for their support of the project.

D. Overview of Version 2

The SIPP Quality Profile is intended to serve two main audiences: "users of SIPP data and those who are responsible for or have an interest in the SIPP design and methodology." The interests of these two groups are different. Users want to know how the errors associated with specific categories or classes of data are likely to affect their analyses. SIPP designers and managers need to know the magnitude of errors associated with specific design features, in order to control the quality of the survey estimates and to guide the allocation of resources available for their improvement. Besides these two primary audiences, it was expected that the publication would be of interest to persons concerned with the design of longitudinal surveys other than SIPP and to two special groups: the ASA/SRM Working Group and a Panel to Evaluate the Survey of Income and Participation, convened by the Committee on National Statistics at the request of the Census Bureau.

Information about the components of error that affect SIPP data comes from four sources:

- Performance statistics, such as unit and item non-response rates and reports based on quality control procedures used in data collection and processing operations.

- Methodological experiments. Both in the developmental period and since the start of survey operations, there have been numerous methodological experiments involving design features such as length of questionnaire, respondent rules, use of respondent incentives, increased use of telephone interviewing and methods of adjustment for nonresponse.

- Micro-evaluation studies. The outstanding example is the SIPP Record Check Study, in which individual survey responses to questions about program participation and benefits were compared with administrative data for each

of several programs.

o Macro-evaluation studies. There have been numerous comparisons of SIPP data with data on the same topics from other surveys, especially the Current Population Survey, and from program records.

Assembling the relevant documentation was a challenge. SIPP has probably generated more methodological documentation than any other survey that has been in existence for a similar length of time. The list of 161 references provided in the updated version of the Quality Profile, which includes only those items that were actually cited in the report, is nearly double the size of the list, included in the first version. The most commonly used sources were: the SIPP Working Paper series; the annual proceedings of the Survey Research Methods, Social Statistics and Business and Economic Statistics sections of the American Statistical Association; the proceedings of the Census Bureau's Annual Research Conferences; and internal Census Bureau memoranda. The report informs readers how to obtain copies of any of the internal

memoranda in which they are interested.

Finding a suitable framework in which to present all of this information about different components of error also presented a challenge. The traditional approach is to organize the material according to the main phases of the survey: sample selection, data collection, data processing and estimation. The core of the Quality Profile (Chapters 3 through 8) is, in fact, organized in that manner, with one chapter devoted to sample selection, three to data collection (covering data collection procedures, nonresponse error and measurement error) and one each to data processing and estimation.

Two important topics did not fit neatly within this framework. Chapter 9, Sampling Errors, covers the procedures used to estimate sampling errors and the relationship between sampling errors and sample size. Chapter 10, one of the longer chapters, is called "Evaluation of Estimates" and covers both comparisons Of SIPP

estimates with data from other sources and indicators of errors of undercoverage. The remaining chapters, 1, 2 and 11, provide an introduction, an overview of the survey and a summary, respectively.

The structure of the SIPP Quality Profile is similar to that of its chief forerunner, the CPS Error Profile. The main differences are the division of the material on data collection (called "Observational Design and Implementation" in the CPS Error Profile) into three chapters, and the addition of the chapters on sampling errors and evaluation of estimates.

Our goal was to provide, insofar as available, quantitative information about overall error and its components. Hence, the report includes 6 figures and 43 tables, a substantial increase over the number included in the first version. Space limitations preclude inclusion of tables in this paper, but for those who may be interested, the numbers of some key tables and figures from the publication ate given in the following section.

E. Summary of Findings

Major sources of error

The SIPP Quality Profile does not contain any broad conclusions about how successful SIPP has been so far in fulfilling its goals. Our goal was to provide enough information about the quality of the survey data so that individuals and groups like the Committee on National Statistics Panel to Evaluate SIPP could reach their own conclusions. The summary chapter does, however, identify what stood out as the three main sources of error in SIPP estimates: nonresponse, differential undercoverage and measurement error.

As in any longitudinal survey, unit nonresponse increases in succeeding rounds (called "waves" in SIPP) of the survey. Table 5.1 (not included with this paper, see the report) shows the data available as of 1989 on unit nonresponse by wave for each panel of the survey (households and individuals in each panel are interviewed 8 or 9 times, at 4-month intervals). The rates are

relatively low -- 4.9 to 7.6 percent -- for the first wave, but increase to over 20 percent at the final wave of each panel. This relatively high attrition is due in part to the difficulty of tracking households and individuals that move, as is required by the SIPP design. The characteristics associated with Unit nonresponse have been analyzed in detail, and these analyses have

guided the development of estimation procedures designed to minimize the biases that result from differences between the characteristics of respondents and nonrespondents.

Item nonresponse has been low for core items on labor force activity, income recipiency and asset ownership. It has been somewhat higher for income amounts, especially self employment earnings and interest. In the topical modules (questions not asked in every wave), especially high nonresponse has occurred for questions on asset amounts.

Indicators of differential undercoverage in SIPP for population subgroups defined by age, race and sex are shown in Table 10.13 of the report. The table shows the reciprocals of the weights that are applied in order to make the simple unbiased estimate for each subgroup agree with an independent estimate that uses the Population Census count as a benchmark. The group most affected is young adult black males. The ratios for black females in the same age group are also quite low. At least for the males, the coverage ratios shown understate the amount of undercoverage, because the ratios do not include any adjustment for census undercoverage which is known to be above average for this population subgroup.

Similar patterns of undercoverage have been observed in the Current Population Survey and other national household surveys. The second-stage ratio adjustments used for both cross-sectional and longitudinal estimates to compensate for undercoverage are believed to reduce both the sampling error and bias of the estimates. The effects of these adjustments on sampling errors can be estimated, but little is known about their affects on biases

associated with undercoverage.

Measurement error takes many forms, but perhaps its most significant manifestation in SIPP has been the seam problem, i.e., a pronounced tendency for survey respondents to report month-to-month changes for months in adjacent waves at substantially higher rates than for adjacent months within a single wave. Figure 6.1 in the report provides a graphic illustration of the seam effect on reports of changes in earnings. Pronounced effects have been noted for most income recipiency and amount variables. Because of the rotation group design used in SIPP, cross-sectional estimates of transitions are not likely to be seriously distorted by this pattern of reporting, but it can affect estimates of the covariance structure and may have adverse effects on multivariate analyses dealing with transitions or length of spells.

Table 6.6 in the report shows some early results from the SIPP Record Check Study. The sample sizes are small, and the table shows results for only two of the four states included in the study. For the State of Wisconsin, significant levels of

underreporting were found for participation in two programs and

benefit amounts in one other program. The full results from the Record Check Study will provide the best direct information so far available on levels of measurement error in SIPP and will be a valuable resource for studying the sources and correlates Of response bias and response error variance.

Current research

An active program of SIPP methodological and evaluation research is continuing. The main areas of research include:

o The design of the questionnaires and the structure of the interviews. Laboratory research is being conducted to study the cognitive aspects of SIPP interviews and how they relate to seam effects and other kinds of reporting errors. Field experiments have been conducted to test the feasibility of providing feedback of prior wave

information and encouraging greater use of records in interviews.

- Interview mode. An experiment with increased use of telephone interviewing is being evaluated to determine whether to adopt the procedures that were tested. For the longer term the Census Bureau is arranging for the development of a prototype questionnaire for use in computer-assisted personal interviewing (CAPI), in order to evaluate the potential effectiveness of this collection mode in SIPP.

- Estimation procedures. The broad goal for this area of investigation is to develop estimation procedures for SIPP that make effective use of auxiliary data available from both the Current Population Survey and administrative records. An initial study of the feasibility of reducing variances by using IRS data as controls in the second-stage ratio estimation procedure showed considerable promise.

Research in these and other aspects of the survey is proceeding at a pace that suggests the desirability of preparing updates of the SIPP Quality Profile on a regular basis.

Areas of research that have been relatively untouched so far include the effects of interviewer variance and the conditioning, effects of repeated interviews on response error. For the latter, the overlapping panel design used in SIPP offers the possibility of comparing cross-sectional estimates for households and persons that have been in the sample for varying lengths of time. There is also a need to update some of the earlier evaluation studies in order to monitor the effects of design changes since the beginning of the survey. Much of the research reported in versions 1 and 2 of the

SIPP Quality Profile, including the Record Check Study, which is the only source of direct information on the site of individual reporting errors, is based on data from the 1984 panel.

F. Conclusions

Judging from some comments by users of the initial version and reviewers of the preliminary draft of the updated version of the SIPP Quality Profile, the systematic compilation and publication of information about the nature and sources of error in a major continuing survey like SIPP, with Periodic updates, is a worthwhile undertaking. A more definitive evaluation of utility will be possible now that the updated version has been published and is being widely distributed. The author believes that the preparation of quality profiles could be valuable in connection with efforts to track and improve the quality of data from other major continuing national surveys, such as the Current Population Survey, the National Health Interview Survey, the National Crime Survey, the Annual Survey of Manufactures and the Monthly Retail Trade Survey. The technique is applicable to both household and establishment surveys.

Maintaining and improving the quality of survey data is a

never-ending job for survey designers and managers, and there is room for a multiplicity of approaches. Some Federal agencies are making a strong commitment to the application, to survey operations, of Deming's philosophy and techniques for total quality management. That approach implies not just measurement of errors and identification of their sources, but modification of the survey process as needed to eliminate or reduce the effects of significant sources of error. The other paper presented at this session (Hanuschak, 1990) provides an example of this model of survey quality management, with active participation and commitment to quality improvement by key managers in the organization. The same commitment to the quality of data can be seen in the work of the sponsors and participants in this Conference and they deserve our thanks for it.

REFERENCES

Bailar, B. and Lanphier, M. (1978), Development of Survey Methods to Assess Survey Practices, Washington DC: American Statistical Association.

Brooks, C. and Bailar, B. (1978), An Error Profile: Employment as Measured by the Current Population Survey, Statistical Policy Working Paper 3, Office of Federal Statistical Policy and Standards, U.S. Department of Commerce.

Conference on Research in Income and Wealth (1958), An Appraisal of the 1950 Census Income Data, Studies in Income and Wealth, Vol.23, National Bureau of Economic Research, Princeton: Princeton University Press.

Dalenius, T. (1974), Ends and Means of Total Survey Design, Stockholm: University of Stockholm.

Energy Information Administration (1983), An Assessment of the Ouality of Principal Data Series of the Energy information Administration (first in a series of "state of the data" reports), Publication DOE/EIA-0292(82).

Hansen, M., Hurwitz, W. and Bershad, M. (1959), "Measurement Errors in Censuses and Surveys", Bulletin of the International Statistical Institute, 38:359-374.

Jabine, T. (1987), Reporting Chronic Conditions in the National Health Interview Survey: A Review of Findings From Evaluation Studies and Methodological Tests, Data From the National Health Survey, Series 2, No. 105, National Center for Health Statistics.

Jabine, T., assisted by King, K. and Petroni, R. (1990), Survey of Income and Program Participation: SIPP Quality Profile, Bureau of the Census, U.S. Department of Commerce.

King K., Petroni, R. and Singh, R. (1987), Quality Profile for the Survey of Income and Program Participation, SIPP Working Paper No. 8708, Bureau of the Census, U.S. Department of Commerce.

Norwood, J. (1987), "What is Quality?" in Proceedings, Third Annual Research Conference, Bureau of the Census, U.S. Department

of Commerce: 215-222.

Subcommittee on Measurement of Quality in Establishment Surveys (1988), Ouality in Establishment Surveys, Statistical Policy Working Paper 15, Statistical Policy Office, U.S. Office of Management and Budget.

United Nations (1982), Non-sampling Errors in Household Surveys: Sources, Assessment and Control, UN Publication DP/UN/UBT-81-041/2, National Household Survey Capability Programme.

U.S. Census Bureau (1974), Standard's for Discussion and Presentation of Errors in Data, Technical Paper 32, U.S. Department of Commerce.

Zarkovich, S. (1966), Quality of Statistical Data, Rome: Food and Agriculture organization of the United Nations.

INITIAL REPORT ON THE QUALITY OF AGRICULTURAL SURVEY PROGRAM

George A. Hanuschak

National Agricultural Statistics Service

I. Background and Introduction

In December 1988, the National Agricultural Statistics Service (NASS) formed a Survey Quality Team (SQT) for its Agricultural Survey Program (ASP). The ASP is a series of integrated multiple sampling frame (area and list) based surveys throughout the agricultural calendar year. Some major items on the surveys are planted and harvested crop acreages, hog, cattle and sheep inventories, crop yields and production and on-farm grain storage. There was a major survey redesign from individual MF surveys to an integrated multiple frame survey program which was implemented over

several years (1984 - 1986). The mission of the Survey Quality Team is to identify and develop statistical process control (SPC) methods for the management of the integrated Agricultural Survey Program. The SPC methods are based upon the fundamentals of total quality management (TQM) techniques developed by Edward Deming, Joseph Juran, Philip Crosby and other well-known TQM developers in the TQM and SPC literature. However, since much of the literature refers to "manufacturing" situations, it was adapted to fit the government agricultural survey situation. Several papers by Ron Fecso developed the basic model of survey quality used by the SQT. The first major milestone of the SQT was to be the development of a baseline "state of the survey" quality report.

The mission of the SQT is quite broad, challenging and critically important to the Agency's long term goal of routinely and continually improving survey quality. The team and the Agency also face this challenge in the light of severe budget pressure, in general, on Federal Statistics programs. However, the team feels that TQM and SPC methods are quite powerful tools, when properly applied, that can aid in measuring and improving survey quality

over time.

One of the first lessons of total process control is to define the major steps in the total process. In the case of the ASP, one needs to first define or identify the major steps or stages of the ASP surveys. The survey quality team had identified the following steps (Exhibit I) as the major 22 processes of the survey. Unfortunately, each one of these survey stages or processes is probably susceptible to some type of errors or biases. The SQT developed the following profile (Exhibit 11) of 24 potential sources of error or bias in the ASP.

Like any good statistical organization, the Agency has tried to minimize the probability of various nonsampling errors occurring

in the survey process. Controls include training, survey manuals and instructions, Agency Policy and Standards Memorandum, quality control checks on enumeration, reinterview studies, etc.

Controlling and measuring nonsampling errors for a complex survey process will remain extremely challenging even with the best efforts at statistical process control. However, in the remainder of this report, the SQT defines and demonstrates how to use statistical process control and total quality management techniques to reduce total survey error over time.

Exhibit I - Major Survey Stages

Survey Clearance

Area Sampling Frame

(Construction, Maintenance and Sampling)

List Sampling Frame

(Construction, Maintenance and Sampling)

Survey Specifications

Design of Questionnaires

(Design, Print and Distribution)

Preparation of Manuals

(Interviewers, Supervisory and Editing)

Prepare Survey Software

(Data Entry, Survey Coordinator, Edit, Analysis, Summary,

Data Base, Mail and Maintenance System, Etc.)

National/Regional Training Schools

Survey Management - Headquarters and State Statistical Offices

(Coordination of Procedures)

Presurvey Coding/Handling/Processing by State Statistical Offices

State Training Schools

Data Collection

Data Collection Quality Control

Manual Data Review and Coding

Data Entry and Validation

Data Edit and Review

Imputation, Analysis and Summarization

State Statistical Office Review of Survey Results

(including submission of estimates)

Headquarters Review and Release Preparation

Post Survey Updating

(Data Base and List Sampling Frame)

Post Survey Evaluations

Survey Research

Exhibit II - Some Potential Sources of Total Survey Error

in the Agricultural Survey Program

Undetected List Sampling Frame Duplication

List Sampling Frame (Old or Incorrect Control Data)

List - Undetected Reporting Duplication or other

reporting/enumeration errors or bias

List Sources of Questionable Quality used for List Sampling Frame

Build/Maintenance

Area Sampling Frame (Outdated Land Use Stratification)

List Sampling Frame (Any large operations not covered by the frame)

Area Sampling Frame (Outdated Sample Segment - Aerial Photography)

Different Farm Operation Description Questions on Different Questionnaire versions

Incorrect overlap/nonoverlap Determination

Incorrect Exception Report Handling (One Type of Survey Weighting Factor)

Incorrect Coding (List Adjustment Survey Weighting Factors, Completion/Imputation Codes, etc.)

Undetected Data Entry errors (pass all the way through the editing system)

Shift in Mix of Data Collection Modes (Telephone, Computer Assisted Telephone, Mail and Personal)

Shift in Mix of Respondents (Operator vs. Spouse vs. Other)

Incorrect Survey Master Records

Questionnaire Design (or Print) Errors

Unmeasured Major Changes in Survey or Estimation Procedures

(Headquarters or State Statistical Offices)

Error in Known Zero Determination (Is Respondent Validly out of Business?)

Overediting/Underediting of Survey Data

Potential Bias in Manual or Machine "Imputation" Procedures

Lack of Formal Outlier Handling Procedures (Non Robust or Non Smooth Time Series Estimation)

Survey Processing Software

Shifts in Characteristics or Skill Level of Work Force {(Enumerators, Statisticians, Programmers, Support Staff) Experience in their current job, survey procedures knowledge, farm knowledge, statistics knowledge, technology skills, etc.}

Farmer or Respondent's level of understanding or grasping of survey reporting concepts and item definitions (Cognitive aspects).

II. The Components of Survey Quality

When faced with the problem of measuring and improving the quality of the ASP, one should consider the components of survey quality. Listing the components defines exactly what is meant by the-term "survey quality" and highlights specific sub-areas that need to be explored.

Figure 1 shows the components of survey quality. It was developed by the Nonsampling Errors Research Section in the Survey Research Branch of NASS and adopted by the SQT. There are four major components related to survey quality accuracy, resources, timeliness, and relevance.

Accuracy is the component that first comes to mind when thinking about survey quality. NASS wants the survey indications to be as accurate as possible. Not only should the sampling errors be small, but also the nonsampling errors should be minimized. In large-scale surveys the relative sampling errors can be smaller than the relative size of the nonsampling errors. Factors such as undetected list sampling frame duplication, nonresponse, questionnaire wording, mode of interview, change in respondent, etc., can lead to substantial nonsampling errors.

The second component of survey quality is resources. Even if a survey organization can control the sampling and nonsampling errors, its ability to do so will be affected by the amount of

dollars that are available to spend on the survey. The amount of dollars has a direct impact on sample sizes, list frame quality, pretesting, reinterview projects, editing programs, summary programs, analysis, etc. Also important is the amount and quality

of staff hours that can be devoted to a survey. Staff hours are affected by salaries, training, hiring practices, long-term career development, and organizational climate; components that are also greatly affected by the amount of dollars available. Most people quickly realize that the crucial problem is to take the fixed set of available resources and use those resources in a way that maximizes the survey quality.

The third component is timeliness. Of course, time could be considered another element of resources -- like dollars and staff. However, timeliness needs to be considered a component by itself because timeliness is crucial in the survey process. The impact and usefulness of survey indications are greatly affected by whether the survey data were collected one month or one year

earlier. NASS has always stressed the need to collect data quickly and to release estimates as close to the survey reference date as possible. Thus, the survey calendar -- which is used to time all the steps of the survey -- is important to the survey quality.

The final component is relevance. Relevance is dependent on the needs of the users of NASS statistics, and those needs change from day to day. It is useless for NASS to collect a high-quality piece of information on farming if that piece of information has no relevance for the users of NASS statistics -- that piece of information simply becomes a product without a buyer. NASS must constantly assess the needs of people using its statistics to make sure that the collected information is relevant. The second aspect of relevance is internal to NASS. An example of internal relevance is whether the Agency wants direct expansion (level) or ratio (percent change) or both types of estimators out of the ASP.

III. Accuracy of Survey Soybean Acreage Estimates

NASS has an expert panel of Agency statisticians called the Agricultural Statistics Board. (ASB) which reviews all survey indications (often multiple indications for any one item), and administrative or check data (such as the amount of soybeans crushed in processing plants) and adopts or sets the official estimates to be published.

Two concepts need to be defined - use and fitness. The ASB's use of the ASP indications was chosen as the primary "use" of the ASP. "Fitness" for use is evaluated by setting a standard for use and measuring adherence to the standard.

Ideally we would have standards for all the components of mean squared error (MSE) for the various commodity indications and administrative data used by the ASB. This would provide the ability to create statistically well defined composites of the data for use as the Board estimate or forecast. As this time we have measures of the variance for most indications, but have only enough

information about MSE's to recognize the importance of developing more extensive MSE measures. This section will provide information for Agency management to assess which areas are most in need of further study or research and/or corrective action.

The ASB's specific need is to have indications which serve as a solid basis for the official numbers. The following chart on soybean planted acreage display the degree to which the ASB has found the ASP indications to be "fit for use."

In reviewing the soybean planted acreage chart on ASB use you will observe the following:

1. The Agricultural Statistics Board finds the area sampling frame based June acreage estimate quite "fit for use."

2. The ASB does not find the integrated multiple frame based June acreage estimate "fit for use." It has an observed substantial upward bias which also changed substantially in magnitude between

1987 and 1988 and stayed at the larger magnitude in 1989 and 1990.

Using Pareto analysis and an expert panel using TQM principles applied to surveys, the SQT identified the major suspected causes of the upward bias in the multiple frame based soybean acreage estimate. These suspected causes are:

1. Different Data Collection Methodologies

The area frame based acreage estimate is based upon a sample of about 16,000 sample segments throughout the U.S. Data collection is done completely by personal interviews using an aerial photograph to locate each crop field and recorded on a questionnaire by the interviewer with the farmers direct participation. Crop acreage data is collected and edited field by field. Farmers are probed to report waste acreage for each field. There are also five specific questions related to defining land operated now to which all the rest of the questions relate to.

On the integrated multiple frame survey, the majority of data collection is done by telephone (both conventional and computer assisted). The crop acreage data is collected for the entire farm (not field by field). Therefore farmers are probed for waste acreage only once, at best, when reporting crop acreage. There is no photographic aid for the farmer to refer to. There is only one or two questions on defining land operated now.

2. Undetected List Sampling Frame Duplication

There are sophisticated record linkage tools to identify and remove duplication on the list sampling frame. However, due to clerical resource constraints and funding to call farmers to resolve differences and the use of multiple list sources some duplication remains. A special study was designed in 1989 to measure remaining duplication and the effect on the estimates. The study showed that approximately 10 percent of the acreage difference was due to obvious list frame duplication.

3. No Formal Documented Outlier Handling Procedurers

While there are several good analysis tools to identify outliers, there is no formal procedure for handling them. The area frame based acreage estimator is quite robust since the average expansion factor is about 200 and the segment size is 640 acres putting an upper bound on "influential observations". For the list sample, expansion factors are considerably larger and farm size does not have much of an

upper bound. Thus it is much easier to get highly influential observations in the list sample. Development of a formal robust estimator for the list sample is highly recommended.

4. Different Imputation Methodologies

There are also different imputation methodologies. All imputation for the area frame is done manually by interviewers observations or statisticians. In the case of crop acreage if a farmer refuses the interviewer can still observe most of the crop fields and the crop. On the list sample, the imputation is a computerized algorithm that uses other reported survey data and list frame control data to impute for nonreported data cells.

5. Undetected Reporting Errors

Since the questionnaire design is different the undetected reporting error structure may also be different. For example, the screening questions on land operated on the area side are more detailed than the list questionnaire and may do a more accurate job of screening out landlords who are not active farmers at survey time. New farm programs may have also led to the formation of more complex farming operations, which may involve a different reporting error structure also.

6. Different Ratio Type Information and Sample Designs

On the area frame sample there is an 80 percent overlap from one year to the next. On the list frame sample (independent from year to year) there is negligible overlap. Thus the area frame sample also provides a paired sample ratio estimator.

It is important to note that there have also been two rather

independent sources of data available to the ASB which also support following the area frame level. These are a Landsat satellite based regression estimator (1980-1987) which for major soybean states had variances at least twice as small as the direct expansion estimator but also were unbiased when compared to the ASB and direct expansion. The second source is the calculation of a soybean balance sheet which the ASB uses as an evaluation tool. A balance sheet takes the carryover from one crop year to the next and adds crop production to that and then subtracts crop utilization including exports from it to get a current balance. These balance sheets also support the area frame based crop acreage level. Thus the agency has attempted to verify the correct crop acreage level using several methods and independent data sources.

Even though there is an observed upward bias in the integrated multiple frame estimator for soybean acreage there are reasons for keeping it and reducing the bias. These reasons are:

1. Later crop season yield and production estimates are tied to the integrated multiple frame (IMF) approach.

2. State and sub-state level estimates from the IMF have much better precision than the corresponding area frame estimates.

3. Solving the bias problem associated with soybean acreage may well improve the entire IMF which is a survey 6 times a year with an average of 20-40 items (multivariate in nature). The Survey Quality Team has performed similar analysis for on-farm grain storage, and cattle and hog inventories. Some of the bias issues are item specific but others are associated with the total survey process or components of the survey process.

4. The IMF approach is substantially more cost efficient and involves less respondent burden than the area frame approach.

Most important is that the Agency is taking actions on all of these expected causes in 1989 and 1990. As previously mentioned there is now an improved list frame duplication adjustment procedure in place starting in June 1989. There is a reinterview research study being conducted in June 1990 to provide initial measures of previously undetected reporting errors. This study will involve the reinterviewing of a subsample of the list sample of farmers and record the crop data field by field and ask the more detailed land operated questions and compare the results. There are also research efforts underway to examine the imputation methodologies and to look at an across year design for list frame based estimators and evaluate several robust estimators. In addition the SQT has provided several quality measures to be monitored on the resource, relevance, timeliness and accuracy dimensions which should become operational in 1990-91.

The Agency is also developing alternative "proxies" to the true item values in addition to relying on the ASB process. An operational reinterview/reconciliation survey is being conducted in six major grain producing states in December 1990. There has also

been an extensive operational soybean yield validation survey (198?
- current) where farmers are asked to harvest specific fields and
take just that grain to a grain elevator to be weighed and
measured.

This "proxies" to true values are important in a survey evaluation program but are also complex and expensive to develop and implement.

As previously mentioned, use of earth resource satellite data has also been used by the Agency to develop more precise and accurate crop acreage estimates.

IV. Summary

It is the claim of the SQT that more consistent and timely process improvements can take place by using the principles of statistical process control and Total Quality Management. More

formal survey quality measurement and monitoring mechanisms will provide the Agency's management with more and critically important information to manage the quality of the ASP. Also, most of these techniques will readily transfer to other survey programs in the Agency such as Prices Paid and Received by Farmers, the Farm Costs and Returns Survey, Objective Yield Surveys, Farm Labor Surveys, and even to new programs such as Water Quality and Food Safety Surveys, the National Animal Health Monitoring System and the Monthly Yield Survey Program.

There are several tools available for such a survey quality management system. First there are numerous charting techniques such as bar and pie charts for resource information, Board standardized indication graphs with standard errors, Gantt charts to display, project management and survey schedule information, upper limit and lower limit control charts, multivariate control charts, Ishikawa fishbone diagrams and Pareto charts and analysis. Many of these were used in an earlier effort by the Nonsampling Errors Research Section when a statistical process control study was conducted on the Soybean Objective Yield Program.

Pareto analysis is one of the most powerful tools in quality monitoring systems. Pareto analysis ranks the potential errors in a system from most serious to least serious. The reasoning is that in many systems and not just surveys, there are a "vital few" and "trivial many" potential errors in the system. Thus, the most important beginning of evaluating the quality of a system is to identify where it is most likely to break down or fail. Once the ranking of potential errors is accomplished, then it is recommended to identify the allocation of resources for each potential error to see if management is allocating resources in a fashion that will truly minimize total survey error. Many Pareto analyses have demonstrated that the resource allocation was not in proper alignment with the true error structure.

Thus, more information on the true total survey error structure and appropriate resource allocations, is being provided to survey managers and administrators to form a basis for future improvements in total survey quality.

Considerable progress has been made by the Agency in

addressing quality issues in its integrated multiple frame Agricultural Survey Program. Many of the discoveries will translate to improved quality on several other major Agency survey programs as well.

References

Beller, N., "Error Profile for Multiple Frame Surveys," Statistical Reporting Service, Research Report, 1979, Washington, DC.

Bosecker, R., "Integrated Agricultural Surveys," National Agricultural Statistics Service, Research Report No. SSB-89-05, Washington, DC, June 1989.

Fecso, R., "Survey Quality," Presented at the 2nd Quality Assurance in Government Symposium, Washington, DC, May 1989.

Fecso, R., Pafford, B., Tremblay, T., Johnson, R., "Quality Profile

for Soybean objective Yield Survey," National Agricultural

Statistics Service, Unpublished Case Study, Washington, DC, 1988.

DISCUSSION

Barbara A. Bailar

American Statistical Association

I. What is a Quality Profile?

The first quality profile was called an error profile and it

concerned the CPS employment statistics. To be more positive, error profiles have now become quality profiles. The purpose is to prepare a systematic and comprehensive account of survey operations, listing the operations, the potential sources of error, and how the error influences the uses of the survey statistics.

Quality profiles are still rare events. When asked why there are not more, survey producers have three main themes:

- The staff resources that would go into producing a quality profile are too great and are in competition with other, more urgent needs.

- Producing a report that tells about the errors in surveys would lead to less credibility in the statistics produced.

- Admitting that there are errors is admitting that we haven't done our jobs well.

In fact there are many benefits to producing quality profiles. Some of these are as follows:

- to minimize total error, not just sampling error within given cost constraints

- to force a thorough documentation of the survey process.

- to guide a user on the effects of possible errors and their impact on specific uses

- to develop a sound quality control program

- to use in training programs for new staff in either operations or research; and

- to use as the foundation for a sound research and analysis program

The development of a quality profile parallels the survey

process and would contain the following elements:

1. Objectives and specifications of the survey

2. Sampling design and implementation

3. Observational design and implementation

4. Data processing

5. Estimation

6. Analysis and publication

Given this as my basic understanding, let me comment on the quality profile for SIPP and the quality assessment of the Agricultural Survey Program (ASP).

The two reports have some differences and some similarities. The SIPP profile summarizes what is known about sources and magnitudes of errors of estimates and addresses accuracy. The ASP report is written from the point of view of total quality management and uses many of the ideas of Deming, Juran, and Crosby.

This report considers resources, timeliness, and relevance as major components of quality, along with accuracy. The aims of the two groups seem to be quite different.

The two reports each identify the same groups as their targets -- the users of the survey data outside the agency and producers of the survey inside the agency.

Another similarity is that both look at major phases of the survey operation, something essential for a quality profile.

A difference in the two reports was that the SIPP report actually identified four main sources of information on nonsampling errors:

Performance data

methodological experiments

micro-evaluation studies

macro-evaluation studies.

The ASP report was more concerned with process and how quality would be assessed. In fact, the report stresses the need not to identify too many sources of error because tracking everything down might take too long. Actually, I think the total quality management movement urges groups to use brainstorming techniques to identify all possible problems and then Pareto analysis to decide where to concentrate one's efforts.

Another similarity is that both reports left out major steps in the survey process. The SIPP report briefly listed the objectives of the survey, but said nothing about the objectives being conflicting. Producing a survey to give both cross-sectional and longitudinal data has been a new experience for the Census Bureau. The two objectives do conflict, at least from the resource point of view. There were some references to different needs in imputation, but the resource needs have probably had more impact on the survey.

The ASP report did not even list objectives of the survey as a

potential source of error. Neither report really addressed the effects of staff training or compared the kinds of training, length of training, etc. It is fairly well known that performance data does not correlate well with interviewer performance on accuracy. Training could make a difference, but almost nothing is known at the present time.

Let me move now to some separate comments on the two reports, starting with the ASP report. There was a large group of people who worked on this survey quality team. Many of them have done excellent work in survey methodology, so I think we can expect great things from this group., The mission of the group is to contribute to NASS's long term goal of routinely and continually improving survey quality.

The focus on quality at NASS has taken on the language of the quality and productivity movement. For example, they use a simple definition of quality, "fitness for use." This led them on a search to decide what that meant and what objective criteria would be. Finally, they decided that they would measure it by comparison with the Agricultural Statistics Board (ASB) estimate. If the ASB

value is within plus or minus two standard errors of the survey indication, then the survey indication is fit for use. And, in fact, they have five ratings: ideal, acceptable, workable, minimal, and out-of-control.

I find it hard to see why the Agricultural Statistics Board estimate would be used as the standard. In some cases, there are long time series and other indicators that the ASB uses to make its estimate. However, for some surveys they have much less information. Perhaps NASS is pushing the ASB to use the survey indicators or explain why they haven't. Though the example given in the paper about the integrated multiple frame based June acreage estimate was interesting, there will not always be that kind of other data available to compare with.

There is nothing about a Board estimate that measures accuracy. In some ways, it is as if the SIPP people looked at one of their macro indicators and said that if SIPP didn't come within two standard deviations of that estimate, then SIPP was not fit for use. At least, with a macro indicator, one might be able to

untangle why estimates differ; that may not be possible to do with the ASB.

Following Deming's principles, I think the careful documentation of every survey for which millions of dollars are spent and on which important decisions are based is important to the profound understanding of which Deming speaks. A quality profile tells you what you know and what you don't know but should.

It was interesting to see that KASS also addressed resources, timeliness, and relevance as major components of quality. However, it was not clear how criteria would be set or measurements taken. The Gantt chart on the QAS was helpful in identifying time periods and overlaps of one round of survey with the next but it did not help individuals who have many surveys to work on identify overlapping periods of high intensity. The sentence "Too frequent use of overtime to correct a process that is out of control usually

has a devastating effect on overall performance" What does out of control mean? How does it affect overall performance? How do you know these things unless you keep careful records on hours worked on a survey, overtime, and have some measure of a downturn in performance?

NASS has several good ideas about looking at relevance, timeliness, and resources as well as accuracy. It is an ambitious undertaking. I have one word of caution in their drive to use total quality management techniques to help them. They focus on several tools available for a survey quality management system including charting methods. I agree that these are useful tools. But what has been most helpful in the manufacturing and service industries where TQM is used is bringing in a team that has hands-on knowledge of all the facets of the survey. The team would include data collectors from states, edit specifications people, estimation people, those who set objectives. The tools would be something the team would be taught to use to help them. They would all need to learn basic concepts of variability. Only when all these people participate, do you get the profound knowledge that

you need to improve a system, not merely tamper with it. As you recall, tampering with a system does not take care of the major changes needed to remove high variability due to special causes.

Let me now move to the SIPP report. This is a good report that gets periodic updating. There are areas not covered in the report, probably because they did not seem as urgent as the areas covered. However, I do believe that we will need to see a section on objectives, meeting multiple objectives, defining concepts, translating concepts into questions, and so forth. At the other end of the survey, something needs to be said about analysis and publication.

Though the Census Bureau does not use the language of total quality management, I know that they have thought along those lines. Using some of the performance measure standards flies in the face of everything Deming preaches. I'm talking about standards for response rates:

Outstanding..... 97.5 - 100.0

commendable..... 95.5 - 97.4

Fully successful..... 91.5 - 95.4

Marginal..... 88.0 - 91.4

Unsatisfactory..... 87.9 and less

Instead of setting arbitrary standards for response rates and production, the Bureau needs to get a deeper understanding of what is possible in each type of area in which it does surveys. For example, response rates can be charted with upper and lower control limits for PSU's in New York City. Probably the response rates there very seldom, if ever, meet the commendable level. However, they may be within normal variability for that area. Only with positive efforts at changing the system can the response rates be lowered. This is partly what Dr. Deming thunders about -- blaming the worker who may be doing the best he or she can when it is the system at fault. Again, this labelling of people's work does not make the interviewer proud, and it is really tampering with the system.

The report gave lots of interesting information on household, person, and item response rates. Some of the non-response rates on asset data are such that it seems questionable that the survey is the right vehicle for collecting the data.

There is also emphasis on the seam problem, but this is nothing new. As I recall, it also showed up in the crime survey. It seems that certain biases are endemic to longitudinal surveys. So far the Bureau has been content to catalog the measured effect. We really need some creative thinking and some money to get some experiments going to look at recall errors, the placement of events in time, and the time in sample problems. Though dependent interviewing may yield more consistent results, they may be no more accurate. Before action is taken to fix a problem, there needs to be a deeper understanding of why the problem exists.

There was very little information available on the extent of editing, what it does, why changes are made, and what we call editing and what we call imputation. Beller made some very pertinent comments in his 1979 error profile for NASS surveys.

"The amount of editing on some questions resulted in changing the level of cattle and calves by an amount two or three times greater than the error caused by sampling. This amount of editing is cause for alarm in that it clearly shows a breakdown in the survey process." In both the NASS surveys and SIPP, we need to get a better picture -- a profound understanding -- of what editing is doing to the data.

One last point on SIPP. The only direct estimates of sampling error were for the third quarter of 1983 using 1984 panel data collected in wave one. The survey at that time was based on the 1970 census. It certainly seems time to recompute variances. Besides having incorrect variances, it seems like gilding the lily when the analysts are making actual and implied comparisons that they multiply by 1.6 times the standard error. The interpretations and the comparisons could be quite far off.

All in all, I enjoyed reading these papers. I think the documentation of SIPP is more complete but I think NASS is farther along in trying to improve quality. They do not want to document only; their real goal is improvement. I believe that is ultimately the SIPP goal too, but no strategy has yet been set forward on how to move in that direction.

Discussion

Nancy A. Mathiowetz

U. S. Bureau of the Census

The data collected by Federal statistical agencies are used to both shape federal policy and change the distribution of federal expenditures; given the magnitude of the impact of these data, the need for high quality goes without question. In developing the Quality Profiles, the agencies responsible for this work are to be commended for continuing to move the discussion of error beyond that of sampling error and into the realm of the measurement of nonsampling error. Although most agencies have for years provided discussion of sampling error with release of their data and research findings, we are just beginning to develop a standard of reporting which includes a discussion of all of the components of total survey error.

Sources of Nonsampling Error

The sources of nonsampling error are many and include:

- the design of the study (e.g. longitudinal vs. cross sectional; length of recall period;

- the questionnaire, both the contents and the structure;

- the interviewer;

- the respondent; and

- the post-survey processing, including coding and keying of data.

Rather than reiterate issues raised in the Quality Profiles, I would like to suggest some other topics of investigation within these sources of nonsampling error. My goal in doing so, is not to criticize the work presented here, but to provide some ideas on where these Quality Profiles could be expanded.

Design

With respect to design, we still know little about the effects of longitudinal designs on the level of error and the error variance structure of reports over time. There has been research

to indicate that respondents suffer from "conditioning" effects, that is the changing of behavior or the reporting of behavior in later interviews resulting from earlier interviews. Some conditioning may improve reporting in that the respondent knows

prior to the interview what are the nature of the questions; conditioning may also result in a reduction in reporting since respondents are now knowledgeable about the sequencing within an interview. In one study, the best predictor of error in reports of functional status in the fourth round of interviewing is the length of time it took to conduct the previous interviews. The finding suggests that conditioning effects may be reduced by something as subtle as reducing the length of an earlier interview. We need further research to understand how conditioning impacts the analysis of change over time and the structure of errors over time.

Longitudinal designs may also be affected by changes in the respondents the interviewer, or even the interpretation and meaning

of critical concepts in the questions, if the panel has a long life. With the proliferation of more longitudinal data collection efforts within the Federal Government, more research into what questions are sensitive and which are resistant to conditioning effects as well as which items are most affected by between interview changes, is necessary.

Questionnaire

As noted in a lecture to the Society of Government Economists, Janet Norwood stated that

> ...the quality of a statistical indicator is sometimes elusive and often difficult to define. Effective measurement requires an underlying conceptual framework and careful identification of the phenomenon to be estimated....
>
> In the past 25 years, we have made great strides in understanding how sensitive response distributions are to minor

changes in question wording. The merging of literatures from cognitive psychology, social linguistics, and social psychology with survey methodology has presented use with new means for attempting to reduce the levels of error associated with the questionnaire. What is now needed in the Federal statistical system is a means for evaluating the various forms by which the "same" information is collected and analyzed among various agencies. For example, in recent years, the proportion of individuals lacking health insurance has been a critical issue. The most widely cited data on insurance coverage comes from the Current Population Survey, which asks whether each person in a household was covered at any time during the preceding year. Persons covered by any source at any time during the year are counted as insured. In 1987, the estimate for uninsured from the March CPS was 17.6 percent. Notice that this question asks whether the person has been covered "at any time" during the previous year. In contrast questions from the 1980 National Medical Expenditure and Utilization Survey (NMCUES) and the 1987 National Medical

Expenditure Survey (NMES), both designed as one-year panel surveys, indicate that point in time estimates of the uninsured (at the time the person was interviewed) are approximately 14 to 16 percent at any one cross-section, but that estimates for all year uninsured are approximately 9 percent.

There is some conjecture that the response to the CPS may reflect a respondent's status at the time of the interview rather than in reference to any time in the previous year, due to the, similarity in the estimates from CPS and the cross-sectional estimates from NMCUES and NMES. From a policy perspective the difference is critical -- whether to provide health insurance for the chronically uninsured, approximately 21 million people, or whether to provide insurance for all individuals ever uninsured, which appears to be approximately 35 million people in a given year. Those attempting to address this issue would benefit from a consistent definition of uninsured as well as a set of questions which asks about a consistent time period.

The use of response rates, hours per completed interview and item nonresponse rates traditionally used as measures of interviewer quality, only begin to capture the errors that are potentially associated with the interviewer's task. While each of these measures provides us with information that we believe is related to quality, we need to employ more measures that could be used with respect to understanding error for individual questions. How well do interviewers understand the concepts underlying the questions they are asking? Do they have sufficient training and understanding to ask non-directive probes when necessary to obtain an adequate answer? The increased movement toward telephone interviewing provides use with a means to routinely randomize interviews across interviewers to obtain measures of interviewer variance. We spend millions of dollars in the training of interviewers and yet know little about the most effective means for training interviewers or determining their ability to conduct the interview as trained. The review of one or more interviews by a supervisor provides some information, but if we believe that

training interviewers to read questions exactly as written is worth the cost, we should be routinely evaluating the association between the delivery of questions and the error associated with the responses.

Editing and Coding

As noted in the SIPP Quality Profile, much of the between wave difference in industry and occupation appears to be a spurious result of either data collection or data processing. A similar problem can be found in the coding of medical conditions and

surgical procedures based on household reported data. Not only coding, but also editing procedures, can contribute to the overall level of error in estimates. For example, Duncan and Mathiowetz (1985), using microlevel validation data, found that trimming estimates of change in income between two years, that is disbelieving levels of change beyond a certain level as reported by household respondents, a procedure often done in editing data from

longitudinal surveys of income, resulted in biased estimates of change and bias in the coefficients predicting income levels and change. Retrospective reports of income were more likely to be correct for those individuals with a large proportional change than for those with little or no change. The finding suggests that editing procedures should be conservative and based on empirically derived principles.

Whereas we have learned to be sensitive to question wording with respect to understanding potential sources of bias, and in doing so demand documentation concerning question wording and study design, few, if any, studies provide information on effects of editing and coding processes. If consumers of the data are to understand all aspects of total survey error, coding and editing decisions need to be researched and documented.

Adjusting for Nonresponse

For the most part, nonresponse adjustments are made using

demographic and segment information and little if any information concerning the nature of the nonresponse is factored into the adjustment. There is a growing body of literature which suggests that using information from call records, specifically separating refusals from those you were unable to locate, in a nonresponse adjustment may prove beneficial, since difficult to locate (but eventually interviewed) sample individuals look similar to respondents who cannot be located.

These comments are intended to extend the excellent work presented in the Quality Profiles. The profiles provide details on the measurement of nonsampling error and the results of several experiments to reduce these levels of error. In addition. I hope that as others consider producing quality profiles these profiles are expanded to cover some of these other issues.

Reference

Duncan, G.J. and Mathiowetz, N.A. A Validation Study of Economic

Survey Data, Ann Arbor, MI: The Institute for Social Research, 1985.

Session 2

PARADIGM SHIFTS USING ADMINISTRATIVE RECORDS

PARADIGM SHIFTS: ADMINISTRATIVE RECORDS AND CENSUS-TAKING

Fritz Scheuren

Internal Revenue Service

There is a lot in the news lately about problems with the 1990 decennial census in the United States. Many opinions have already been offered about what went wrong and what should be done. Indeed, a paradigm shift may be needed in census-taking.

This brief note talks about the possible role administrative

records might play in a new paradigm. To get things started, the word "paradigm" might deserve some elaboration: a paradigm is a way of thinking and then doing; a pattern of belief and behavior; a way of seeing reality and using that sense to accomplish something. Paradigms are common -- the way we get to work would be a humble example. Conventional census-taking, under this definition, could be characterized as a major scientific and technical paradigm.

As long as our paradigms work well for us, we tend not to change them. Occasionally, however, paradigms break down and have to be replaced; e.g., the bridge goes out and we need to find another route to work. As Kuhn pointed out in his seminal book on the structure of scientific revolutions, paradigms break down in, science, as well (Kuhn, 1970). Perhaps the most famous example of this is the revolution in the thinking of astronomers that occurred when the Ptolemic earth-centered view of the universe was replaced by the Copernican view of an earth that revolved, with the other planets, around the sun.

If we look at the problems the U.S. Census Bureau has

encountered with the 1990 decennial census, it can easily be argued that one of the major barriers to overcoming these obstacles is the conventional census-taking paradigm. Kish, in a recent paper he has written for Survey Methodology (1990), considers at length some possible alternatives. My objective here will be to focus on two of those areas -- rolling censuses and administrative registers and to explore a new paradigm for the U.S. decennial census.

Conventional Census-Taking

Conventional censuses, like those in Canada and the U.S., continue to do many things very well (e.g., Hammond, 1990). Indeed, at present, we have no adequate substitute for them;

nonetheless, the need for at least some change seems compelling. Rising costs are a big factor. There have been many improvements in census-taking in this century; still, in both Canada and the U.S., total costs and even costs per person have risen significantly:

- The 1990 decennial census in the U.S. is budgeted at about $10 (U.S.) per person. Even adjusting for inflation, this is a four-fold increase over what the per capita expenses were in 1960. Item content differences between the two censuses are small and essentially not a factor in explaining the difference. Both the 1960 and 1990 Census, for example, asked only 7 population questions of everyone (U.S. Bureau of the Census, 1989). The Census long-form sample in 1960 contained 35 questions and was to be completed by 25% of the population. For 1990, the Census long-form sample was given to 16% of U.S. households and had 33 questions.

- The situation in Canada is similar with regard to the

costs of census-taking. For example, the 1991 Canadian Census is budgeted at about $9.50 (CAN) per person. Like the U.S. Census, there are again just 7, albeit somewhat different, population items that are asked of everyone. Like the 1990 U.S. Census, questions on housing are included for everyone (2 in Canada and 7 in the U.S.). In Canada, a 20% long-form sample will be employed in 1991. The Canadian long-form questionnaire has 45 items for 1991. The 1961 census in Canada was quite different from that planned for 1991 and thus meaningful cost comparisons are hard to make. Nonetheless, looking back 30 years in Canada, the same long-term trend in census-taking costs seems to exist; however, per capita costs have been roughly the same -- even declining slightly in the last two or three censuses.

The U.S. Census Bureau has looked at the growing cost of conventional census-taking and concluded that a major change may be needed (Browne, 1989). Labor costs have grown appreciably in recent decades in both Canada and the U.S. Technological improvements have not been great enough to offset these costs,

though some, like TIGER (Topologically Integrated Geographic Encoding and Referencing) and CATI (Computer-Assisted Telephone Interviewing), offer promise. Greater attention in the U.S. to improved population coverage is another important factor (Anderson, 1990). The degree of public cooperation in the census also seems to be dropping, at least as reflected by the poorer than

anticipated mail response rate for the 1990 U.S. Census. (It should be noted that this same tendency is not clearly apparent in Canada.)

Increasing cost is not the only major problem facing conventional census-taking. Perhaps of even greater importance is the growing rate of obsolescence of the information collected. The combination of rising costs and growing information obsolescence has had the effect of reducing the benefit/cost ratio for conventional censuses steadily and dramatically.

To obtain more frequent small area data, some countries have introduced quinquennial censuses. For example, in Canada this was first done nationally in 1956. Budget problems led to the 1986 Canadian Census being cancelled and then reinstated. Indeed, it is unclear whether there will be a Canadian census in 1996. While a quinquennial census was also legislated in the U.S., funds were never made available.

Rolling Censuses

Conventional census-taking, of necessity, must sacrifice both timeliness and item content (on a 100% basis) to achieve complete spatial detail and high population coverage.

One of the alternatives that Kish asks us to look at is a "rolling census." His proposal envisions the sampling of a country over a decade in such a way that every area is eventually covered. In its purest form, space and time become a single dimension and content remains fixed, such that, at decade's end, we have obtained

cumulative information on the entire country for a given set of items.

The chief advantage of a rolling census is that it can avoid the problem of information obsolescence at national and major subnational levels. For small geographic areas, though, there would, of course, still be only one observation per decade. Unlike a conventional census, comparisons among small geographic areas would be very difficult to interpret because the data are being collected at different points in time (Fellegi, 1981).

For a rolling census or survey, unit costs could be higher, as Kish notes, than in a more conventional enumeration (indeed, ceteris paribus, maybe even higher than the cost of existing survey efforts) . In an age of fixed or declining resources, therefore, it might not be possible to do a complete "enumeration" each decade, even if content were significantly scaled back. Rolling samples would seem to have their greatest attractiveness not as a replacement for conventional censuses, but, say, as part of a strategy to link together census-taking with ongoing surveys and

local area population estimates for the intercensal years (Herriot, Bateman and McCarthy, 1989).

Both the United States and Canada employ monthly surveys to estimate the national (and some subnational) labor force characteristics. The Canadian Labor Force Survey (LFS) of 64,500 households covers 0.67% of the total Canadian population each month. "Given the rotation pattern in effect for the LPS, the 0.67% sample per month rolls up into a 6.7% sample of unique households over a 5-year period" (Drew, 1989). In the Canadian context, at least, Kish's proposal may be feasible. A sample survey vehicle could be designed, with some reduction in the month-to-month household overlap, which could achieve many of the benefits he has stated for a rolling sample, while also meeting the information needs currently met by ongoing household surveys (Drew, 1989). This sample would not replace the 100% census count data, itself, but, might be a partial substitute for Canada's 20% long-

form census sample.

Because the United States has a population about 10 times larger than Canada, the tradeoffs involving rolling samples and overall country coverage are not as favorable as they are in Canada. The U.S. Current Population Survey (CPS), for instance, at about 60,000 households, covers only .06% of the total U.S. population monthly. Even if cumulated over a whole decade (but, with no change in its rotation pattern), the CPS would cover just. roughly 1% of all U.S. households. This does not compare well in size to the overall 16% long-form sample being conducted as part of the 1990 U.S. Census.

To bring the rolling sample population coverage nearer to the 1990 U.S. decennial sample, major changes in the CPS rotation pattern would be needed. Other U.S. Census Bureau surveys might also have to be redesigned if the objective were to achieve even a partial substitute. Despite these changes, moreover, the resulting decade-long sample would still be only a small percent of the total U.S. population -- perhaps, at best, in the 2% to 3% range,

assuming resources and other requirements remained essentially fixed.

In both Canada and the U.S., the likely higher unit costs of a rolling sample may need to be addressed by changes in survey procedures: how area segments are listed (Royce and Drew, 1988); how first contact with households is made, etc. Where is it written, for example, that a personal interview contact is heeded before using other modes of collection?

It will be no mean challenge to keep effective sample sizes equal for the major level and change components now obtained from ongoing surveys (e.g., Tegels And Cahoon, 1982). Some compromise may be needed, moreover, in the extent to which the basic content of the current long-form census samples can be included. Despite

these challenges, or perhaps because of them, rolling samples deserve continued serious attention and should be the focus of

extensive practical experimentation.

Administrative Registers

With the flowering of scientific sample survey methods in the 1940's (Bailar, 1990), the use of administrative records for statistical purposes became relatively less important in many national statistics programs. By the early 1980's, however, at least in the developed countries, the pendulum had begun to swing back. Philip Redfern has been the major chronicler of this phenomenum internationally (Redfern, 1987). While the Danes seem to have gone the farthest (Jensen, 1983 and 1987), major efforts have been made in Canada (e.g., Statistics Canada, 1990) and even some in the U.S. (e.g., Alvey and Kilss, 1990).

A good summary of most of the key barriers to the greater use of administrative registers for census-taking is found in Redfern (1989) including the extensive discussion published with that paper. Perception barriers by the citizens (e.g., in Germany) are mentioned as problems. Psychological barriers by the national

statistical service may, however, be of equal or even greater importance. Major scientific "paradigm shifts" generally have this problem (Kuhn, 1970). Certainly, this seemed to be part of the reason for the reception given to the proposal (made by me in 1980) to explore the feasibility of making administrative records an integral part of the U.S. Census of Population. While a sketch of such a proposal was eventually given at the 1982 American Statistical Association meetings (Alvey and Scheuren 1982), it seems, with a few fairly limited exceptions (e.g., Irwin, 1984-Citro and Cohen, 1985), that serious interest at the Census Bureau has been notably lacking.

Suffice it to say that in the U.S. very little of the needed research has been undertaken. This is true, despite continuing efforts to give the proposal prominence (Jabine and Scheuren, 1985 and 1987) and to get it discussed widely (Butz, 1985). Sadly, therefore, it appears that, in the United States, at least for the year 2000, we should not expect administrative registers to replace censuses.

The 1990 U.S. decennial census could have been used as a

proving (or disproving) ground for some of the needed research into administrative record alternatives. Why that didn't happen is a matter that can only be speculated about. A contributing factor, quite possibly, is a case of "paradigm paralysis" (Barker, 1988). The literally decades-long controversy about whether to adjust census "counts" seems to have locked the U.S. Bureau of the Census into what some, at least, would call an increasingly sterile intellectual position (Fienberg, 1990). The viewpoint that they

have adopted makes it very hard for them to see any alternative, like a (partial) administrative record approach, that starts out with the notion that adjustments would be required.

The situation is different in Canada. Since the late 1970's, Statistics Canada has assembled many of the building blocks needed to conduct an administrative record census (e.g., Drew, 1989; Podoluk, 1987; Verma and Raby, 1989). While much remains to be done, such a change could even happen as early as 1996. For

example, the coverage of the Canadian tax return system, alone, is quite high and growing. In 1987, for instance, it has been estimated that the coverage was about 94% -- i.e., about 3% less than the 96.8% coverage achieved in the 1986 Canadian Census. By 1991, the tax return coverage, alone, should be up to about 97% or better, with overall administrative record coverage still higher and likely to grow further in the 1990's. (See Table I for more details oh administrative record coverage in Canada and the U.S.)

One concern often raised is that administrative registers, even after they become adequate in quality and coverage, will be

58

limited to only a few, bare demographic variables: head counts, age, sex and little more. An immediate observation concerning this remark is that conventional censuses do little more than this, themselves at least for the 100% items. It is also evident that, while the variables on administrative records are not the same as those collected in a traditional census, there is more already available than critics may realize (e.g., Meyer 1990; Alvey and Scheuren, 1982).

More important even than any current item content comparison is the need to emphasize that the proposal to use administrative registers in census-taking does not envision that administrative records have to be used as they are. Administrative records will need to be changed. In my personal opinion, limited optimism about achieving needed changes is justified. However, without a doubt

it is too much to expect of administrative records that they will be able to capture exactly the same concepts now measured in censuses and surveys. Additionally, there almost certainly will need to be special efforts, using existing census-taking techniques, to separately enumerate certain groups. The efforts in the 1990 U.S. Census to count the homeless would be one such example.

Censuses and administrative records each have inherent limitations. Unavoidable Conceptual differences will be a major barrier to any shift from one medium to another. Administrative feasibility is another issue; however, some hard-to-duplicate census concepts (e.g., households) may not be as important to the measurement process as formerly.

Shifts in methodology (from conventional census to administrative records) for some uses would potentially be accompanied by a parallel shift in the underlying concepts measured. some concepts may alter or expand in meaning, including our ability to measure them (e.g., families). We also must ascertain the extent to which respondents answer survey questions

the same way they fill out administrative forms that may have real direct impact in their lives.

In recent years, traditional survey methodology has been enhanced by new tools from the field of cognitive psychology. These cognitive research tools could be used to understand any conceptual differences between the meaning of terms when they are used in surveys or drawn from administrative records. We may not have what we think we have anyway (Bates and DeMaio, 1989). In any case, there is already an extensive body of cognitive research that can be drawn on (e.g., Dippo, 1987; Fienberg and Tanur, 1989; Jobe and Mingay, 1990).

It should also be pointed out that, most likely, administrative registers will not be able to completely meet the demands of modern society for richer sources of statistics. Such

demands, of course, appear to be insatiable. Even if they were

not, administrative records will never have the flexibility and responsiveness of surveys. Registers, however, (including partial ones like those that exist in the U.S.) when linked to survey data, can be extremely important as auxiliary variables in making improved direct national survey -- and even subnational survey -- estimates. The U.S. Census Bureau's Survey of Income and Program Participation research on the use of Internal Revenue Service data for improving the precision of national survey estimates is a good recent example (Huggins and Fay, 1988). Indirect (e.g., synthetic) estimates for small areas would still be needed for variables not on the administrative registers (Platek, Rao, Sarndal, and Singh 1987). The registers, though, might provide a source of valuable symptomatic indicators.

Concluding observations

The case for considering a "paradigm shift" in census-taking seems compelling, at least in developed countries like Canada and the U.S. The rolling census alternative Kish proposes is probably

too expensive to fully implement as a complete substitute for a census. Rolling samples do offer real promise, however, if they can be integrated into the current ongoing survey operations of Canadian and U.S. national statistical programs. Such samples could provide a needed link in addressing small area estimation needs that might otherwise not be met. Less promising, but still possible, is their use as a (partial) substitute for the census long-form samples.

As far as administrative registers are concerned, critics may have been unduly pessimistic. The Canadian situation, however, differs from the United States:

o In Canada, it is already within the realm of feasibility to combine rolling samples with administrative records as an alternative to conventional census-taking. This is not to say that enormous practical challenges don't remain. The 100% count portion of the Canadian census, though, could be done with administrative records as a starting point, augmented by a large-scale survey to

measure and potentially adjust for undercoverage. The Canadian 20% census long-form sample might be, at least partially, replaced by a rolling sample. The content of the Census long-form is considerably richer than that of household surveys, but the content differences could be made up through additional questions "piggybacking" the on-going surveys at regular intervals. Coverage issues surrounding the use of administrative records could also be addressed directly with rolling samples, especially to calibrate for changes in administrative records between censuses.

- In the United States, the U. S. Census Bureau has begun to look at alternatives other than conventional census-taking (Bounpane, 1988). Unfortunately, the research needed to look at an administrative register Alternative has barely begun. Whether the Census Bureau will find a better approach than the use of administrative records and rolling samples remains to be seen. Whatever other

alternatives they study, however, the use of administrative registers as a partial replacement for the conventional 100% counts definitely needs to be considered. A preliminary research agenda updating earlier ideas will appear in Scheuren, Alvey and Kilss, 1990.

Naturally, with such radical proposals, the answer is uncertain. Like Kish, I believe that "the balance of variance components" favors a change from conventional census-taking in most cases. However, as Kish states, "theoretical as well as empirical investigations will be needed to decide matters" (Kish, 1990).

In a change as big as the one proposed here, the "balance" that needs to be struck goes, of course, well beyond looking at variance (and bias) components. One issue that needs to be emphasized more is that some aspects, at least, of the paradigm shifts being considered could go to the heart of the social contract that exists between national statistical agencies and the people that those agencies have a mission to serve. For instance,

in the U.S. Constitution, there is a requirement that an "enumeration" of the population take place every ten years. Would the use of administrative records or rolling censuses fit within this "Constitutional paradigm?" Perhaps the starting place is to adopt a broader definition of "enumeration."

Another example where social contract issues arise is the extent to which the greater use of existing (or expanded) administrative data for statistical purposes might be seen as an unwelcome increase in the intrusiveness of the State into the private lives of its citizens (Grace, 1989). As legitimate as concerns about "intrusiveness" might be, though, there is no evidence in a North American context, at least, that they pose an insurmountable barrier. On the contrary, there have been virtually no adverse public reactions to past U.S additions to administrative records for statistical purposes (e.g., of residential address, information in 1972, 1974 and 1980 tax returns). To my knowledge the issue, so far, has not come up directly yet in Canada, at least at the Federal level.

In summary, to make these kinds of changes there is the need for a lot more scientific research. Studying the implementation technologies will be an even bigger job. Finally, the issues go beyond our profession and may well be settled in other arenas. Wherever they are decided, it is incumbent on us, as statisticians,

to frame the debate in terms of feasible options. Hopefully, exchanges such as ours today will help lead the way along that path.

References

Alvey, Wendy and Kilss, Beth (eds.) (1990). Statistics of Income and Related Administrative Record Research, U.S., Department of the Treasury, Internal Revenue Service. See also Kilss, Beth and Alvey, Wendy (eds.) (1984). Statistical Uses of Administrative Records: Recent Research and Present Prospects, vols. 1 and 2, U. S.

Department of the Treasury, Internal Revenue Service.

Alvey, Wendy and Scheuren, Fritz (1982). "Background for an Administrative Record Census," 1982 American Statistical Association Proceedings, Social Statistics Section, pp. 137-146.

Anderson, Margo (1990). According to Their Respective Numbers ...' for the Twenty-First Time," Chance, vol. 3, no. 1, pp. 12-18.

Bailar, Barbara (1990). "Contributions to Statistical Methodology from the Federal Government," Survey Methodology, vol. 16, no. 1, Statistics Canada.

Barker, Joel Arthur (1988). Discovering the Future: The Business of Paradigms, Institute for Information Studies.

Bates, Nancy A. and Demaio, Theresa J. (1989). "Using Cognitive Research Methods to Improve the Design of the Decennial Census Form," Proceedings of the Fifth Annual Research Conference, U.S. Bureau of the Census, 267-285.

Bounpane, Peter (1988). "A Sample Census: A Valid Alternative to a Complete Count Census?" 46th Session of the International Statistical Institute.

Browne, David (1989). "U. S. Bureau of the Census: Facing the Future Labor Shortage," Asian and Pacific Population Forum, vol. 3, no. 4.

Butz, William (1985). "Comment: The Future of Administrative Records in the Census Bureau's Demographic Activities," Journal of Business and Economic Statistics, vol. 3, no. 4, pp. 393-395.

Citro, Connie and Cohen, Michael L., eds. (1985). The Bicentennial Census: New Directions for Methodology in 1990. National Academy Press, Wash., DC.

Dippo Cathy (1987). "A Review of Statistical Research at the U.S. Bureau of Labor Statistics," Journal of Official Statistics, vol. 3, no. 3, pp. 289-297.

Drew, J. Douglas (1989) "Address Register Development and its Possible Future Role in Integration of Census, Survey and Administrative Data," A paper presented at the U.S. Bureau of the Census/Statistics Canada Interchange. (Unpublished)

Fellegi, I.P. (1981). "Comments," Discussion of a paper by Leslie Kish entitled "Population Counts from Cumulated Samples," Using Cumulated Rolling Samples to Integrate Census and Survey operations of the Census Bureau, An Analysis, Review, and Response, Congressional Research Service, the Library of Congress.

Fienberg, Stephen (1990) "An Adjusted Census in 1990? An Interim Report," Chance, vol 3, no. 1, pp. 19-21.

Fienberg, Stephen and Tanur J. (.1989). "Combining Cognitive And Statistical Approaches to Survey Design," Science, 243, pp. 1017-1022.

Grace, John W. (1989). "The Use of Administrative Records for Social Research," Statistics Canada Workshop, December 12, 1989, Ottawa, Ontario.

Hammond, Robert B. (1990). "The 1990 Decennial Census: An Overview," Conference Proceedings, Advanced Computing for the Social Sciences, sponsored by the Oak Ridge National Laboratory and the U.S. Bureau of the Census, April 16-12, 1990, Williamsburg, Virginia..

Herriot, Roger; Bateman, David V.; and McCarthy, William F. (1989). "The Decade Census Program -- New Approach for Meeting the Nation's Needs for Sub-National Data," to appear in American Statistical Association Proceedings, Social Statistics Section.

Huggins, Vicki and Fay, Robert (1988). "Use of Administrative Data in SIPP Longitudinal Estimation," American Statistical Association Proceedings, Section on Survey Research Methods.

Irwin, Richard (1984). "Feasibility of an Administrative Records

Census in 1990," Special Report on the Use of Administrative Records, Committee on the Use of Administrative Records in the 1990 Census, unpublished Census Bureau report.

Jabine, Thomas B. and Scheuren, Fritz (1985). "Goals for Statistical Uses of Administrative Records: The Next Ten Years," Journal of Business and Economic Statistics, vol. 3, no. 5, PP. 380-391.

Jabine, Thomas B. and Scheuren, Fritz (1987). "Statistical Uses of Administrative Records in the United States: Where Are We and Where Are We Going?" Proceedings of an International Symposium on Statistical Uses of Administrative Data, J.W. Coombs and M.P. Singh (eds.), Statistics Canada, December 1988, Ottawa, pp. 43-72.

Jensen, Poul (1983). "Towards a Register-Based Statistical System -- Some Danish Experiences," Statistical Journal of the United Nations Economic Commission for Europe, vol. 1, no. 3, pp. 341-365.

Jensen, Poul (1987). "The Quality of, Administrative Data from a Statistical Point of View: Some Danish Experience and Consideration, Proceedings of an International Symposium on Statistical Uses of Administrative Data, J.W. Coombs and M.P. Singh (eds.) Statistics Canada, Ottawa.

Jobe, Jared B. and Mingay, David J. (1990). "Cognition and Survey Measurement: History and Overview," Applied Cognitive Psychology, in press.

Kish, Leslie (1990). "Rolling Samples and Censuses," Survey Methodology, in press.

Kuhn, Thomas S. (1970). The Structure of Scientific Revolutions, Second Edition, Enlarged, The University of Chicago Press, Chicago.

Meyer, Bruce (1990). "The Tax System: Comparisons of Demographic, Labour Force and Income Results for Individuals and Families," Small Area and Administrative Data Division, statistics Canada.

Platek, R.; Rao, J.N.K.; Sarndal, E.E.; and Singh, M.P. (1987). Small Area Statistics, New York: Wiley-Interscience.

Podoluk, J. (1987). "Administrative Data as Alternative Sources to Census Data," Proceedings of an International Symposium on Statistical Uses of Administrative Data; J.W. Coombs and M.P. Singh (eds.), Statistics Canada, December 1988, Ottawa, pp. 273-290.

Redfern, Phillip (1987). "A Study of the Future of the Census of Population: Alternative Approaches," Eurostat Theme 3 Series C, Luxembourg: Office for Official Publications of the European Communities.

Redfern, Phillip (1989). "Population Registers: Some Administrative and Statistical Pros and Cons," The Journal of the Royal Statistical Society, Series A (Statistics in Society), vol. 152, pt. 1, pp. 1-41.

Royce, Don and Drew, J. Douglas (1988). "Address Register Research: Current Status and Future Plans," 1991 Research and

Testing Project, 1991 Census, Statistics Canada, Ottawa.

Scheuren, Fritz; Alvey, Wendy and Kilss, Beth (1990). "Paradigm Shifts: The Integration of Administrative Records and Surveys," a paper delivered at the 151st Annual Meeting of the American Statistical Association, August 7, 1990, in Anaheim, CA.

Statistics Canada (1990). "Research Papers and Reports," Bibliography Small Area and Administrative Data Division, Ottawa, Ontario. (unpublished)

Tegels, Robert and Cahoon, Lawrence S. (1982). "The Redesign of the Current Population Survey: The Investigation into Alternate Rotation Plans," Proceedings of the American Statistical Association, Survey Research Methods Section.

U.S. Bureau of the Census (1989). 200 Years of U.S. Census Taking:

Population and Housing Questions, 1790-1990, Superintendent of Documents, U.S. Government Printing Office, Washington, DC.

Verma, Ravi B.P. and Raby, Ronald (1989). "The Use of Administrative Records for Estimating Population in Canada," Survey Methodology, vol. 15, no. 2, pp. 261-270.

AN ADMINISTRATIVE RECORD PARADIGM:

A CANADIAN EXPERIENCE

John Leyes

Statistics Canada

1.0. Introduction

In 1979, Statistics Canada began a formal review of the potential of using administrative records for social statistical applications for small area data (Statistics Canada, 1979). Based on this review, it was concluded that the highest coverage of the population and the greatest potential for social administrative data would arise through the use of the personal income tax records. With few exceptions, then, this paper considers data derived from the personal income tax file in Canada.

The Canadian tax system differs from the U.S. tax system. For example, in Canada, there is no joint filing; and the tax system is used as an instrument to provide benefits to persons and families with low incomes. The personal income tax return is known as the T1. The Tl serves a purpose similar to the IRS' Form 1040.

In its earliest days, Statistics Canada's work with the personal income tax file was subject to a number of expected a

priori shortcomings. These shortcomings represented an adminis-

trative records paradigm (or rules of the game). The shortcomings

included the following:

- Population Coverage. The income tax system is based on individuals only. Since only 60% of Canadians were filing tax returns in the mid-1970's, coverage was deemed inadequate for social statistical applications.

- Population Coverage Bias. The age profile of taxfilers differed from the age profile of the population. This was judged to be an unacceptable bias.

- Income Coverage. Not all income received by Canadians is taxable; hence, the income coverage of the Tl was considered incomplete.

- Income Distribution Coverage. Since both the elderly and the young frequently have low incomes and do not file tax returns, data from the tax file would be inadequate for

public policy purposes directed at these target groups.

- Dimensionality of Variables. Since any single administrative record has a specific and narrow application in program administration, the range of data variables were also judged inadequate as a source of social data.

- Concepts and Definitions. The concepts and definitions used in household surveys and censuses of population can only be approximated through the use of an annual tax file.

Each of the above represented a limitation or shortcoming for data derived from administrative records in general, the Tl in particular. In spite of these shortcomings, the work began, and

this paper is directed at a few findings that resulted from work in Canada with the personal income tax records in the development of family data since 1984. Perhaps this paper may even indicate some of the potential of using the Tl as a source of small area data in post-censal periods in Canada.

2.0. The Development of Taxfiler Family Data

The taxfiler family concept has been designed to emulate the census family concept. A census family:

> "[r]efers to a husband and a wife (with or without children who have never married, regardless of age), or a lone parent of any marital status, with one or more children who have never married, regardless of age, living in the same dwelling. For census purposes, persons living in a common-law type of arrangement are considered as now married, regardless of their legal marital status; they accordingly appear as a husband-wife family in census family tables." (Statistics Canada,

1982, p. 29)

This concept is suitable for household collection methods since respondents are asked to report on the relationships between all residents of a dwelling. With administrative records, secondary information such as reported marital status, value of exemptions/tax credits, ages of taxfilers, addresses, child care expenses, and so forth, are used for forming families.

It has not, therefore, been possible to emulate the exact census family concept. The major sources of difficulty arise with older children (whether they have ever been married or not when they reside with their parents) and with common law couples. In

general, the census family concept works reasonably well for families with dependent children, and some success has been achieved in estimating single parent and common law families, as

can be seen in Table 1.

In 1984, Statistics Canada began estimating families from the individual taxfiler (Tl) data. The creation of families from the Tl is based on a six-step process:

i. Taxfilers, reporting the Social Insurance Numbers (SIN) of their spouses, are matched to form husband-wife families;

ii. Other husband-wife families are formed from taxfilers who declare themselves married but do not report spousal SINs;

iii. Non-dependent filing children who reside with their parents are matched to their parents;

iv. There is an intermediate step to unduplicate records, to identify one-filer husband-wife family units, to assign a unique postal code to family members, and to assign a

family composition type to each family unit;

v. Common law spouses are matched from the pool of individuals classed as single parent families and non-family persons; and

vi. In Step 6 non-filing family members are imputed.

With this brief introduction and description of the taxfiler family data, it is now possible consider some data findings.

3.0. The Coverage Shortcomings, Some Empirical Findings

3.1. Population Coverage Comparison: 1985 Taxfiler Family File (T1FF) to 1986 Census of Population

The taxfiler family data have been placed into four classifications: husband-wife families, single parent families, common

law families, and non-family persons. The data in Table 1 reflect the first three of these classifications (common law families are noted twice, once as husband-wife families, and then separately).

In creating the taxfiler family (T1FF) data, a record is created for each family member and for each non-family person. Thus, there is a record for a taxfiler and a record for each person that is imputed. Line three of Table 1, therefore, is an estimate of the TIFF population that can be identified through the tax system.

69

There are several highlights in Table 1:

o The T1FF population has varied between 93.7 and 95.8 percent during the 1982 and 1987 period (line 6 of Table 1);

o The number of total tax family records (taxfilers plus imputed) increased at a slightly lesser rate than the number of taxfilers alone (i.e. lines 2 and 4);

o Coverage of husband-wife families was slightly higher than coverage of total family records (i.e., lines 6 and 9); and

o single parent overcoverage decreased for 1986 and 1987.

3.2. Population Coverage Bias, 1985 T1FF to 1986 Census

In Table 2, some broad age range comparisons have been included. The first age range is, perhaps, a bit unusual since it includes the population 29 years of age and below. This age range resulted from an arbitrary decision, namely, that the maximum age of a matched filing child could be 29. Furthermore, for imputed, non-filing dependent children, there is limited age information and no gender information. Thus, children, whether imputed as dependents or identified as taxfilers who reside with their parents, have been placed into one.age range.

In reviewing column 4 of Table 2 (i.e., % ratio), it can be

noted that the coverage of the T1FF to the 1986 Census was approximately 50% or higher for age ranges under 60. The T1FF coverage of the 1986 Census population declined more rapidly for the population 65+.

3.3. Coverage of Aggregate Sources of Income

In conducting the 1986 Census of Population, sources of income data were collected for the 1985 calendar year. Table 3 contains a sources of income comparison between the 1985 T1FF and the 1986 Census.

For both data sources, the largest component of income was wages and salaries. The T1FF estimate was 96% of the Census estimate. In the government transfers section of Table 3, considerable variability existed, primarily because some transfer payments were either not subject to taxation or were received by individuals with low incomes who did not file a Tl.

3.4. Income Distribution Coverage

Table 4 includes a time series comparison of median incomes between the T1FF and the Survey of Consumer Finances (SCF) for the period 1982-87. The T1FF medians were lower for all years. Moreover, the medians were about 95% for the first four years. In the

fifth and sixth years, the medians declined to about 92%. This decline can be partly attributed to the introduction of a refundable Federal Sales Tax Credit. This credit was available to

individuals and families with low incomes some of whom may only file a tax return to obtain this credit.

Since it is generally assumed that taxfilers have higher incomes than non-taxfilers, one would expect the SCF to have lower

medians since some respondents would have low incomes and not file tax returns. Clearly, these findings are inconsistent with such an expectation.

3.5. Dimensionality of Variables

Since any single administrative record (for example, the Tl) has a specific and narrow application in program administration, the range of data variables might be judged inadequate as a source of social data. Although the T1FF data are oriented to the income tax system, Table 5 indicates (mainly by reference to the footnotes) that some comparability in the variables existed between the 1985 T1FF and the 1986 Census.

From Table 5, it is clear that the T1FF data,lack the richness of the census data. T1FF has a low coverage of non-taxable sources of income and a low coverage of those taxable sources of income received by low income persons who do not file tax returns.

4.0. Major Directions for 1989+

Two new initiatives have been started.

o The development of a pilot Longitudinal Administrative Database (LAD) to enable research studies of poverty/welfare/income dynamics in Canada for the period 1982-86. The LAD was designed as a 10% sample to parallel the Panel Survey of Income Dynamics (PSID) that was begun by

the Survey Research Center, University of Michigan about 20 years ago. (Duncan, 1984)

o The development of an Administrative Record Consolidation File (ARC) through the linking of multiple records on a sample basis for the purpose of (a) improving the coverage of the population and (b) improving some of the variables on the taxfile.

5.0. Summary and Related Observations

The T1FF data possess some positive characteristics. The data are annual and small area estimates can be produced. Furthermore, if 95% is high coverage, the comparisons in this paper have indicated a fairly high coverage of the population by the T1FF.

73

One potential benefit of administrative data seems to lie in the domain of longitudinal databases. While longitudinal surveys

can only be created in the future, based on current decisions and funding, longitudinal administrative databases can be created retrospectively, based on current decisions and funding. For example, a decision was made in late 1988 to begin creating a longitudinal database for the period 1982 to 1996. While the database has not yet been completed, early indications are that the database will be a source of useful information for the development of social policy and and for the analysis of income dynamics.

To conclude, this paper has been prepared to illustrate some findings that may not be widely known. In preparing this incomplete report on an evolving new paradigm in Canada, it is hoped that members of the research and statistical community will provide comments and insights that will improve and stimulate the continued evolution of this work.

References

Duncan,, Greg J. (1984) Years of Poverty, Years of Plenty. Survey Research Center, Institute for Social Research, The University of Michigan, Ann Arbor, Michigan. 1984.

Statistics Canada. (1979) Final Report of the Task Force on Administrative Data Development. UnPublished. Ottawa, Canada. March 16, 1979.

Statistics Canada. (1982) 1981 Census Dictionary. Statistics Canada, 1981 Census of Canada. Catalogue Number 99-901. Ottawa, Canada. May 1982.

Statistics Canada. (1988) Postcensal Estimates of Families, Canada. Provinces and Territories. Catalogue Number 91-204. Ottawa, Canada. 1988..

Statistics Canada. (1989) Total Income: Individuals. 1986 Census of Canada. Catalogue Number 93-114. Ottawa, Canada. March 1989.

Statistics Canada. (1990) Bibliograrphy: Research Papers and Reports, Small Area and Administrative Data Division. Small Area and Administrative Data Division, Statistics Canada. Ottawa,

Canada. Unpublished. February 22, 1990.

Statistics Canada. (Annual) Family Incomes: Census Families.

Survey of Consumer Finances. Catalogue No. 13-208 (Annual).

Ottawa, Canada. 1989.

Vigder, Michele and John Leyes. (1989) Administrative Data to

Mid-Decade Census Data Comparison. Small Area and Administrative

Data Division, Statistics Canada. Ottawa, Canada. Unpublished.

January 31, 1989.

DISCUSSION

Gerald Gates

U.S. Bureau of the Census

The theme of these papers by Fritz Scheuren (IRS) and John Leyes (Statistics Canada) is shifting the paradigm of census taking to allow for more frequent detailed information for small geographic areas at reasonable costs. Scheuren points to two weaknesses in the U.S. census taking process -- 1) the increasing costs of enumeration and 2) the increasing obsolescence of the information between censuses. He discusses new paradigms that have been proposed by Kish and others which employ rolling samples and other techniques to obtain mote frequent small area data. His primary focus, however, is on administrative registers that could be modified to serve a census function as well as their intended administrative uses. His intent is to frame the debate for feasible options that will lead to a lot more scientific research on this topic.

Leyes addresses the census paradigm in terms of research undertaken by Statistics Canada using administrative records. Primarily, he describes the development of a family tax file representing approximately 95% of the census in terms of population covered. He looks at coverage of this file in comparison with the census; with surveys conducted by Statistics Canada; and with administrative data maintained by other agencies. Finally, he describes a project to develop a linked administrative file that would allow Statistics Canada to estimate the characteristics of the population missed in the family tax file. The work Leyes describes has implications for shifting the census paradigm to address cost, accuracy, and timeliness issues.

Turning first to the Leyes' paper, I have a few specific reactions to the role Statistics Canada plays with Revenue Canada and with the content and coverage of the family tax file. The family tax file could only have been created with a great deal of cooperation from Revenue Canada. The Canadian tax form contains demographic characteristics such as age, sex and marital status that have no practical tax program application. In addition, all information from the tax return is available to Statistics

Canada -- this is not the case in the U.S. Another major difference between the two countries is the negative income tax provisions in Canada which increases coverage of the tax file

(from 89% in the U.S. to 96% in Canada). Despite these advantages, the Canadian tax form, like the IRS Form 1040, contains mailing address rather than physical address.

I also found the Canadian work on record linkages to be quite impressive, especially as it relates to creating retrospective longitudinal databases to deal with emerging issues (e.g., income and health care issues relating to the elderly). Also, these linkages permit, as Leyes states, adjustment of the family tax file for undercoverage. This feature allows Statistics Canada to use the family tax file as an independent source for producing

population estimates between census years. Since these linkages are only done on a sample basis due to privacy concerns, their utility is diminished somewhat. In the current U.S. situation, the reduced coverage and content of the Form 1040 file makes l00% record linkages critical, while similar privacy concerns need to be addressed. (I should add that Form W2 earnings records could improve the coverage possible with only 1040 tax returns, but this will continue to miss nonworkers and omit some of the detail available on the 1040.)

Scheuren's paper raises some important issues regarding the need for research on alternatives to the traditional once-a-decade enumeration. I complement Fritz on his persistence over the years to explore traditional census alternatives. His current paper addresses the need for a census alternative to deal with "problems" facing the 1990 census in terms of costs (low mail response rates) and increasing data obsolescence. Although administrative records remain his primary focus, Fritz sees a need for research in other areas, such as rolling samples. He believes that rolling samples offer real promise if they can be integrated into current ongoing

survey operations. Although the "rotating" sample techniques have been proposed for 2000 census planning (Herriot, Bateman, McCarthy, 1989), little research has been done and we have no plans to incorporate this technique into the current surveys. There are several reasons for this which reflect the different goals of current surveys and intercensal estimates

- a rolling design will create inefficiencies because of increased interviewer travel (and reduced workloads) which will come from abandoning Primary Sampling Units (PSUs) in favor of more geographically disperse samples;

- survey procedures that, as a cost saving feature, incorporate an alternative to the traditional first time personal visit, could result in lower response rates (telephone) or delays in the interviewing process (mail)

- for surveys such as SIPP, the sample may be too small to spread out geographically;

o sponsors may not want long form questions added to their questionnaires nor want intensive sample in areas with small population.

The second major point I would like to address regards, as Fritz put it, our "missed opportunity" to use the 1990 census as a proving ground for the use of administrative records in the census process. The 1990 census Research and Experimentation (REX) program considered many applications for administrative records including all uses made in 1980 plus an administrative records census and a coverage improvement program designed to enumerate parolees and probationers through state administrative records. All of these uses were abandoned because of resources available and the expected minimal improvements given the costs. (The parolee/probationer operation was accomplished by parole officers who distributed and collected questionnaires from persons in their charge.) An additional use, which was tested on a small scale as part of the 1988 dress rehearsal, involved supplementing the Post

Enumeration survey (PES) with names obtained from administrative records in order to improve the PES as a coverage measurement tool. (Wolfgang, 1989) An evaluation of this test, which will be released shortly, may encourage further research in this area.

Several administrative records uses that were adopted in the 1990 census include:

- use of local lists of shelters and street locations to assist in enumerating the homeless;
- use of vendor lists for developing the mail register;
- macro-level consistency checks for content evaluation;
- encouraged use by local jurisdictions as a way of improving outreach activities.

Like Fritz, I believe that more extensive use of administrative records, in a productive way, will require changing administrative records. But, it will take more than that. It will take institutional changes in the way administrative agencies view their role in the census statistical process.

By way of tying this challenge to the future research activities of the Census Bureau, allow me to expand slightly on Fritz' paradigm analogy and relate it to the environment in which we operate. Both Leyes and Scheuren see administrative records playing a key role in shifting the "census" paradigm. Under this assumption, I suggest that, rather than a single census paradigm, there are actually three interrelated paradigms that require equal consideration. These are the once-a-decade enumeration, intercensal population estimates, and administrative records information systems.

Before we consider approaches to shifting these three paradigms we need to think about the role the public and bureaucracies may play. We need to consider the social contracts that exist between the government and the American people. The statistical agency has a specific obligation to census respondents to ensure privacy (confidentiality) and reduce burden to the extent possible. In addition, the statistical agency must fulfill its obligation to the American taxpayer to use its resources in the most efficient manner in providing the information needed by society. Balancing these tradeoffs will determine which direction the paradigm shift takes.

Shifting the administrative records paradigm also requires a new partnership between Federal agencies and, possibly, between Federal agencies and the states. The administrative agency must accept new unrelated tasks that are not part of its primary mission. Traditionally, agencies avoid taking on tasks that differ significantly from those that are at the heart of the organization's mission. (Wilson, 1989) Even within an administrative agency, the statistical functions often take a back

seat to administrative functions. Despite laws and additional funds that reflect these new tasks, when push comes to shove, the primary mission (in the case of the IRS, collecting taxes) will most likely win out.

A census reliance on administrative records requires a commitment by the administrative agency to the census function which heretofore has not existed. Where information is lacking, such as physical address and household relationships, change must be encouraged. Where change in the administrative process could negatively affect the census use, accommodations must be made. In the past, changes have occurred but they have not always been anticipated or beneficial. For example,

- Physical location information was added previously to the Form 1040 by the Census Bureau for the General Revenue Sharing Program.

- The 1986 Tax Reform Act required the IRS to collect SSNs for children (a plus) but eliminated the personal exemption for persons 65 or older (a negative).

o The SSA recently introduced a program of assigning SSNs to infants at birth using state birth records. Despite Census Bureau objections and concerns of its own statistical office, SSA did not require that race of child (or mother) be part of the application process.

If we assume that planning for paradigm shifts is good which I think we must -- then we need to consider, as Fritz suggests, which options are feasible. First let me discuss options as they relate to the traditional census. The basic Constitutional

requirement for apportionment requires an actual enumeration every ten years. Seven items are requested from each resident to provide: 1) the basis for the apportionment of Congress; 2) a sampling frame for use in the next decade; and 3) a base for developing intercensal estimates. To obtain this information from administrative records (i.e. an administrative records census), may

require a Constitutional amendment in addition to changes to the way administrative agencies do their jobs. Research on this aspect should concentrate on the most useful sources of information with the least amount of change required.

A second component of the census consists of the housing questions asked of every household. The Census Bureau is exploring the possibility of obtaining this information in future censuses from records of city or county tax offices, accessors offices, or recorders offices. Such an option has the potential to reduce burden and costs of census taking while offering comprehensive coverage of the nation's housing. One of the key requirements for such an operation would be fostering interest in the local jurisdictions to change/standardize their information systems to maintain the items needed for the census. This could be done by promoting the changes as an improvement to existing administrative systems and as a rich source of data for administering housing related programs. In this way, we win acceptance for the changes needed for statistical purposes through the administrative benefits

they provide.

The final component of the census is the long form sample questions. This component provides a source of detailed information for small geographic areas -- but only once a decade. As Scheuren suggests, these data could come from a rolling census design in the event that the basic census (count) is done through administrative records, but there are many problems as I have noted.

The intercensal estimates paradigm is certainly tied to the census paradigm and any change to the census will most likely necessitate changing the way we do intercensal estimates. The current population estimates program was a byproduct of the General Revenue Sharing Program. We will evaluate alternative designs in the years ahead to see if the current program is meeting the needs of users. The work of Statistics Canada on developing family tax files definitely needs to be considered. In addition, recently proposed legislation would put greater reliance on currently available population estimates for funds allocation formulas which

will in turn put pressure on the Census Bureau to expand the utility of these estimates. A possible alternative which is being

given some consideration by the Census Bureau would involve conducting a large sample survey at mid-decade and modeling the results to administrative records linked to TIGER geography. The administrative file could be constructed by linking the tax returns obtained by Census with the social security number applicant file

to be obtained from the Social Security Administration (assuming we can address the privacy issues).

In conclusion, greater reliance on administrative records in the census process needs public acceptance and a commitment from all those affected to make it work. Perhaps the increasing costs and respondent burden involved in traditional census taking will encourage this change. Scheuren and Leyes have shown us some options. We will need to explore these and others -- and fund the necessary research -- so that, as we move into the 21st Century, we

are able to avoid the pitfalls and take advantage of the opportunities that lie ahead.

References

Herriot, R.; Bateman, D.; and McCarthy, W. (1989). "The Decade cade Census Program -- New Approach for Meeting the Nation's Needs for Sub-National Data," to appear in American Statistical Association Proceedings, Social Statistics Section.

Wilson, J. (1989). Bureaucracy -- What Government Agencies Do and Why They Do It; New York: Basic Books, Inc., p. 190.

Wolfgang, G. (1989). "Using Administrative Lists to Supplement Coverage in Hard-to-count Areas of the Post-Enumeration Survey for the 1988 Census of St. Louis." Proceedings of the Section on Survey Research Methods, American Statistical Association.

DISCUSSION

Edward J. Spar

Market Statistics

The Scheuren paper, is provocative and challenging. At the same time, some of the ideas presented here should be challenged back. For example, Scheuren mentions how expensive the decennial census has become - $10 per capita. But based upon what is this expensive, in other words, as compared to what? If each individual, has to spend about $1 a year for the decennial census, is this still considered to be too expensive? Maybe we should have a check off box on the 1040 form for those who wish to contribute

a dollar to the census instead of presidential election campaigns. Money better spent.

Scheuren also points out the problem of the decline of public cooperation. However, when all the bodies are counted, what figure makes a successful census. In 1980, 98.6 percent of the population was counted. Let's say that this time 96.8 percent of the population is accounted for. Does this make the decennial census effort a failure? This will depend upon the differential undercount. Should we begin to find other ways to reach people based upon this? We will still have for the very most part usable small area data to work with. Most decisions will not change at all if the response rate does not decline drastically. Perhaps adjustment will adequately solve much of the undercount problem.

We should certainly accept the possibility of the need for "paradigm shifts". But there seems to be a problem. The paper tells us that the rolling sample approach and the use, of administrative registers just won't do the job that's needed, and all things being equal, might even be more expensive.

If accurate data are needed not only for redistricting and reapportionment, but the allocation of funds for over 100 federal programs, and if local communities need information to update their plans and allocations, you immediately have to fall back on some intensive decennial census activity. And what about private sector uses? Correct market decisions based upon detailed information is still what pays the bills, including the tax bill. If you eliminate detailed information for local areas, efficiency will decline, which is something we as a nation cannot afford.

As Fritz knows, I'm a very strong supporter of the use of administrative records for making intercensal estimates. And it has been shown in this country, and in Canada as the next paper shows, that excellent work can be done in linking administrative data sets.

Therefore I believe that our best approach so far is not to throw out the present paradigm. Instead, we have to find ways to convince the American people that they have an important stake in knowing what their about. Further, we have to convince the policy makers that once in ten years is far too infrequent, a point that Scheuren makes quite well. Also, we mustn't abandon the concept of a quinquennial census, and we have to convince the policy makers that more intercensal work is needed.

For the first time in many years, you, the statistical agencies have a special opportunity. Over the years, there has been no one in very high circles who had a real interest in statistics and was also close to the decision makers. At present, the Chairman of the Council of Economic Advisors has the ear of the president. We know that he believes in the need for timely accurate data. Therefore, the Federal statistical system needs his support and you should ask for it.

On to the Leyes paper, which was a pleasure to read. This paper portrays a cogent attempt to build a file over time which

will eventually yield excellent information between census efforts. However, the Canadian Privacy Act seems to limit the use of these data.

Statistically, however, this is kind of model where different data sets are linked, that we in the United States should explore to make better intercensal estimates. Perhaps this is where the paradigm shift should take place. Finally, I wonder if the private sector in Canada has taken advantage of these files for marketing purposes? How does the private sector in Canada interact with these data, if at all?

Two points on both papers. First, both discuss the inability to generate household information. I think that this would be harmful to both the public and private sectors, and I urge more work be done to solve this shortcoming.

Second, the private sector has developed many linked files, some good, some bad. There are claims that over 80% million households can be reached with at least one of these files, and demographic data are attached to these files. I suggest that your

agencies, at the very least, learn what has been done in the private sector and maybe take advantage of it by getting us all together and sharing our knowledge.

Session 3

SURVEY COVERAGE EVALUATION

CONTROL MEASUREMENT, AND IMPROVEMENT OF SURVEY COVERAGE

Gary M. Shapiro

Bureau of the Census

Raymond R. Bosecker

National Agricultural Statistics Service

I. Introduction

Coverage errors can cause serious biases in estimates based upon sample survey data. Undercoverage may be substantial in many surveys, especially of selected subpopulations. For example, the

estimated undercoverage of Hispanic males aged 14 and over is 23 percent in the Current Population Survey (Hainer et al., 1988). In economic surveys, new businesses may be missed at a higher rate than older ones. If the characteristics of the missed portion of the population are very different from those of the covered portion, serious biases in the survey estimates for the total population will result.

This paper is a condensation and editing of Survey Coverage, Statistical Policy Working Paper 17 (U.S. Office of Management and Budget, 1990). The 115-page working paper was prepared by the Subcommittee on Survey Coverage of the Federal Committee on Statistical Methodology. Subcommittee members are Cathryn S. Dippo (Co-chair), Gary M. Shapiro (Co-chair), Raymond R. Bosecker, Vicki Huggins, Roy Kass, Gary L. Kusch, Melanie Martindale, and D.E.B. Potter. Robert Casady, Charles Cowan, John Paletta and Richard Pratt also wrote parts of the working paper. This paper has numerous unattributed quotes from the full working paper. Although the authors of this short paper accept responsibility for all errors, credit for the good ideas and concept of the paper

belongs to all subcommittee members. We would also like to thank Melanie Martindale and Vicki Huggins for their useful comments on this paper and Cora Wisniewski, Sue Chandler and Bessie C. Johnson for their typing.

The purpose of both this paper and the full report is to heighten the awareness of survey program planners and data users concerning the existence and effects of coverage error and to provide survey researchers with information and guidance on how to assess and improve coverage in sample surveys.

This report utilizes a broad definition of coverage error. This is defined to include all possible sources of error which are not classified as observational or content errors (U.S. Department of Commerce 1978).

Section II of this paper discusses selected major sources of coverage error. IIA discusses errors which might occur before the first stage of sampling and IIB those that might occur after the first stage. Issues associated with the creation and maintenance of sampling frames, the choice of sampling frame and strategy, field listing and interviewing are included. Section III discusses selected methods for preventing, reducing and evaluating coverage errors.

II. Major Sources of Coverge Error

A. Sources of Coverage Error Before Sample Selection

(1) Conceptual Issues -- The importance of thinking carefully about the research goals, concepts, and targeted populations) for a survey cannot be overemphasized. Coverage errors can be inadvertently designed into a survey from the beginning by incorrect specification of the concepts to be measured or the

population(s) to be targeted by the survey. Vague definitions of populations and concepts tend to create coverage errors because they lead to inappropriate unit inclusions on, or exclusions from, a frame and even to naming a population which cannot be adequately represented by a frame.

(2) Frame Construction -- Once a decision is made concerning the target population, either the sample design must be based upon available sampling frames or a frame must be constructed specifically for the study. Dalenius (1985) notes the following three important properties of a frame:

- Makes it possible to compute estimates concerning a population which is sufficiently "close" to the target population.

- serves to yield a sample of elements which can be unambiguously identified.

- Makes it possible to determine how the units in the frame

are associated with the elements of the (sampled) population.

The first stage of sampling is usually dependent upon a frame consisting of a physical listing of units. This may be a list of names of individuals, establishments, institutions, counties, cities, streets, etc., or a list of numbers attached to city blocks, land area segments, houses, pages, or any number of unique,

definable entities. However, as Kish (1965, p. 53) notes, a "Frame is a more general concept: it includes physical lists and also procedures that can account for all the sampling units without the physical efforts of actually listing them." Deming (1960) cites one exception to a list of units. This occurs when a watch is used to sample time intervals during which customers leaving a store are interviewed.

The units listed in the initial frame may not correspond to

the units about or from which information is sought. Often, additional frames are needed for successive stages of sampling in order to progress from available sampling units to the units to be contacted or measured. For example, areas may be selected from a listing or array of all blocks in an area frame. Housing units inside sampled areas may then be listed and sampled in order to achieve a listing of persons to be sampled that are members of the target population from which information is sought.

A more complex example is the procedure for selecting items to be priced in the Consumer Price Index. The sample of priced items is selected from items sold by a sample of outlets which, in turn, was selected from a list of outlets created from information provided by interviews with consumer units in addresses sampled from the decennial census, new construction permits, and area listings. In this case, interviews are conducted in a sample of housing units to create a sample frame of establishments, not a population frame, from which a sample is selected. Within the sample outlets, probability methods are used to select increasingly more detailed classes of goods until a particular item is selected. A complete list of all the items available for sale is never

constructed.

(3) Frame Errors -- Kish (1965) states that a "frame is perfect if every element appears on the list separately, once, only once, and nothing else appears on the list," and classifies possible frame errors into four types: missing elements clusters of elements appearing on the list, blanks or foreign elements, and duplicate elements.

Missing elements is the frame error which causes greatest concern. Because they are missing, no examination of the sample from the frame will reveal the nature of that component of the population. Often, conclusions are erroneously extended beyond an incomplete frame on the tenuous assumption that missing units are like or very similar to those represented on the frame.

The initial sampling units may contain clusters of subunits which must be incorporated into the sampling design. An example is a list of farm operator names of which the vast majority represent a one-name/one-farm relationship but some represent a one-name/

multiple-farm relationship. In this situation, there is a distinct

possibility for coverage error unless the interviewer has been thoroughly trained.

If a frame is created or an existing list modified for a particular one-time survey, elements on the list which are blank or are not members of the population of interest should be removed. If they are not removed, those appearing in the sample must be identified and properly handled in the survey process.

Duplication of units on the frame may result in overcoverage, i.e., some members of the population are represented more than once. Population totals may then be overstated and means could be biased.

4) Frame Maintenance -- Frame maintenance procedures are

discussed as they relate to the classes of coverage error just described. These procedures can be classified as follows:

- Adding new frame elements or births,

- Eliminating or identifying inactive frame elements or deaths,

- Correcting misclassified frame elements,

- Identifying existing frame elements no longer in scope, or in scope for the first time, and

- Determining whether or not elements have combined with other elements or have split from existing elements (e.g., change in ownership, mergers, and divestitures in an economic setting).

When the research population is dynamic, it is important that a frame which represents it be updated to reflect births. Section

III discusses several methods for doing this.

The failure to identify deaths on a sampling frame does not necessarily imply a bias, since any deaths sampled would be representative of the universe of deaths. But, biased sample estimates can result if an inactive element is sampled and imputed for when no response is obtained.

A problem associated with many frames is not that elements are missing, but that they are misclassified or are not classified at all with respect to one or more variables. This assumes importance if the variable or variables that are misclassified determine either the elements eligible for sampling or the subpopulations for which estimates are produced. Housing occupancy status (vacant or occupied), geographic codes, SIC codes, etc., are examples of such variables.

Closely related to the problem of misclassification is the

problem of out-of-scope elements, i.e., elements that if properly classified would not be part of the universe of interest. As with death elements, the presence of out-of-scope elements on a sampling frame does not result in any biased sample results should they be sampled (assuming the sample process identifies them as out-of-scope).

The composition of elements comprising a frame will often change over time. This is especially true for economic-based frames, where, for example, individual plants are bought and sold by companies, two or more companies merge, or companies divest. From a coverage point of view, ownership is important because the continued sample status of a sold establishment often depends upon the status of the buying company.

B. Sources of Coverage Error After Sample Selection

The full Survey Coverage report discusses three broad kinds of error occuring after the initial selection of a sample from a

frame: (1) incorrect association of sampling to reporting unit; (2) editing errors; and (3) other nonsampling errors. We discuss only the first of these in this paper.

Misclassification of occupied housing units as vacant units is a frequent type of classification error in household surveys. In many surveys, the population of interest consists of occupied housing units, but the frame consists of other types of units as well. In the Current Population Survey (CPS), for example, an interviewer is generally given specific addresses for interview. When an interviewer is repeatedly unable to find anyone home at an address (s)he must classify it either as a vacant noninterview (out of scope) or as a noninterview unit occupied by persons eligible for interview. In October 1966, the CPS reinterview concentrated on measuring this type of coverage error (U.S. Bureau of the Census 1968). This research revealed that more than 10 percent of the units classified as vacant were actually occupied by eligible persons.

In two separate evaluation projects in the 1970 Decennial

Census, 11.4 percent and 16.5 percent of the units initially enumerated as vacant were misclassified (U.S. Bureau of the Census 1973).

We believe that error in listing persons within interviewed households (within-unit) is the most serious source of coverage error occurring after sample selection. Alexander (1986) has estimated that within-unit error results in overall undercoverage of four percent for persons 12 and over in the National Crime Survey. Within-unit error is probably more serious for blacks and Hispanics. Hainer et al (1988) point out that in the CPS, black

female undercoverage is close to the overall undercoverage of seven percent, but black male undercoverage is about 20 percent, suggesting that most of this undercoverage results from within-unit error.

There are several instances in which authors have speculated

on large biases caused by within-unit error. One example of this is discussed by Hainer et al. (1988): "... Cook (1985) presents evidence suggesting that the National Crime Survey may underestimate the number of gun assaults by as much as one-third. He offers the explanation that the National Crime Survey does not adequately cover the kinds of people criminologists believe are most likely to be involved in the life of the streets (including participation in criminal activity...)" (Cook 1985, see also Martin 1981).

Hainer, et al. (1988) discuss at length the ethnographic research that has been done on household survey coverage. They suggest there are two main causes of respondent reporting error resulting in missed persons:

- Some people, especially black and Hispanic males, are deliberately omitted because of potential loss of household income if their presence in the household were known to authorities.

- There is a lack of correspondence between survey

definitions of household residency and how people actually live.

III. Methods for Dealing with Coverage Errors

The previous discussion focused on sources of coverage error in selecting and maintaining sampling frames. Solutions to problems arising from the limitations of available frame sources are a major challenge to the survey design statistician. Some options, however, are available for dealing with coverage error. The options discussed are: Questions to specify concepts, current sampling frame, updated frame for births, random digit dialing, multiple frames, reinterview, estimation procedures, and evaluation methods.

A. Preventing Incorect Specifications of Concepts

To avoid coverage errors caused by incorrect specifications of

concepts, it is useful to ask a series of questions:

o To what population(s) of units does this problem refer?

Distinguish among populations about which information is sought, those which will be frame units, and those which may be reporting units, if different from the frame units. For example, suppose one wished to do research on "the scholastic achievement (as measured by grades) of children of recent immigrants." In this case, "children of recent immigrants," more suitably specified perhaps as, "persons aged roughly 5 to 17 enrolled in Grades 1 through 12 of the U.S. public schools and living in a household in which at least one related head has been resident in the United States 5 or fewer years," would be the population about which information is sought. However, it seems likely that one might need to construct two or more frames in order to reach this population.

One of the frames might have U.S. public schools as units, while another might consist of residential addresses to be screened. In this example, reporting units might well consist of two groups, school recordkeepers and parents or guardians.

o Is (are) this (these) populations) observable or potentially measurable? How?

Continuing from the example above, one can see that the suggested specification of "children of recent immigrants" takes account of some of the presumably unobservable "children of recent immigrants", such as those who may be homeless and those who may not be currently enrolled in school. Among recent immigrants, those who entered the country illegally may not be observable, as well as those who died following entry, leaving school-age dependents. Sources for obtaining U.S. public schools and residential addresses might be lists from various agencies. Thinking through all possible categories of the populations of interest should reveal those subsets which cannot be measured or reached; those whose measurement (observation) might be achieved;

and those which seem reachable with some existing or proposed methodology. Thus, the "children" may be reached by means of a household survey, school survey, and/or institutional survey (hospitals, orphanages).

o Are there one or more subsets of this (these) populations) which cannot be measured/observed in some way? What are these? Would they ever be measurable?

Continuing the example of "children of recent immigrants of some of the unobservable components of the populations discussed have already been mentioned. The potentially measurable components might be those who cannot be reached now but who might be reached using a methodology that is prohibitively expensive, such as scanning all death certificates or other sources of information to identify deceased recent immigrants. Thus, it may be useful to distinguish the inherently unobservable from the practically unobservable components of populations of interest.

o Does time enter into the answer to one or more of the questions above, in the sense that the measurable populations) may change or may have changed?

Continuing the example of "children of recent immigrants," one may find that a change in a legal boundary or definition can turn "internal migrants" to "recent immigrants" or vice versa. This would happen, for example, if Puerto Rico became a U.S. state, thus solving the problem of how technically to classify migrants to the mainland, who would become "internal migrants". Such a change might force a redefinition of the size and location of the populations of interest.

o Have previous efforts been made to build a frame of this (these) population(s)? What problems were encountered in frame construction? Was one of these faulty conceptualization? Which of these problems has been solved?

This series of questions focuses on the need to locate previous research, to attempt to contact those who designed and conducted the research, or to obtain procedural histories about it and to evaluate carefully the definitions and language used by others. An assessment of previous research often reveals use of frames built for other purposes by still earlier researchers, especially when the frames are very expensive to assemble. Information needed for adequate frames may now be available (such as improved school lists) due either to improvements in information processing or to changes in laws regarding availability of administrative data.

B. Current Frames and Updating Old Frames

Use of old frames can result in serious coverage problems, because births may be partially or totally excluded and other units may be misclassified. An obvious but important solution is to use current or recently built or updated frames whenever possible.

When an old frame must be used, it is important to have updating procedures to include births. One effective method for detecting new units is to periodically canvass the existing frame elements. As an example, all of the larger multiunit companies and some of the smaller companies on the Standard Statistical Establishment List are canvassed on a yearly basis. Companies are questioned as to whether or not they have started new operations.

A second method of identifying new units results from coverage maintenance operations performed for samples selected from the frame. As part of the questionnaire administration process in nearly all surveys, inquiries are made about the status of the sampled units and whether any changes in their status have occurred

since the last data collection period. Although the inquiries are targeted to sampled units believed not to be births, sometimes incidental information about other units (including births) can be

obtained.

Several methods can be used for including new units in household surveys. The Bureau of the Census includes most new housing starts in its household surveys by sampling from building permit files. This is an efficient procedure, but building permit files do not identify illegal new construction, conversions, and new mobile home placements; nor do they identify new special places, such as dormitories, fraternity houses, boarding houses, and public housing. To illustrate, it was estimated for the 1985 American Housing Survey that approximately 25 percent of all new mobile homes were missed (Schwanz, 1988).

C. Random Digit Dialing

One household sampling method employed in an attempt to avoid omission problems is random-digit dialing (RDD) (Waksberg, 1978). The use of telephone directories as sampling frames often results in unacceptable levels of undercoverage because they omit unlisted numbers for some nontypical portions of the population. With RDD,

a sample of telephone households is located through the use of randomly generated telephone numbers. In this way only those households without telephones are omitted. For many surveys, this could be considered a trivial exclusion. In others, differences between telephone and nontelephone households may have a profound impact on the characteristics being measured. For example, measures of poverty and income from entitlement programs would most likely be biased.

D. Multiple Frames

Coverage may be improved through the use of multiple frames. Sometimes no single frame fully covers the target population and merging independent source lists would be impractical. In this case separate probability samples from different frames can be used to expand coverage beyond any available single frame.

The application of overlapping multiple frame sampling most commonly found in Federal surveys is the use of an area frame and

an overlapping list frame. The area frame is generally designed to provide complete coverage by including all U.S. land parcels as sampling units. The list frame is nearly always incomplete (a common attribute of lists), but its use provides certain sampling efficiencies which enable the multiple frame survey to provide the same precision at a much lower cost than would an area frame survey, alone.

E. Reinterview

Reinterview can often be profitably used for both evaluation and control of coverage error. In the CPS, the regular reinterview program is able to detect misclassification of occupied housing units as vacant units, errors made in listing housing units in area segments, and errors made in missing persons within interviewed units. However, the CPS reinterview program serves many purposes and consequently fails to detect a number of these errors. A

special intensive coverage check was done in the 1966-67 CPS reinterview. This check was much more successful than regular reinterview in detecting vacant unit misclassification and area segment listing errors, but still found few instances of within-unit errors (U.S. Bureau of the Census 1968).

A type of reinterview can also be used for nonresponse follow-up. A subset of original noninterviews can be more aggressively pursued to obtain complete or at least partial, interviews, or alternatively, refusal households can be sent a very brief mail questionnaire asking why they refused and collecting basic demographic information.

F. Estimation Procedures

Estimation procedures may also be used to decrease the bias of survey estimates relative to the target population. One such procedure is the use of ratio estimation or benchmarking. The Bureau of Labor Statistics employs a benchmarking procedure to revise monthly employment estimates from the Current Employment

Statistics survey. (U.S. Bureau of Labor Statistics 1989) Sample estimates are compared each year with later summarizations of mandatory UI reports filed by employers. The UI data, which serve as a benchmark, are an aggregation from the same source as the micro-data used to construct the frame from which the sample was selected, except that the benchmark data are one year newer. Hence, the benchmark file takes into account new firms or changes in industrial classification to ensure more accurate coverage. The completeness of the UI administrative data affords the opportunity to analyze and adjust for frame deficiencies (Thomas, 1986).

G. Macro and Micro Level Evaluation

Evaluation methods to independently determine the representativeness of the sampling frame(s) used are very useful for quality control. One method of measuring the degree of frame coverage error is comparative analysis. Comparative analysis can occur at two levels. The first is a macro level evaluation, which compares known population values with totals derived from summing characteristics for each sampling frame unit. The second type of analysis is performed at the micro or individual sampling unit

level. This most often involves matching of data available from different sources for individual units.

The Bureau of the Census utilizes a macro-level approach for frame completeness evaluation called demographic analysis. With this method, demographic data from various sources are used to develop expected values for the population as a whole and by race, age, and sex to compare with the census counts.

On a micro-level basis the Bureau of the Census matches census returns against Administrative records for drivers' licenses from State departments of motor vehicles and against registers of resident aliens supplied by the Immigration and Naturalization Service.

IV. Conclusion

This paper has presented many of the major points treated in the full Survey Coverage report, whose purpose is to provide information about the types and effects of coverage error in surveys and guidance on how to assess and improve survey coverage. We found few studies, however, which actually measure coverage errors in surveys and even fewer which address the impact of coverage error on survey estimation. The paper implies that significant resources should be allocated to the conceptual and planning stages of surveys, and that procedures providing for the evaluation of coverage and for minimizing and controlling coverage error be clearly established and included in the survey design.

As to the seriousness of coverage error, the largest single source of coverage error identified in the full Survey Coverage report for an economic survey is a 20 percent underestimate in the 1988 Economic Census statistic of receipts for nonemployer establishments due to misclassification. For household surveys, large single source of overall coverage error is an estimated 4 percent undercoverage in the National Crime Survey estimates of persons aged 12 and over due to within housing unit listing errors.

(Undercoverage from this source for some subgroups is much worse.) Since we know that single sources themselves can be significant, the overall effect of all sources of coverage error on survey products is of great concern.

Several leading methods for identifying and assessing coverage error and for improving coverage have been mentioned here. The full report treats these and other methods in detail. It also provides case studies of specific Federal surveys which illustrate various frame and coverage issues.

The methods that apply to most surveys and which can lead to significant improvements in data quality are the use of multiple

frames to improve coverage at the sampling stage and weighting adjustments to reduce bias from coverage error.

References

Alexander, C. (1986), "The Present Consumer Expenditure Survey's Weighting Methods," in Population Controls and Weighting Sample Visits, Washington, DC: U.S. Bureau of Labor Statistics.

Cook, P. (1985), "The Case of the Missing Victims: Gunshot Woundings in the National Crime Survey," Journal of Quantitative Criminology, 1, 91-102.

Dalenius, T. (1985), "Elements of Survey Sampling," Notes prepared for the Swedish Agency for Research Cooperation with Developing Countries (SAREC).

Deming, W. (1960), Sample Design in Business Research, New York: John Wiley and Sons, Inc.

Hainer, P., Hines, C., Martin, E., and Shapiro, G. (1988), "Research on Improving Coverage in Household Surveys," Proceedings of the Fourth Annual Research Conference, U.S. Bureau of the Census, pp. 513-539.

Kish, L. (1965), Survey Sampling, New York: John Wiley and Sons, Inc.

Martin, E. (1981), "A Twist on the Heisenberg Principle: Or, How Crime Affects Its Measurement," Social Indicators Research, 9, 197-223.

Schwanz, D. (1988), "1985 Type-A Unable-to-Locate Rates for the AHS National Unit Samples," Internal memorandum, U.S. Bureau of the Census.

Thomas, A. (1986), "BLS Establishment Estimates Revised to March 1985 Benchmarks," Washington, DC: U. S. Bureau of Labor Statistics.

U.S. Bureau of the Census (1968), "The Current Population Survey Reinterview U.S. Program, January 1961 through December 1966," Technical Paper 19, Washington, DC: U.S. Government Printing Office.

U.S. Bureau of the Census (1973), "The Coverage of Housing in the 1970 Census," Report PHC(E)-5, Washington, DC: U. S. Government Printing Office.

U.S. Bureau of Labor Statistics (1989), Employment and Earnings, 36 (12).

U.S. Department of Commerce (1978), "Glossary of Nonsampling Error Terms: An illustration of a Semantic Problem in Statistics," Statistical Policy Working Paper 4, Washington, DC: U. S. Government Printing Office.

U.S. Office of Management and Budget (1990), "Survey Coverage", Statistical Policy Working Paper 17, Washington, D.C.

Waksberg, J. (1978), "Sampling Methods for Random Digit Dialing," Journal of the American Statistical Association, 73, 40-46.

QUALITY OF SURVEY FRAMES

Judith T. Lessler

Research Triangle Institute

1. Introduction

This paper focuses on the quality of sampling frames with particular emphasis on the relationship of the sampling frame to the overall error of survey estimates. It also presents some examples from studies that have been conducted by the Research

Triangle Institute (RTI).

The frame is a fundamental element of scientific survey research. Probability sampling involves selecting a subset of units from a finite collection of units in a manner that lets one determine the probability of obtaining that subset. The sampling frame is the finite population of units to which the probability sampling mechanism is applied. Thus, the type of frame used for a survey and any deficiencies or inefficiencies in it affect the total error of the survey estimates.

2. Definition of a Frame

The population of frame units is not necessarily equivalent to the population for which information is to be collected. In this paper, I refer to the population the survey researcher wishes to make measurements on as the target population and the individual components of that population as elements. This population may not be the same as the inferential population. For example, the National Human Monitoring Program of the U.S. Environmental Protection Agency (EPA) conducted a special study of mirex residues

in human adipose tissues (Leininger et al., 1980). Mirex is a persistent insecticide that has been used to control fire ants. Human adipose tissue specimens were collected from selected surgical patients and cadavers and chemically analyzed for the presence of mirex residues. The inferential population in this study was not the sick and the dead, but, rather all persons living in the areas subject to application of the insecticide.

Just as the target population is not necessarily the same as the inferential population, neither is the population of frame units the same as the population of target elements. Thus, noting this distinction and the role of the frame in survey sampling, I once defined a frame as follows:

> "The frame consists of materials, procedures, and devices which identify, distinguish, and allow access to the elements of the target population. The frame is composed of a finite set of units to which the probability

sampling scheme is applied. Rules or mechanisms for linking the frame units to the target population elements are an integral part of the frame. The frame also includes auxiliary information (measures of size, demographic information) used for (1) special sampling techniques such as stratification and probability proportional to size sample selections, or (2) special estimation techniques, such as ratio or regression estimation."

I like this definition because it clearly recognizes that different types of frames support different types of sampling and estimation procedures.

However, I think that it fails to recognize a key aspect of sampling frames, namely, the types of measurement designs they support. To illustrate, if a survey is to be conducted by asking questions or by gathering information from records, the reporting units are not always equivalent to the target elements. For

example, suppose we wanted to know the family income of all children who attended the Saturday afternoon swimming classes at Sometown Community Park. A sampling frame consisting of a list of all swimming classes and the times that they met would provide us easy and efficient access to the target population of children. We could go to the class and identify each child; however, this would not be very helpful because few children know their family incomes. Thus, we need to insert a key word in the above definition -- measurement -- yielding:

> "The frame consists of materials, procedures and devices which identify, distinguish, and allow access to and measurements on the elements of the target population. The frame is composed of a finite set of units to which the probability sampling scheme is applied. Rules or mechanisms for linking the frame units to the target population elements are an integral part of the frame. The frame also includes auxiliary information (measures of size, demographic information) used for (1) special sampling techniques such as stratification and

probability proportional to size sample selections, or

(2) special estimation techniques, such as ratio or regression estimation."

3. Components of Quality

Researchers who are choosing a sampling frame for a survey need to consider a number of factors when making that choice. These include:

- coverage of the target population

- efficiency of the sample designs that are supported by the frame

- effect of the frame on nonresponse errors

- types, costs, and quality of the measurement designs supported by the frame

- cost of constructing the frame

- accuracy of information on the frame

Coverage of the target population: It is widely recognized that several aspects of a sampling frame can cause bias in survey estimates. Missing target elements, inclusion of nontarget population elements, unrecognized multiplicities, and failure to account for the clustered nature of frames during sampling and estimation can all introduce bias in survey estimates.

Efficiency of sampling and estimation: The structure of the frame, the information it contains, and the quality of that information will determine the types of sample designs and estimation procedures that can be used in a survey. Simple frames lacking auxiliary information support simple sample designs; complex frames containing auxiliary information support more

complex designs, which are generally more efficient. Frames used for sampling business establishments are a good example. Lists that also include information on the size of the establishment will permit sample designs that are much more efficient than those that could be designed using a simple listing of establishments.

Effect of the frame on nonresponse errors: The type of frame that is chosen also has a major impact on nonresponse errors. Often, a frame that provides efficient access to large segments of a target population will also be guarded by "gatekeepers" who can deny access to the target elements. For example, if one would like to conduct a survey of young people aged 12-17, using a school based sampling frame rather than an area household frame will provide more efficient access to the great majority of this target population. To use such a frame, one usually needs permission from school district personnel who can, in a single decision, deny access to large segments of the target population. In a national survey, failure to obtain cooperation from the large city school districts can have a devastating impact on our ability to control nonresponse errors.

Types of measurement designs supported/cost of making measurements: The frame that is chosen for the survey also affects the types of measurements that can be made. Frames of telephone numbers using random digit dialing provide access to a very large part of the household population. Using this frame, however generally limits one to making measurements by asking questions.

One cannot weigh the person or collect a blood sample although one could, of course, obtain the person's address and make the direct measurements in subsequent visits. These subsequent visits would cost more than using an area housing unit frame because the sampled elements would be widely dispersed.

RTI recently completed a survey for the Food and Nutrition Service of participants in WIC (Women, Infants, and Children Feeding Program). Much of the information that needed to be collected could be abstracted from the WIC records; however, other

information required an interview with the WIC participant. A sampling frame that consisted of lists of WIC agencies and lists of persons served would have been the most efficient for collecting the record data; however, it would have been very inefficient for conducting the interviews. Because of this, we developed procedures for listing people as they arrived at WIC clinics for their initial enrollment into the program.

Cost of constructing the frame: When assessing the relative quality of various sampling frames, we must consider the cost of constructing the frame. A frame that includes "size measures" for the units may be permit more efficient sampling; however, it may be too costly to determine the size of the units. The money spent on constructing the frame might be better spent in increasing the sample size.

Accuracy of information on the frame: If the auxiliary information on the sampling frame is inaccurate, the efficiency of sample designs and estimation procedures that make use of this information will be reduced.

4. Examples

RTI has conducted many types of surveys using many kinds of sampling frames including area household surveys and random digit dialing surveys, as well as surveys of schools, businesses, military personnel and families, nursing homes, hospitals, and so on. We also do a number of environmental surveys, and I will describe two of these to illustrate the points discussed earlier.

4.1. Of Flowing Waters

The first example shows how a frame can influence in several ways the quality of a survey's estimates. The goal of the 1982 National Fisheries Survey was to measure the biological quality of the Nation's flowing waters. After some discussion of exactly what was intended by the phrase, the "Nation's flowing waters," the statisticians on the project turned to the task of developing an operational definition of sampling units and target elements for use in the survey. It turns out that the EPA has developed a

cataloging system in which each body of water in the United States

is segmented into well-defined units called reaches, described according to the following definition (Horn, 1981):

> "Most reaches represent the approximate centerlines of streams and extend between points of confluence with other streams. The reaches constructed within open waters are generally straight lines connecting tributary streams with assumed transport paths through the open waters."

In addition, the U.S. Geological Survey (USGS) has a system in which the United States is divided into nonoverlapping areas based upon the configuration and sizes of watersheds. There are 2,100 cataloging units (CUs) contained in larger regions called water basins or hydrologic regions. When we designed the survey, EPA maintained a River Reach File, that contained some 68,000 reaches

defined within these CUs. This file was not complete because it was estimated that the total number of reaches was around 179,000. Moreover, a clustered design was not needed to control data collection costs because the survey was to be conducted by mailing questionnaires to local fisheries biologists who were familiar with each waterbody. In addition, a very accurate (but costly) digitizing procedure for identifying reaches and for measuring their length was available. Thus, staff decided to select the sample in two stages: (1) sampling CUs, then (2) reaches within CUs using maps to identify the reaches.

We established the following operational definition of the target population:

All reaches of rivers and streams that were:

a. contained in the 48 contiguous States;

b. shown on 1:500,000 USGS maps;

c. including watercourses shown on the maps as being seasonally intermittent, impoundments, reservoirs, canals and constructed channels, and waterways; and.

d. excluding the Great Lakes and other lakes, marine waters, estuaries, and wetlands (Glauz, 1984).

One interesting feature of this definition is the specification of the map scale. The scale 1:500,000 is in inches -- one map inch for every 500,000 inches. Because reaches are defined by points of confluence, maps with higher resolution would show more reaches and maps with lower resolution fewer reaches. Smaller-scale maps were not available; thus, our definition of the target population was limited by the materials we had available for identifying its elements (given the available budget).

Measures of size were constructed for the first-stage sample by obtaining maps of all the 2,100 CUs and measuring the length of

all the eligible waterways using a map meter. Grids were drawn on the maps to facilitate keeping track of the measurements, and cataloging units were randomly assigned to the staff performing the measurements.

A first-stage sample of 302 CUs was selected with probabilities proportional to size. Within this first-stage sample, a second-stage frame was constructed using automated digitizing equipment to trace, list, and record the size of each reach in the 302 selected cataloging units. A total of 1,303 reaches were selected from this second-stage frame.

This example illustrates several ways in which the frame influenced the quality of the survey estimates. First, the materials and procedures that could be afforded for constructing the frame limited the target population to reaches that were visible on the 1:500,000 scale maps. Smaller reaches could have been identified by selecting a sample of areas and using a counting and listing procedure; however, the budget for the survey did not permit such an activity. Second, the use of size measures for

selecting first- and second-stage sampling units increased the efficiency of sampling. Third, the cost of constructing a complete list of all reaches required the use of a two-stage design.

4.2. Of Passing Time

The second example illustrates the relationship between the frame, the definition of a target population, and the measurement design. RTI recently completed the National Alachlor Well Water Survey (NAWWS) that required distributing a sample in both time and space (Whitmore et al., 1990). The goal of the survey was to estimate the frequency of occurrence of the herbicide alachlor in private rural wells used for domestic consumption. Because the water in wells is not static, sample wells could not be monitored at arbitrary points in time without introducing an unknown temporal bias into the sample. Data were to be collected over a 1-year period; thus, one of the first tasks was to decide on a definition of the target population by dividing the year into units for which it was possible to collect measurement.

A major constraint on the choice of a time period for the survey was the amount of time it would take to make a measurement. A year into months, weeks, days, hours, minutes, and so on. The lower limit would be the time required to draw and package the amount of water required for an accurate chemical analysis from the well -- a few hours. Partitioning the year into hours and selecting a sample of hours, however would have required a survey team to be at the well head standing at the ready while they waited

for the sample hour. In truth, the entire process of collecting water samples for the survey was much more complicated.

The survey team needed to contact the owner of the well, obtain his or her consent to draw water from the well, make an appointment with the resident (not necessarily the owner) for obtaining the water sample, travel to the site, identify a tap or hole for collecting the water (collection of water before any treatments was preferred), measure water temperature by running water through a

flow-through cell, continue to run the water until a stable temperature was achieved or 10 minutes had passed, fill three large sample bottles, collect an additional water sample and mix it with a stabilizing reagent, and package the water bottles for shipping. In addition, observations and photograph(s) of the well site and surrounding area were needed as were questionnaire data on water use, well characteristics, and the surrounding area. After considering the time required for the survey teams to implement the entire measurement process, we decided that (with the resources available) dividing the year into observational units smaller than a month would not be feasible. Therefore, the target population for the survey was defined as well-months.

An assumption that underlay all NAWWS estimates was that the herbicide concentrations would be stable for the entire month. Dividing the year period into smaller units would have reduced measurement error; however, this would have also resulted in more missing data because the data collection team would have had severe difficulty in obtaining the measurements at the prescribed time.

To increase the chance that the concentrations were stable for

the sample month, temporal strata were formed based upon ground-water recharge conditions. Prior information was used to classify each month into a historically low, medium, or high recharge stratum. Because the first-stage spatial sampling units were counties, temporal strata were created for each county.

References

Glauz, W.D. (1984) 1982 National Fisheries Survey. Volume II: Survey Design. FWS/OBS-84/14. U.S. Fish and Wildlife Service, Washington DC.

Horn, C. Robert (1981) The reach file: a digital base of streams and lakes (memo).

Leininger, Carol, Donna L. Watts, Charles Sparacino, and Stephen Williams (1980) Mirex Residue Levels in Human Adipose Tissue: A Statistical Evaluation. RTI Project Report, Contract 68-01-5848. U.S. Environmental Protection Agency, Washington, DC.

Whitmore, Roy W., et al. (1990) National Alachlor Well Water Survey. Volume I: Survey Design and Data Collection Final Report. RTI Project No. RTI/3895/04-03F. U.S. Environmental Protection Agency, Washington, DC.

DISCUSSION

Fritz Scheuren

Internal Revenue Service

Judith Lessler and Gary Shapiro and Ron Bosecker deserve our thanks today for their thorough "coverage of coverage." They have very ably reminded us of the important quality features of this aspect of a survey.

General

Taken together, the two papers provide a valuable summary of current practice. The papers complement each other nicely. In particular, we have been given two viewpoints today -- one, from the public sector and, the other, from the private sector of survey research. Differences in emphasis arise due to these perspectives. One example would be the degree to which frame construction is ad hoc (private sector) versus ongoing (public sector). More specifically, maintenance of frames is covered in detail in the Shapiro-Bosecker paper, but only touched on in the Lessler one.

A key issue in frame construction arises when we have a target finite population, but our real purpose is in making inferences about an ill-defined superpopulation. Judy's phrase "of flowing waters" says it all. Frame construction is part of learning what is already known before conducting a survey. It is part of connecting the measurement process with the "thing" to be measured. Coverage adjustments have this flavor of connection, too.

The cognitive research movement needs to be at least mentioned in the context of survey coverage issues, if only because of the conceptual challenges in defining the target population and the even more difficult challenge of "defining" the population of inference. Just look at the problem of within-household undercoverage, for example. Maybe Judy Lessler or our Chair, Cathy Dippo, would like to comment on these cognitive aspects, since they have been heavily involved in this emerging area.

Both speakers have constructed somewhat different taxonomies of survey coverage errors. One could profitably relate and refine their approaches; however, I found both useful as is.

On the whole, the papers do an excellent job of describing (albeit in broad terms) the main technological aspects of frame construction, maintenance and coverage. I have only one quibble: I was surprised by the complete omission of any mention of record linkage.

Finally, one last point of a general nature: the Shapiro-Bosdcker paper should whet your appetite for the larger effort conducted by the Federal Committee on Statistical Methodology (FCSM). The FCSM subgroup led by Gary and Cathy Dippo conducted an excellent series of case studies (Subcomittee On Survey Coverage 1990). These studies are, however, largely descriptive, rather than proscriptive -- a point I will turn to at the end of these brief remarks.

Quality

This two-day workshop is supposed to be about quality, so I would like to connect the present papers somewhat more to that theme than has been done already. In doing this, I want to shift the focus from PRODUCT quality to PROCESS quality and look more at how to improve the processes that we use to construct frames and conduct surveys.

At, IRS, we are following an action-oriented quality management approach advocated by Juran (1986), Deming (1986) and others. This. is in contrast to the mainstream statistical emphasis which has long focussed more on measurement and perhaps not enough on improvement. Anyway, Juran divides quality, like Gaul, into three parts:

o Planning. -- The steps to be taken to prepare, including establishing the desired level of quality (implicitly or explicitly).

o Control. -- The steps needed to implement and to achieve the desired level of quality.

o Improvement. -- The efforts undertaken to make further improvements in quality over those initially planned.

Figure A provides a generic example giving you some typical steps taken at each of these three stages of quality management. This is an approach that we, at IRS, have begun to use to help the Census Bureau avoid a repetition of the 20 percent underestimate (for 1987) in the economic census statistics on receipts for nonemployer establishments -- among the largest coverage problems mentioned in the Shapiro-Bosecker paper (Greenia 1990; Konschnik and Moore 1990).

Conclusion

Let me conclude by making some recommendations on possible next steps for a follow-up to the fine FCSM efforts to study survey

coverage quality issues:

o Complete the learning from each of the FCSM case studies by subjecting them to a checklist like that in Figure A, to summarize for each case what the quality management steps were for survey coverage. -A

o Choose the "best of the best" approaches. The Japanese, word here is DANTOTSU. This (partly subjective) step is the beginning of an initial conjecture on a prescription for potentially system-wide improvements.

o Use some of the results of this proscriptive exercise to initiate improvements and to gain (back) a deeper knowledge of the once-American-now-partly-Japanese ideas that surround the second quality revolution.

In the last session, I talked about paradigm shifts in census-

taking. I am unable to resist doing so again. In particular, I would like to refer you to an excellent article in scientific American (Gomory 1990) on two improvement paradigms: ladders and cycles. My belief is that a big -- or ladder -- paradigm shift (like cognitive methods) may not be needed in the coverage area (unlike in census-taking). But, whether it is or not, we must make better use of small -- or cycle -- paradigm shifts and learn faster from each other's successes (and failures). The Federal Committee's work, as summarized today by Gary and Ron, plus Judy's ideas, offers a platform for at least some of the improvements needed.

References

Deming, W. Edwards (1986). Out of the Crisis, Cambridge: Massachusetts Institute of Technology.

Gomory, Ralph (1990). "Of Ladders, Cycles and Economic Growth," Scientific American, June, 140.

Greenia, Nick (1990). "Sole Proprietor/ship IMF/BMF Connection: An Application of the Juran Trilogy," Statistics of Income working paper (unpublished), Internal Revenue Service.

Juran, J. M. (1986). "The Quality Trilogy," Quality Progress, American Society for Quality Control, Inc., August, 19-24.

Konschnik, Carl A. and Moore, Richard A. (1990). "EC-14, A Study of the Methodology for Removing Employer Duplicates from the Nonemployer Universe for the 1987 Censuses of Retail and Services," Business Division internal memorandum (September 19), Bureau of the Census.

Subcommitte on Survey Coverage, Federal Committee on Statistical Methodology (1990). Survey Coverage, Statistical Working Paper 17, Office of Management and Budget.

DISCUSSION

Joseph Waksberg

Westat, Inc.

1. Content of the Two-Papers Presented

The two papers present a good review of issues relating to sampling frames. Their emphasis is on coverage, but they are not exclusively devoted to coverage. They would be useful reading for anyone developing a design for a new survey, or reconsidering sampling and related methods for a continuing survey. Although much of the material in the two papers covers the same subjects, there is considerable difference in focus. As a result, the authors provide a well-balanced discussion of options normally available and considerations that should be kept in mind in choosing among

alternatives. Shapiro and Bosecker mostly describe properties of frames that affect sample designs. Judy Lessler places more emphasis on how the frames can affect measurement methods, and conversely the way measurements can influence the choice of frames. The two papers thus complement each other nicely. The papers contain definitions, properties of frames, important problems inherent in some frames, and in some cases suggestions and recommendations for dealing with the problems. I'd like to discuss in more detail several of the points made in the papers.

2. Minimizing Total Means Square Errors

The authors of both papers imply, although they do not specifically say, that efforts to improve coverage by choice of suitable frame and procedures for working with that frame, are all part of attempts to minimize the total mean square error of survey estimates. Although the minimization usually cannot be done in precise mathematical form, it is almost always part of the background thinking in developing survey procedures. Judy Lessler

discusses the relationship of the frame to measurement methods. In practice, the situation is even more complex,, with frame, measurement methods, sample design, and sometimes estimation methods intertwined. All four frequently have to be taken into account in decisions on choice of frame and intensity of efforts to improve coverage. Let me give some examples:

a. About 25 years ago,, the sample design for the CPS and the other Census-conducted national population and housing surveys changed from using area sample frames to list samples in most of the U.S. The list samples consist of the set of addresses in the preceding census, plus building permits issued for new construction since the census date. In considering pros and cons of the two types of frames, it was clear there were biases in both

systems. Building permits do not quite cover all additions to the residential stock of housing, even in

areas requiring permits for construction. In addition, there is some loss because permits cannot always be located in the building permit office. Finally, in theory, the building permit frame should consist of permits for units constructed after the date of the Census. The time period is somewhat fuzzy and permits issued in the year or so preceding the census cannot be unambiguously classified on whether they were included in the census (at least not without an inordinate effort and cost). Area frames have other types of bias. The maps Census has used over the years are frequently outdated and many are difficult for interviewers to use. In addition, experience over the years indicated that interviewers cannot locate all units in area segments and a small loss consistently appeared. This undoubtedly affects the quality of the frame although how much and in what direction are difficult to quantify. However, one aspect of the comparison of two frames is quite clear. The list sample had a smaller variance. This is because over the 10 to 15 years following each census, the measures of size of the area segments became seriously

out of date. Starting a few years after each census, the area segments became quite variable in size, and this variability increases progressively over the years. The list sample provides relatively consistent segment sizes. The change from area to list sample was mainly introduced to reduce the variance arising from variability in segment size. It appeared probable that coverage would also improve, although the evidence on this was weak.

b. Westat has carried out three cycles of the National Survey of Family Growth for Health Statistics. The sample designs for first two were based on traditional area samples. For the third cycle, the National Health Interview Surveys (NHIS) was treated as the sampling frame, and the sample consisted of a subsample of eligible persons in the NHIS in the preceding year and a half. The original purpose of this revision in the frame was to reduce the cost of the extensive screening necessary to locate the required number of eligible persons. In order to keep the screening costs in the two

earlier cycles in check, a complex sample design with variable sampling rates was necessary. The NHIS permitted the elimination of most of the variable rates resulting in substantial reductions in variances for many statistics. Although it was recognized that there would be a small loss in coverage from inability to locate some of the persons who moved after the NHIS interview, it was felt that the reduction in variances compensated for it. There was a side benefit to the procedure adopted. The

NHIS contained considerable data on social and health characteristics of the persons in the frame. This information was very useful in the nonresponse adjustment procedure.

c. Random digit dialing (RDD) is, of course, much cheaper than face-to-face interviewing, especially when screening for a target population is necessary. The difference in

cost is so great that except for the major complex national surveys requiring an extraordinary degree of accuracy and surveys requiring physical measurements, most surveys both in the government and private sectors are now carried out over the telephone. Although RDD is presumably only a sampling device and the sample persons can be interviewed over the telephone or in home visits, telephone interviewing is so much cheaper that researchers generally pick it. The frame thus influences the choice of measurement methods. It's interesting that the emergence of RDD has spurred research into the quality of telephone and face-to-face interviews, and the findings have made telephone interviewing a more respectable measurement method.

3. Narrowing Definition of Target Population

Shapiro and Bosecker mention that in some circumstances it is useful to narrow the definition of the target population to one

that permits use of a more accessible frame. In some sense, this is almost always done. Surveys using area samples implicitly define the target population as those persons who are normally reported in area samples, thus excluding the undercoverage normally found. Business surveys frequently use businesses with one or more employees instead of all businesses, etc. I'd like to discuss two aspects of a narrower definition.

3.1. Risks of Narrowing Definition

I think most researchers would agree that the redefined target population should satisfy two criteria:

a. It accounts for a very high proportion of the true target, preferably 85 to 90 percent or more.

b. Characteristics relating to the subject of the study should not be wildly different in the narrower population and the missing piece.

The second criterion is quite important. It's not always recognized that even if the missing part is a small part of the

inferential population, in some cases it can have big effects. Let me give some examples.

RDD telephone surveys are probably the most common method by which a population is restricted to permit use of less expensive sampling and interviewing methods. About 93 percent of the U.S. population live in telephone households, so that the first criterion is satisfied. The extent to which the second criterion is satisfied depends on the statistic being studied. For example, in examining the feasibility of using RDD for a study of school drop-outs, the following results emerged. Figure 1 shows drop-out rates in telephone and nontelephone households for 14-21 year old youths. The shaded and cross-hatched boxes represent all drop-outs, and youths who dropped out in the past year. It can be seen

that drop-out rates in nontelephone households are about five times the rates in telephone households. The discrepancy is large enough to substantially affect the total, even though the nontelephone households only account for seven percent of all household. In fact estimates of drop-out rates from telephone households alone would understate the actual drop-out rates by about 25 percent. These estimates can be improved somewhat by post-stratifying the telephone household results, but they still seriously underestimate the true drop-out rates. Figure 2 shows drop-out rates for telephone households as a percentage of drop-out rates for the total population, and similar ratios when post-stratification is used to compensate for known deficiencies in using telephone households as a surrogate for all households. The post-stratification cells comprised single years of age, race/ethnicity, and highest grade attended by the head of the household. As can be seen, post-stratification improves these rates considerably. The ratio for-total drop-outs goes from 77 to 85 percent, but the rates are still much below the actual numbers.

Telephone households showed up much better for other

statistics studied in the same feasibility study. An analysis of enrollment in education programs for three- to five-year olds showed only trivial bias in restricting a study to telephone households. Figure 3 shows ratios of enrollment rates in telephone households to all households. As can be seen, post-stratification practically eliminates whatever bias exists in the data.

Thornberry and Massey similarly report wide differences among health-related items in the extent to which telephone households can be considered to represent all households. For the vast majority of items, there is no problem, but problems exist for items related to income. For example, estimates of the number of persons with private health insurance would be overstated about four percent if it were based only on telephone households. Most other health items would be affected only slightly.

3.2. Population for Which Estimates Are Prepared

When a survey uses a frame that does not include the total target population, there should be a clear and unambiguous statement on how the sample was selected. However, the estimation methods should attempt to adjust the narrow population so that inferences can be made about the broader population. Some researchers feel that there is something wrong in expanding the results beyond the boundaries of the frame. I don't think it makes any sense to tell data users who are interested in a specific population, that because it is cheaper or easier you've done a study of another group and they can't infer anything about the population they're interested in from the study.

Of course, no one would make such a strong statement. However, there is an implication that the results tell you about the inferential population but as a scientist you're not allowed to say so. It seems to me that since the only reason for having done the survey was to shed light on the inferential population, it makes sense to do whatever is necessary to produce the best estimates you can for that population. This is, in fact, a

commonly accepted procedure. The weighting or imputing procedures used to reduce nonresponse biases implicitly assume that one wants to produce statistics for the total rather than the respondent population. Similarly results of telephone surveys are usually inflated up to the level of the full population.

There are some real dangers in not taking the trouble to produce estimates for the inferential population. Let me cite an example where even the producers of the statistics forgot the statistics referred to a narrow population.

In November 1989, the Census Bureau issued a report on the Black population in the U.S. One of the statistics cited in the report was that the black female to male ratio was l00 to 88 compared to 100 to 96 for whites. The difference is startling, and if true has serious social implications. However, the text statement of this statistics is followed by a sentence which mentions that the ratios may be affected by greater census undercoverage of males than females. Elsewhere in the report is a footnote stating that the numbers reflect only the civilian

noninstitutional population. The term may be affected is a gross understatement of the effect. If one takes coverage and institutional population into account, the discrepancy in the sex ratios between blacks and whites is cut by more than half. The full report gave no hint that the sex ratios are affected that much by these two factors. Furthermore, by the time a press release was issued by the Bureau of the Census, the fine line between the population actually covered in the CPS and the total population was lost, and the numbers were described as reflecting the difference between the total black and white population. The only way one can avoid these kinds of misinterpretations is to make the best

adjustments one can to have the data reflect the population that readers of the report assume is referred to.

References

1. J.M. Brick, and J. Burke, "Undercoverage Bias in the Field Test for the National Household Education Survey," report by Westat Inc. to the National Center for Education Statistics, 1990.

2. 0. T. Thornberry and J.M. Massey, "Trends in U.S. Telephone Coverage Across Time and Subgroups," Telephone Survey Methodology, ocrv, edited by R.M. Groves et al, John Wiley & Sons, 1989.

119

120

Session 4

TELEPHONE DATA COLLECTION

121

QUALITY IMPROVEMENT IN TELEPHONE SURVEYS

Leyla Mohadjer

David Morganstein

Westat, Inc.

1. Introduction

The use of telephone as an alternative mode of data collection in surveys has become very popular in recent years. Considerable research has been dedicated during the past decade to evaluate the quality of data collected in telephone surveys and to compare that with data collected by face-to-face interviewing. Simultaneous to the increased use of this methodology has been efforts at improving its efficiency and reducing the total error of telephone survey estimates. This paper provides a summary of recent methods for improving the quality of telephone surveys and reviews the recent literature on the results of these efforts.

Below we discuss several aspects of telephone surveys that fall into the category of "quality improvement." Most of these issues are design decisions that affect the expected total survey error. From its very beginning, the choice of telephone sampling over face-to-face sampling was one of improved efficiency. That is, the cost per complete in almost every case is significantly less than that of face-to-face sampling while the 'quality' of the results, as measured by total survey error, is little if any reduced. By way of comparison, mail-out surveys may have a very low cost per complete, but they suffer from large and generally unknown biases. Increasing efficiency is a traditional argument for system changes, such as the choice of telephone sampling over face-to-face interviewing, whose principal purpose is that of quality improvement.

In the following sections, we discuss several aspects of telephone survey operations in which the quality of a telephone sample design is established. We begin with decisions regarding the survey methodology. These decisions typically include the,

trade-off of greatly reduced survey cost for what might be, at most, a small increase in mean square error (MSE). Less quantifiable in cost terms is the reduced time to completion of survey operations afforded by a telephone methodology and improvements in the level of quality assurance.

Next, we discuss a number of sample design aspects which impact on the survey cost, schedule and error. We mention the much discussed issue of coverage and the general problem of frame construction as they relate to total error. A number of sample design improvements have been developed in the past few years which can decrease the expected number of wasted calls needed in the

process of identifying eligible respondents. These are described and compared.

As contrasted to other methodologies, telephone surveys contain a number of operational features which result in improved quality. Among these are aspects of management and supervision and

of direct data entry through Computer Assisted Telephone Interviewing (CATI). We discuss and quantify some of the benefits which accrue from these approaches. Lastly, we review the topic of estimation as it relates to minimum mean square estimates. Several estimation procedures are required by the sample designs which are worthy of note.

2. Overview of the Properties of Telephone Surveys

Telephone surveys have become an often selected alternative to face-to-face interviewing for several reasons. Telephone surveys can be conducted at a much lower cost when compared to face-to-face interviewing. They also allow for the sample results to be available more quickly than face-to-face surveys. There are greater opportunities for quality control through more rigorous supervision and through frequent monitoring of the interviewing staff. Also, telephone interviewing makes it possible to contact otherwise hard-to-reach respondents such as those living in difficult to visit or dangerous neighborhoods, in bad weather conditions, or late at night (Groves and Kahn, 1979). The sample

design effects for estimates derived from telephone surveys are smaller than those coming from more heavily clustered area probability designs. Finally, telephone surveys have smaller interviewer effects. Discussions on these issues are provided in different sections of this paper.

Considerable research has been dedicated to improving sampling techniques, to increasing response rates, and to reducing noncoverage bias. Research has also focused on the issue of data quality, a comparison of collection modes, and the influence of collection mode on the quality of the data. Several authors such as Groves (1979) and Jordan (1980) have stated that one of the causes of lower performance for telephone surveys when compared to face-to-face surveys is the lower degree of operational experience with telephone surveys. Leeuw and Zouwen (1988) have analyzed the results of a number of studies in this area. Their work confirmed that the difference between the face-to-face and telephone interviews is becoming smaller over time.

Leeuw and Zouwen (1988) integrated findings on interviewing

mode differences and have provided a review of this topic. The method of analysis they used made it possible to present an overview of mode differences found with respect to data quality and estimate the size of these differences. The main conclusions of their paper are the following:

- Response rates are generally higher, for face-to-face interviews than for telephone interviews;

- The majority of studies did not find statistically significant differences in modes. When differences were found, however, they were in favor of face-to-face interviews; and

- Only small differences were found between random digit dialing (RDD) and face-to-face, and the differences have become smaller over time.

Leeuw and Zouwen (1988) also point out that one major

difference between the two modes is the lack of visual support in telephone surveys. This makes the respondent's task of answering some questions difficult in telephone surveys. It also results in reduced control over the respondents behavior in telephone surveys. On the other hand, since the questions come through the phone, responses are meaningless for other persons in the same room with the respondent, especially for closed questions. This reduces the potential influence of "bystanders" on the respondents.

The fact that telephone interviewing can be contained in a small area offers many potential benefits. Interviews done by telephone are subject to more supervisory control than field surveys, resulting in a positive effect on the quality of data from telephone surveys. Unlike the face-to-face mode, supervisors can monitor telephone interviewing anonymously and frequently with little impact on survey costs. This allows for rapid modification of questionnaire wording found to be problematical. In addition, they can arrange for needed interviewer re-training or they can make appropriate re-assignments if an interviewer is observed to be unsuitable for their assignment. In addition, with CATI systems, it is much easier to put checks and probes in different parts of

the interview to insure that answers provided by respondents are consistent throughout the questionnaires. All of these features should result in reduced non-sampling error.

Two disadvantages of telephone surveys are the noncoverage of persons living in households without telephones, and lower response rates when compared to face-to-face surveys. Section 3 provides a discussion of undercoverage in telephone surveys and the methods available to compensate for the undercoverage. Section 4 discusses nonresponse issues in telephone surveys.

3. Undercoverage in Telephone Surveys

Households without telephones are not included in telephone surveys since the sampling frames do not include such households. A considerable amount of information has been published on the nature of possible biases resulting from the use of a telephone sampling frame. Thornberry and Massey (1988) have analyzed trends

in telephone coverage in the U.S. across time and subgroups of the population. They indicate that estimates for the entire U.S. population, may experience only minor biases because of the high rates of telephone usage, about 93 percent of the population can be reached by telephone.

Although overall telephone coverage has risen to a very high level, it is not uniformly distributed across the population. Thornberry and Massey (1988), Groves and Kahn (1979), and Banks (1983) have shown striking differences between telephone and non-telephone households with respect to demographics, economics, and health characteristics.

As might be expected, telephone coverage correlates highly with income. Massey (1988) points out that other variables such as employment status, education, marital status, and race are also correlated with income and thus affect telephone coverage. More lower-income persons tend to be missed in telephone screening. This, in effect, results in higher telephone penetration for whites than blacks. Telephone coverage is lower in the South than in the

rest of the U.S., and it is lower in rural than urban areas.

Massey (1988) points out that noncoverage bias is a function of the noncoverage of a telephone survey frame, and of the difference in characteristics between the covered and uncovered population. Even though the percentage of households with telephones may increase and the overall noncoverage rate becomes smaller, large differences between telephone and nontelephone households can result in significant noncoverage bias. Surveys which focus on income or variables related to income may experience high noncoverage bias. It is true that the estimates of characteristics for the total population may not be drastically affected by the omission of nontelephone households, however, for some subdomain estimates there could be large biases due to the exclusion of households without telephones.

3.1 Methods to Compensate for Undercoverage

Several methods are available in telephone surveys to address

the problem of noncoverage bias. One approach which may eliminate certain kinds of undercoverage bias is the use of dual frames. Dual frame, mixed mode surveys use a combination of RDD and face-to-face samples to overcome the noncoverage of households without telephones. Research in the area of such mixed mode surveys include Sirken and Casady (1988), Groves and Lepkowski (1985), Lepkowski and Groves (1984), Biemer (1983), and Casady et al. (1981).

Sample weighting adjustments in the form of post stratification factors can be used to decrease the effects of noncoverage. The post-stratified weights are frequently employed

in national surveys to compensate for noncoverage bias. The subgroups established for the purpose of post-stratification are specifically tailored to each study. Subgroups are defined on the basis of variables thought to be correlated with the major statistics to be obtained from the survey as well as variables correlated with telephone penetration and nonresponse distribution.

Massey and Botman (1988) have investigated the impact of post-stratification survey adjustments in national surveys. They discuss several post-stratified weighting adjustment methods for RDD surveys, and show the effect of these adjustments on the estimates. Other work done in this area includes Banks (1983), Banks and Undersign (1982), and Thornberry and Massey (1978). Their results show that, although these methods reduce the effects of undercoverage, they do not completely eliminate the bias.

3.2 Within Household Coverage

The main focus of research in the area of coverage in random digit dialing surveys has been on sampling frame inadequacies, i.e., the exclusion of nontelephone households from the frame, as discussed earlier. However, there is another cause of undercoverage that arises from failure to obtain complete listings of household members in responding households. This is usually referred to as within household coverage. Within household coverage also exists in face-to-face surveys. Maklan and Waksberg

(1988) used two surveys conducted by Westat and compared their within-household coverage rates with those obtained by the Current Population Survey (CPS). They concluded that the coverage of persons in households with telephones generally available in RDD surveys is at least as good, if not better, than that provided by CPS.

4. Nonresponse Issues in Telephone Surveys

Groves (1988) gives an overview of nonresponse issues in telephone surveys, and distinguishes between those factors common in both face-to-face and telephone surveys and those factors that are specifically related to the selection mode. Factors such as length of the questionnaire, subject matter (topic of the survey), sensitivity of the questions, refusal conversion and callback routines are common in both modalities. The differences in response rates that Groves (1988) cites between face-to-face and telephone surveys are that refusal rates are higher for telephone surveys, and relatively more of the refusals take place immediately

after interviewers have introduced themselves prior to describing the purpose of the survey. However, as pointed out earlier, Leeuw and Zouwen (1988) have shown that these differences have become smaller over time. Researchers have varied the introductory section in an effort to reduce early refusals. A number of researchers have reported some improvements by using advance

letters to alert sample persons about the survey and the upcoming telephone call.

5. Choice of a Frame, List vs. RDD

Essentially, there are three types of sampling frames available for telephone surveys. List frames use information available in telephone directories, or other frames based on telephone directories, to generate telephone number sample. This is the alternative with the greatest undercoverage problem.

Second, random digit dialing provides a frame of all possible telephone numbers, and thus covers both listed and unlisted numbers. Third is a multi-frame approach which uses both directories and RDD. Lepkowski (1988) provides a description of these frames and methods of sample selection used with them.

6. Two-Stage RDD

Random digit dialing was originally developed to overcome the coverage problems inherent in directory samples; however, surveys of residential respondents were burdened by the excessive effort required to filter many nonworking or business telephone numbers. The Mitofsky-Waksberg cluster sample technique eliminates much of this inefficiency by utilizing the manner in which the telephone industry initiates new phone exchanges which is to assign a prefix to either a business or a residential clientele. Accordingly, it is possible to select a probability sample that is significantly richer in residential numbers than would be obtained by conducting a simple random sample of telephone numbers.

6.1 Waksberg Method for Reducing Effort

The method frequently used for large scale residential telephone surveys is a two-stage cluster procedure. This method was originally developed by Mitofsky (1970) and Waksberg (1978), and is usually referred to as the Mitofsky-Waksberg method. In a 1978 article, Waksberg demonstrated mathematically that this procedure provides a probability sample of households with telephones, in which all telephone numbers have the same probability of selection. Further, the method was shown to require a smaller number of telephone calls than the sampling procedures previously used for RDD, and thus, as a quality improvement, significantly reduces the cost and time involved in such surveys in comparison with dialing numbers at random.

The majority of numbers dialed completely at random are nonworking, business and other nonresidential numbers. Current estimates are that about 75 percent of the potential numbers within

existing telephone prefixes are nonworking and another three percent are businesses or institutions of some type. Given that only about 20 percent are residential numbers, a typical RDD simple random sample requires that five calls be made to locate a single household. In some cases, the telephone companies do not provide a message that the number dialed is not a working number. Additional checking necessary to distinguish between not-at-homes and nonworking numbers adds further to the cost of achieving completed interviews.

The Mitofsky-Waksberg sampling method is designed to reduce the number of nonproductive calls. It takes advantage of the fact that a high proportion of nonworking and commercial numbers occur in consecutive sequences. Essentially, the procedure involves two steps: first, "household cluster identification" (identifying and, selecting a sample of blocks of 100 numbers called "telephone clusters," which contain working, residential telephone numbers); and second, dialing random numbers within the clusters. Users of this technique typically locate three residential numbers for every five attempted within each cluster, a significant improvement in

efficiency for minimal additional effort.

6.2 Modified Waksberg Method

The "standard" Mitofsky-Waksberg method, which produces a self-weighting sample, involves designating a desired number of household clusters, and sampling a constant number of households per cluster. There are, however, some awkward operational features arising from the requirement for a constant number of households per cluster. For example, before the need for more telephone numbers in specific clusters can be determined it is necessary to wait until the required number of households have been identified and interviewed. Since a large number of calls are required to determine whether a telephone number is residential and, if so, to obtain the cooperation of the household, the standard method is rather time-consuming.

To improve the data collection process and to reduce the data collection time, researchers have come up with different ways to

speed up the data collection (for example, refer to Alexander [1988], Potthoff, JASA [1987], and Potthoff [1987]). The modified Waksberg procedure that Westat sometimes applies is based upon a fixed number of telephone numbers (instead of households) per clutter. There is thus no necessity to wait until the original sample of clusters has been completed to determine whether the desired number of households within clusters has been achieved. With the modified method, sample size becomes a random variable and the tight control on sample size offered by the original procedures is loosened. What is more, the modified procedure results in a sample that requires sample weights to adjust for differential probabilities of selection. Accordingly, its reduced data

collection time is purchased at the price of increased sampling error.

7. Efficiency of Estimates Derived from RDD Studies

In designing a two-stage RDD sample, the number of sample clusters and the average number of sample households per cluster must be specified. The choice of the sample sizes is usually made on the basis of cost and variance considerations. The extent to which the variances are increased due to clustering depends on the intraclass correlation between households within cluster and the average number of eligible households per cluster.

Clustering generally reduces survey data collection costs. The magnitude of the cost, however, is very different for face-to-face than it is for telephone surveys. The cost savings brought about by reduced travel costs is a virtual necessity in face-to-face surveys wherein they could comprise a substantial portion of the total survey cost. In telephone surveys, clustering is used to reduce the cost of dialing and reaching telephone numbers that belong to households. Considering the minimal cost of dialing telephone numbers (especially when compared to travel cost in face-to-face surveys), cluster sizes in telephone surveys need not be as large as they are in face-to-face surveys. As a result, statistics derived from RDD surveys are generally more efficient (have smaller

variances) than those coming from face-to-face surveys.

8. Improvements in Locating Rare Populations in Telephone Surveys

Studies of specific subgroups of the population that comprise relatively small proportions of the total population have always been the focus of many research efforts. With any method of sample selection, surveys of rare populations almost always require a considerable amount of screening. The frame generally used for RDD, a computer file provided by AT&T;, comprises all telephone households. Subsets cannot be determined except as part of a screening procedure. Extensive screening is necessary to locate members of the rare population, and as a result, it is usually very costly to sample rare populations through telephone surveys,

One efficient option for sampling members of rare populations is to use a commercially available tape (the Donnelley tape) that contains census population characteristics for prefix areas.

Mohadjer (1988) provides an evaluation of the quality of the information on this tape. Furthermore, Mohadjer (1990) discusses the effectiveness of using the Donnelley tape to improve the sample efficiency in an education study. She shows that sampling efficiency is greatly improved by using the Donnelley tape to oversample blacks and Hispanics.

9. Interviewer Effects in Telephone Surveys

Many studies have compared interviewer effects in telephone and face-to-face surveys. They mainly speculate that the interviewer effect is smaller for telephone surveys than for face-to-face surveys. In this section we examine the potential causes for interviewer effects and the way these causes relate to the data collection mode.

Stokes and Yeh (1988) give the following as the potential causes for interviewer effect:

1. Not following directions exactly;

2. Variations in personalities, tone of voice;

3. Respondent's reaction to characteristics of the

   interviewers; and

4. Different response rates for different interviewers.

   The main belief is that the variability among interviewers is smaller in centralized telephone surveys than in face-to-face surveys. The reason of ten given is that these effects can be controlled better by monitoring and supervision in a centralized data collection facility. Telephone interviewers can be much more easily monitored and training can be more uniform as well as more frequent. Furthermore, interviewers have the opportunity to observe and learn more from each other in a centralized facility such as a telephone center. This makes interviewer behavior more uniform in telephone surveys than in face-to-face surveys. Differences between interviewers can be detected much easier,

especially in centralized facilities. When differences are observed between interviewers, steps can be taken readily to reduce, them. For example, changes in training or instructions to interviewers can be implemented more quickly.

The interviewers personality and the respondents reactions to the interviewer also have smaller effects in telephone surveys. The tone of voice is the only variable that is thought to have a higher effect in telephone surveys than in face-to-face surveys. This effect is suggested because of the lack of visual contact in telephone surveys (the lack of visual contact increases the effect of tone of voice on respondents).

A number of steps can be taken to limit these interviewer effects even further. There are several quality control measures which can provide a quick assessment of interviewer performance and which can identify the need for action. Strict supervision is especially important in the early stages of data collection to

insure that all interviewers are following directions and have a clear understanding of the survey purpose and instrument. Group meetings to emphasize important aspects of the procedures and individual conferences with weaker interviewers should be used to limit the effect of interviewer differences. All interviewers should be monitored when they first begin data collection. Staff who fail to meet or exceed standards should not be allowed to continue until they have undergone remedial training.

10. Computer-Assisted Telephone Interviewing (CATI)

The use of CATI was a quantum step in telephone survey quality improvement. Survey organizations have used CATI with increased frequency in recent years because of its many benefits. It is believed that CATI improves the quality of the data collected, it reduces the cost of data collection, and it increases the timeliness of telephone surveys.

A CATI system has the potential for providing clean data immediately after interview completion. Three CATI features contribute to this capability. First, most data edits can be done on-line as the responses are entered. A CATI program can prevent interviewers from entering out-of-range responses ("hard range check") and can be programmed to require verification of unlikely responses ("soft range check"), e.g., such as an age of 100 years. Second, consistency checks are possible in the CATI program appropriate to inconsistent responses verified during the interview. Third, CATI can be set up so as to prevent the interviewer from leaving a question incomplete. If the interviewer has difficulty recording an answer (e.g., difficulty categorizing the answer into the precoded choice on the CATI screen), they can be trained to enter a "comment" explaining the circumstances. A quality control monitor can be responsible for reviewing all interviewers comments on a daily basis to resolve difficulties and to update the data files as required.

It was previously observed that a telephone interview methodology helps to reduce between-interviewer variances because

of greater opportunity for monitoring and supervision. Since interviewers can be easily observed without disturbing the interview process, frequent monitoring can be used to uncover interpretation and presentation difficulties, all of which contributes to reduced interviewer variance. A CATI approach represents yet another step in this same direction. Between interviewer differences in understanding the flow of the instrument can be virtually eliminated.

Often sampling efficiency can be improved through the use of complex respondent selection procedures. Unfortunately, complexity can breed errors especially when interviewers are tired or are dealing with a difficult set of questions. Through the use of

CATI, very complicated sampling rules can be implemented, virtually without error. The interviewer enters the household composition and the software selects a random respondent using pre-specified sampling rates.

As pointed out by Nicholls (1988), additional advantages of the CATI systems can be summarized in the following way:

- Rather than being managed by the interviewers, the status of each sampled case is available in the computer, thereby improving sample management.

- The scheduling and assignment of cases are done by computer. The scheduler schedules the appropriate time to call respondents taking into account the time differences across the U.S.

- On-line interviewing makes it possible to display the instruction to the interviewers, display the survey questions, and response categories without any need to use paper and pencil.

- Answers to closed questions which are not in the permissible range can be determined at the instant the response is entered. The software can prompt the

interviewer that this answer contradicts another response given by the same respondent at an earlier point in the interview and a correction can be made immediately. This reduces the need for data retrieval.

- Branching or skipping to the next item is done by the computer. This improves the quality of data collected for more complex data collections that involve complicated skip patterns and subsampling at different stages of data collection.

- Interviewers may interrupt, resume or repeat some of the sections. Also they can go back and correct previous answers or write notes on the screen in appropriate places.

- The system improves supervision. The screen and the telephone conversation can be seen and heard with no disturbance to the interviewing process. The telephone conversation with the respondent can be monitored. All

of these advantages result in faster reaction to the needs for clarification, re-training or re-assignment.

- The survey results are virtually ready for weighting and tabulation upon completion of the data collection phase. This more timely data collection makes possible survey schedules that could not have been met in the past.

- The CATI system maintains records of on-line calls, outcomes of the calls, response rates, and the amount of time spent by interviewers. It can also be used to time different parts of the questionnaire if the survey length becomes a problem.

Since effective CATI interviewers must be able to perform a number of demanding tasks simultaneously, the task of training suitable staff is more challenging. Interviewers must establish

rapport with a respondent, accurately read the questions shown on the terminal screen, correctly code the response, and enter messages to the respondent's file indicating that a probe (e.g., reading a question, prescribed clarification of an item, etc.) was required. In addition, they must record verbatim a respondents comments oh a question, and keep the respondents interest long enough to complete the interview. This is a set of qualifications that require interpersonal, computer, and typing skills that surpass those of traditional telephone interviews. Fortunately, the improved ability to monitor telephone interviews conducted via CATI assist in assuring that suitable staff is adequately prepared for the survey.

11. Summary

Face-to-face interviewing has long been the standard data collection method selected when the highest quality survey results were required. The authors have reviewed those features of telephone surveys which can result in improved survey quality, that is, reduced total survey error for the same, or even for less, cost

as other modes such as face-to-face interviewing. After review of these features, survey designers are better able to choose between a telephone sampling approach and a face-to-face methodology.

References

Alexander, C.H., "Cut Off Rules for Secondary Calling in a Random Digit Dialing Survey," Telephone Survey Methodolocry, edited by Robert M. Groves et al, John Wiley & Sons, 1988.

Banks, M.J., "Comparing Health and Medical Care Estimates of the Phone and Nonphone Populations," Proceedings of the Section on Survey Research Methods, American Statistical Association, 1983 page 569-574.

Banks, M.J., and Anderson, R.M., "Estimating and Adjusting for Nonphone Coverage Bias Using Center for Health Administration Studies Data," in National Center for Health Services Research, Heath Survey Research Methods: Proceeding of the 4th Biannual

Conference, National Center for Health Services Research Proceeding

Series, Department of Health and Human services Publication No. PHS) 84-3346, 1982.

Groves, R.M., and Lyberg, L.E., Tdlephone Survey Methodology, edited by Robert M. I Groves et al, John Wiley & Sons, 1988, page 191.

Groves, R.M., and Kahn, R.L., Surveys by Telephone: A National Comparison With Personal Interviews, New York, Academic Press, 1979.

Jordan, L.A., Marcus, A.C., and Reeder, L.G., "Response Styles in Telephone and Household Interviewing: A Field Experiment," Public Opinion Quarterly, Vol. 44, No. 2, Slimmer 1980, page 210-222.

DeLeeuw, E.D. and Van der Zouwen, J., Telephone Survey Methodology,

edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 283.

Lepkowski, J.M., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley Sons, 1988, page 73.

Maklan, D., and Waksberg, J., Telephone Survey Methodology, edited by Robert M. Groves et al. John Wiley & Sons, 1988, page 51.

Massey, J. T., and Botman, S.L., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 143.

Massey, J. T., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 3.

Mitofsky, W., "Sampling of Telephone Households," unpublished CBS memorandum, 1970.

Mohadjer, L., "A Study of the Effectiveness of Oversampling

Telephone Clusters with High Concentrations of Blacks and Hispanics in the NHES Field Test," unpublished report to NCES, 1990.

Mahadjer, L., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 161.

Nicholls, W.L., II, Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 377.

Potthoff, R.F., "Some Generalizations of the Mitofsky-Waksberg Technique for Random Digit Dialing." Journal of the American Statistical Association, Vol. 82, No. 398, June 1987, pp. 409-418.

Potthoff , R. F. , "Generalizations of the Mitofsky-Waksberg Technique for Random Digit Dialing: Some Added Topics," Proceedings of the

Section on Survey Research Methods, American Statistical Association, 1987, pp. 615-620.

Sirken, M.G., and Casady, R.J., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 175.

Stokes, L., and Yeh, M., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 357.

Thornberry, O.T., Jr., and Massey J.T., Telephone Survey Methodology, edited by Robert M. Groves et al, John Wiley & Sons, 1988, page 25.

Thornberry, O.T., Jr., and Massey J.T., "Correcting for Undercoverage Bias in Random Digit Dialed National Health Surveys," Proceeding of the Section on Survey Research Methods, American Statistical Association, 1978, page 224-229.

Waksberg, J., "Sampling Methods for Random Digit Dialing," Journal of the American Statistical Association, Vol. 73, No. 361, March 1978, page 40-46.

COMPUTER ASSISTED SURVEY TECHNOLOGIES IN GOVERNMENT:

AN OVERVIEW

Marc Tosiano

National Agricultural Statistics Service

Introduction

CATI is an acronym for computer assisted telephone interviewing. It it the interactive use of computers to assist in

data collection activities typically performed in a centralized telephone facility of a survey organization.(27)(31) CATI is only one use of the computer in the growing realm of computer assisted survey work. Other uses of computer assisted surveys include: 1) computer assisted personal interviewing (CAPI), 2) computerized self administered questionnaires (CSAQ,), 3) computer assisted data entry (CADE) of information on paper questionnaires into a electronic format.(31)(26) Each of these computer assisted survey techniques may be used alone for a survey or in combination depending on survey management requirements and the various modes used to collect data for a given survey.

Features of computer assisted surveys

During an interview, the minimum use of computer assisted survey technology is the presentation of survey questions and their response categories on the computer screen. Interviewers read the question to the respondent and key the answer on the screen by using the computer keyboard. However, computer assisted survey

techniques offer many capabilities above and beyond the traditional paper questionnaire. These features include enhancements to the interview proper as well as the automation of survey management activities. Obviously, the features available depend on the software chosen for computer assisted interviewing.(10) Common features offered by various computer assisted survey software include: (26)(10)(24)

On-line interviewing:

- Instructional or reference information appears on the screen or is available via help screens to assist the interviewer.

- Fills are used to customize question wording by inserting input from records prior to the survey or from answers to previous questions.

- Answers to closed questions are checked against permissible entries. Some software offers multiple responses as well.

- Numeric answers are checked against a pre-defined range.

- Consistency checks are made against data collected earlier in the interview.

- Answers detected as invalid can invoke an error correction routine or additional probing questions.

- Formats are available for special answers, e.g., date, time, money, zip code, etc.

- Open-ended questions or interviewer notes are answered by typing text.

- Question order as well as response categories may be randomized to reduce order effects.

- Item-based design offers one question per screen or multiple related questions per screen; the interviewer is forced to answer the questions in a pre-determined sequence.

- Form-based design presents a screen that simulates a paper form. The interviewer is free to move the cursor around the form and fill in-the form in any order.

- Automatic branching is done based on input from records prior to the survey, previous answers in the interview, logical conditions, or arithmetic checks.

Creating the computer assisted questionnaire.

- Some packages offer a menu driven approach to building the questionnaire while others require the use of a special programming language.

- Some packages come with their own editor to write or change the questionnaire, but other editors or word processors may be used as well.

- Questionnaire debugging tools of various strengths may be available.

- A paper copy of the questionnaire including screen prints and a flow chart of the questionnaire may be available.

Survey management:

- The sample is stored in computer media and the software maintains the status of each questionnaire.

- Sampling procedures may be available including random digit dialing facilities.

- Call scheduling delivers the next case to be called by the CATI interviewer. The call scheduler prioritizes and sequences the calls made in the CATI environment. This includes the retrieving of cases at the appointed time for a call-back, establishing follow-up calls for busy signals or no answers, and targeting groups of cases such as strata or replicates.

- Survey managers may generate reports including such things as: completions, response rates, refusal rates, time per interview or question, call-back appointments, etc. These reports may be by interviewer, by day, by shift, cumulative, etc.

- Monitoring individual CATI interviews may be done by

viewing the interviewer's screen at a supervisor's workstation where audio monitoring may be available as well.

Data handling and analysis:

- Post-survey processing may be done to review, edit, clean, or code each interview.

- A codebook may be created containing questions, the variable names and location in the dataset, etc.

- An audit trail may be maintained with all previous answers if an answer is changed.

- Output files are created in a form ready for the next processing stage, these could include SPSS and SAS datasets.

- Some packages offer their own statistical analysis

packages, including histograms, distributions, regression, Analysis of Variance (ANOVA), etc.

The features listed above are not available in all computer assisted survey software. A survey organization procuring software for a computer assisted application would have to decide which features are important and select software accordingly. In

addition, all software packages will not operate on all computer hardware; a problem for all computer systems which must be resolved is the matching of software to hardware.

Computer assisted survey software is relatively new and constantly evolving; enhancements are usually inspired by the needs and requirements of end users.(10) Therefore, another consideration in choosing software might be the existence of a user support group and the willingness of the software company to enhance the system as new features are requested by users and the cost of these

modifications.

These different features of computer assisted survey software have various effects on costs and quality of data. For example, the use of interactive interviewing may improve the quality of data, but without call scheduling, the productivity of interviewers may be unaffected.(26) If improved interviewer efficiency and the elimination of paper callback records is important, software with call scheduling would be more attractive. However, systems with call scheduling may not be strong in other areas such as form-based design or software portability across various hardware. Evaluating these trade-offs is a difficult but critical task in choosing (or developing) this type of software.

Costs and Data Quality

The CATI concept was originally proposed by the American Telephone & Telegraph Company; in 1971 they sponsored the first CATI survey to measure customer satisfaction.(26) After this experience, CATI was believed to have three advantages over

conventional data collection methods: "accuracy, speed and reduced costs".(22) Since then there have been many studies and papers evaluating the accuracy or extent of the validity of these original beliefs.(26,22,9,19,35,41,37,16,17,18,33,36,38) Some authors have also reviewed the impact of CATI on survey administration and the internal structure of survey organizations.(5,6,13,21,50,320 This section of the paper does not intend to review all of these sources but to briefly review some of the implications and consequences that arise by using this new computer assisted survey technology. Some of the topics discussed here are not easily definable as advantages or disadvantages; it often depends on the methods used to implement this new technology.

The first set of reasons for implementing computer assisted surveys is to expedite surveys and thereby reduce costs.(24) There is always the initial cost of procuring and maintaining hardware and software. This overhead cost could be alleviated by utilizing the hardware and/or software for projects other than computer assisted surveys.(35) Some of the hardware configurations used in the past have been 'dumb' terminals attached to a centralized mainframe

computer.(14) Later, terminals or 'intelligent' microcomputers were attached to a minicomputer.(27,35) The latest hardware innovation used is microcomputers used in a stand-alone or in a Local Area Network (LAN) environment.(8) After these items are procured, there are the costs for training the staff to implement the new technology, and training the interviewers on use of the system. Interviewer training costs also depend on the turnover rate of interviewers. CATI questionnaire design will take longer than paper design because it employs many of the features listed previously such as automatic branching, use of fills, interactive editing and consistency checks, interviewer 'helps' and special processing needed for other activities previously done on paper.(20) This special processing includes resolution of busys, no answers, refusals, arranging callbacks and other administrative activities. As with other programming, CATI-questionnaire designers typically 'steal' code from previous studies whenever possible. This efficient use of previous code is enhanced by the use, of modular or structured programming. The CATI questionnaire setup for some

surveys could be faster and simpler than creating a paper questionnaire, but only if the CATI instrument emulates the paper which is seldom the case.

Once past these overhead costs, there are other cost considerations. Interviews typically take longer with computer, assisted surveys because of the edit checks and additional questions generated to probe for corrections or clarifications; another reason could be the interviewer's lack of familiarity with the keyboard, especially if there is a lot of text to be entered. These higher costs are somewhat offset by other features of computer assisted surveys. The use of an automatic scheduler can improve interviewer efficiency and reduce the cost of supervision by eliminating voluminous and tedious paper shuffling; supervisors are freed to do more real supervising rather than managing callbacks.(50) Status systems automatically keep track of each case in the sample including its current disposition and any actions taken on the case. Immediately after each interview, the data is already in electronic medium; this eliminates the data entry stage necessary in conventional data collection. At any time during the

survey, output files are available for preliminary analysis and/or administrative reports needed to allocate resources during the remainder of the survey period.

The second and probably the more important set of reasons to implement computer assisted surveys is to improve survey data quality and enhance the ability to implement complex surveys.(24,9) One major source of improved data quality is the ability to perform on-line edit and consistency checks which means corrections can be made during the interview with the help of the respondent. Post-survey edit checks can be eliminated or greatly reduced.(35) Many times, post-survey corrections to the interview are done without re-contacting the respondent; this results in more unknown or

imputed data. Computer assisted surveys also result in increased standardization among interviewers, especially in a central telephone facility.(35,15) This standardization may help reduce some

interviewer effects typically seen in paper questionnaires such as following proper question sequence.(16) However, there are sources of error possible which did not exist in the paper environment such as simply keying the an incorrect number for an answer while using touch typing. Some of the benefits of complex instruments include: creation of multiple versions of a questionnaire within the same instrument, inclusion of pre-programmed probes, use of historic data from previous surveys, table look-up routines, and other techniques difficult to employ in a paper questionnaire. In addition, computer assisted technology permits easier implementation of research than does its paper counterpart. Some examples are: randomizing questions and answer categories, use of historic data, use of randomized probes to check respondents understanding of questions, re-interview and reconciliation studies, and item-based versus form-based questionnaire design.

Government CATI Implementations(39)

Early CATI systems were developed by United States market

research organizations in the late 1960's and early 1970'S.(14) University survey research centers became involved in this technology in the middle 1970's.(27) U.S. government agencies did not begin work with CATI until 1980 when both the Census Bureau and the National Agricultural Statistics Service (NASS) each established working groups to investigate this technology.(2,25,35)

The largest installations of CATI in the federal government are in operation in four agencies: Bureau of Labor Statistics (BLS), Census Bureau, National Agricultural Statistics Service (NASS), National Centers for Disease Control (CDC). BLS has about 70 workstations in 14 sites. This includes a 10 workstation test site for developing CATI methods for the Consumer Price Index (CPI) Surveys which is planned for expansion to a 50 workstation production facility by 1994. Another 20 workstation site is in BLS headquarters for special surveys of the BLS Office of Employment and Unemployment Statistics. Their largest use of CATI is 40 workstations in 12 sites for the monthly establishment survey supporting the Current Employment Statistics Program. CATI is used for interviewing, non-response follow-up, and failed edit reconciliation. If successful, BLS plans expansion of these 12

sites to all 51 State offices with about 200 workstations in 1994.

The Census Bureau has two CATI sites with about l00 workstations. one site of 30 workstations is the Field Division's Hagerstown Telephone Center which collects data for surveys of household residents and small surveys of industry. This site is expected to expand from 30 to between 250 and 300 workstations by

1994. The second Census CATI site is in Jeffersonville, Indiana where 70 workstations ate used to collect data from establishments for the Retail and Wholesale Trade Industries. Here, CATI is used for telephone interviews, data capture from paper questionnaires and failed edit follow-up.

The National Agricultural Statistics Service (NASS) surveys farm operators and agricultural businesses with the largest CATI network in the Federal government. NASS has about 200 workstations

operational in 14 State offices. Four additional State offices have recently installed the hardware and software for CATI and will soon become operational. This brings the NASS CATI capabilities to about 260 workstations in 18 State offices. Current plans are to install Local Area Networks in 42 State offices by 1992; this will increase the CATI workstation count to about 750 nationwide. While mostly used by CATI interviewers after business hours, these same workstations will also be used by the office staff during the day for normal office operations. These daytime operations include survey activities (e.g., data capture of paper questionnaires, interactive error detection and correction of data collected, survey management) and all other office work (e.g., word processing, spreadsheet operations, graphics).

The National Centers for Disease Control (CDC) operates about 150 workstations in 21 State offices to collect data for the Behavioral Risk Factors Surveillance Survey and other random digit dialed household surveys. Little expansion of the CDC CATI network is expected over the next few years because data collection is commonly contracted out to other survey organizations.

These Federal agencies are expanding their CATI capabilities and Plan to complete their initial CATI implementation by 1994. Unlike many private and university survey organizations, government CATI installations are Not generally implemented in a national or regional centralized telephone facility. Most of the federal resources are directed toward smaller State offices where the same equipment is used for other survey related activities and office automation (BLSL CDC, NASS). Even with this increase in CATI activities, CATI will not become the only mode of data collection. Mailed questionnaires are still important in the mixed mode method of data collection in NASS and BLS. Personal interviewing is still important to all agencies as well, often as part of mixed mode data collection; the field interviewing staff numbers about 3,000 in the Census Bureau and about 2,800 in the National Agricultural Statistics Service. With these large field staffs, implementing CAPI may be the next large task facing computer assisted survey work in these agencies.

In private and university survey organizations the use of

CATI is generally associated with a single centralized telephone facility. CATI encourages a centralized facility to benefit from some of the features listed earlier such as automatic call

scheduling, monitoring, and administrative reports.(13,40,1) A central facility is better suited to computer assisted operations because of the shared hardware, software, sample, and technical support. While CATI improves standardized interviewing and quality control (by automatic branching, tailored question wording, and probes for on-line edits) , centralization contributes to survey management with consolidated and more standardized training and supervision of interviewers. One disadvantages of centralization may be that the interviewers do not have the local knowledge, and cultural understandinG which local interviewers may share with the respondents.(21)

A major challenge to federal agencies implementing CATI involves the resolution of the associated issue of centralized or decentralized interviewing. Many agencies already have national,

regional, and/or State offices with commitments to Federal State agreements, office staff, and an interviewer staff including office and field interviewers. These commitments may have as much impact on implementation decisions as the goals of operational efficiency and maximizing data quality. The Census Bureau has transferred the Retail and Wholesale Trade survey from, the traditional regional telephone calling to one centralizeD CATI facility in Indiana. the other previously mentioned three agencies have maintained their dispersed data collection techniques by implementing CATI in the existing regional or State offices. However, these dispersed CATI facilities can be used as central sites as well. For example, if a given sample is so widespread across the country, one or more State offices can be designated as regional CATI centers for that survey.(5) NASS has successfully tested the centralization of CATI interviewing in regional centers while personal interviews were still administered from the State offices. However, this mixed mode with centralization for only part of the data collection requires strong communication, coordination, and overall survey management.(5)

Other organizational considerations revolve around the question, "How do computer assisted techniques fit in with the current mode of operations?" Some of these considerations may be specific to a survey or addressed for overall computer assisted operations. A few examples follow: What is the role of the supervisory interviewer? Should CATI edit checks during the interview or during post survey processing totally replace existing batch edits? How should technological advances in software and hardware be incorporated into an existing CATI operation? When a mailed questionnaire is followed up with CATI or CAPI, how closely should the interview instrument follow the paper questionnaire?"

In many cases, the difficulties of implementing computer assisted techniques in government agencies arise from organizational requirements, not the technology itself. Some of the problems encountered with CATI are due to use of a central

facility; these problems would be the same if paper questionnaires

were used in the same central environment. Therefore, it is important to understand the source of potential problems when advocating or implementing a computer assisted system technological and organizational.

The Future of Computer Assisted Technology(39)

As these four government agencies are approaching full CATI implementation, newer technologies are developing which go beyond telephone interviews and some re-evaluation is necessary. Very little research has been done to measure the cost, timeliness, and data quality of surveys done with these new approaches. This paper reviews some of the major new technologies, and their possible use by survey organizations. These technologies can be divided into five groups: computer assisted personal interviews, computer assisted self administered questionnaires, geographic and communication technologies, voice technology, and artificial intelligence.

Computer Assisted Personal Interviews

Now that computers are getting smaller and smaller, computer assisted personal interviewing (CAPI) is the next natural extension of computer assisted interviewing beyond CATI. As mentioned before, personal interviewing is still important in federal survey agencies; CAPI can be used to benefit from the advantages listed earlier and also improve the data transfer between personal and telephone interviewing for mixed mode surveys. Unlike the course of CATI development, government agencies are in the forefront of CAPI development both for their own use and in sponsoring CAPI investigations by universities and the private sector. Also, the government's implementation of CAPI is proceeding rapidly compared to CATI. CAPI investigations have found that CAPI data collection is acceptable to most respondents and that most experienced field interviewers can be trained in its use.(7,23,42,44,3,43,49)

In addition to the organizational considerations of implementation of CAPI there are some technological problems which need to be addressed. Assignments and questionnaires must be given to CAPI interviewers and completed interview data must be sent back to the office. National Analysts have used the mail, UPS, And

courier services for this transmittal during the Nationwide Food Consumption Survey.(44) Another method is to use automated telecommunications with modems attached to computers. Research Triangle Institute (contracted by the Envirormental Protection Agency) and the Netherlands Central Bureau of Statistics have used telecommunications with some success.(43,49) However, the Netherlands is returning to the use of mail for data transmission as a simpler

and less costly approach.(46) If a workable solution is found, rapid telecommunications between the office and interviewers may be especially advantageous when operating on tight deadlines and using mixed mode methods. The Census Bureau and NASS plan to investigate an integrated CATI-CAPI system where cases can be transmitted between CATI interviewers in central or state offices and CAPI interviewers dispersed throughout the field.

The software used in CAPI is typically the same as used for CATI or personal interviewing in an office environment. However,

personal interviews in the respondents home or at the doorstep can be more demanding and distracting. This may call for special software features for question formats, entry modes and questionnaire movement commands which are easier to use. These features specific to CAPI interviewers have not yet been determined or shared.

The hardware used for CAPI applications is still evolving and being investigated. Machines must also be evaluated based on the environment expected for conducting interviews: on a table top, standing and holding the machine, or both. The machines generally available for CAPI include laptop computers, hand held computers, and slate computers (handwritten character recognition devices). The laptops are generally 4 to 15 pounds and have various sized and types of screens and keyboards. Hand held computers are much smaller but offer very small screens, keyboards, and limited computing power which eliminates some software packages. Slate computers range from 3 to 4 pounds and are held like a clipboard while the interviewer reads questions and writes the answers on the screen with a stylus. This device emulates paper questionnaire

data entry and some machines are able to recognize special functions such as tallies, diagrams, maps, and signatures. Unlike a year ago, these devices now run DOS based systems and NASS can run both BLAISE and CASES computer assisted software for CAPI applications on the Gridpad machine.

The weight of these machines is an important factor in an interviewer's acceptance of using a machine as a data collection tool. Most recent tests of CAPI have been qualitative reports with inconsistent findings.(42,3,43) However, recent laboratory research has studied ergonomic properties of CAPI, interviewer attitudes, and logistical features of the technology. This work investigated the maximum weight of laptop computers which would lead to the acceptance of CAPI by interviewers for doorstep interviewing; further research is being done to include newer lighter laptops and slate computers.(45)

Once the technological problems are resolved, survey organization and management will require review and modification to meet the needs of a computer assisted survey environment. CAPI may

change the methods of assigning, conducting, supervising, checking-

in, and reviewing interviews. These changes will affect staffing requirements and how to most effectively organize and manage survey personnel. For example: 1) CAPI field supervisors must cope with hardware, software, and telecommunications problems in addition to interpersonal skills. 2) Interviewer training must include machine maintenance, CAPI interviewing, and transmission of assignments and data. To reduce costs, some of this training could be done as home study with on-line tutorials. 3) The software will eliminate survey specific errors such as inappropriate skips or data inconsistencies; however, supervisors will need to identify interviewers needing further training in CAPI operations. 4) Field supervisors and office staff must use new techniques to check-in, review, and edit CAPI interviews. Due to computerization, some of these-functions may also disappear requiring clerical staff to be

replaced with technical staff. 5) With better communications and data transmission, the relationship of State, regional, and headquarters staff may change as well. Data and messages could travel directly between field interviewers and headquarters. All these possibilities and more will affect how CAPI is implemented in the various survey agencies.

Computerized Self-Administered Questionnaires

Establishment surveys usually collect brief numeric responses from the same respondents time after time. New technologies may be welcomed by these respondents if it results in reduced respondent burden or is perceived as such. This area is ripe for the investigation of computerized self-administered questionnaires (CSAQ).

BLS is experimenting with voice simulation of the questions and touchtone data entry of the answers by the respondent.(51,29,4) When respondents have prepared their reports, they dial a local telephone number at a nearby BLS office; a voice simulation module

requests the entry of their identification number on the telephone's touchtone pad. The voice module then asks survey questions that the respondent answers by keying the numeric response on the touchtone pad. Since this procedure operates 24 hours, this interaction can be done at the respondent's convenience; without a telephone interviewer and data entry staff, costs are minimal. Of course, a BLS interviewer is still needed to call non-respondents after a cutoff date or to resolve data inconsistencies. A further extension of this project is voice recognition of the respondent which would eliminate the need to key answers on the touchtone pad.(48)

The Energy Information Agency (EIA) is investigating CSAQ by using respondents' personal computers.(34) Respondents who have access to personal computers are given diskettes containing the monthly CSAQ, menu-driven procedures to obtain the necessary

information from other files, and programmed procedures to

electronically transmit the completed questionnaire to the EIA ccomputer.

Geographic and Communication Technologies

Other technological developments may assist field interviewers in some of the administrative work accompanying personal interviews. These include automobile telephones, beepers, and navigational and position-recognizing systems to provide reliable geographic coordinates. This technology could be used to: 1) assist rural field interviewers in locating sampling units by using coordinate position of landmarks and buildings; 2) update maps by driving through new streets not on current maps; 3) define, coordinates of area frame boundaries for sampling because these coordinates are not affected by changes in physical boundaries or political borders; 4) recording precise locations of dwellings and establishments to allow summation of data to any area definable by geographic coordinates. On recent examination, the Census Bureau found that current systems are not sufficiently accurate, reliable

and cost-effective for typical survey applications.(39)

Voice Technology

The National Bureau of Standards has recommended that this technology be investigated by the Census Bureau as the next step in computer assisted methods.28 This technology includes both voice simulation and speech recognition. It could be used to conduct telephone interviews without human interviewers or as an auxiliary computer tool to reduce the keyboard skill necessary for interviewers using computer assisted methods. As mentioned earlier, some voice technology is being investigated for gathering data from establishments at their convenience. For household or other personal surveys, acceptance of a fully automated computer interview seems to depend upon respondent acceptance. However, the potential cost savings possible from voice technology will probably stimulate further research in this area; survey agencies will need to evaluate these new systems as they become available to judge their applicability to surveys.

Artificial Intelligence

Artificial Intelligence is a computer discipline which builds computer programs that perform tasks requiring intelligence when done by humans. This discipline is used to develop expert systems for problem solving which involve the use of appropriate information acquired previously from human experts.(11) This technology has been used by Westat for computer assisted coding of

open-ended questions on a paper questionnaire. Initially, humans do all the coding which is recorded by the computer and from this human input, the computer "learns" how to do this coding as well. As the coding process continues, the computer program can code increasingly more open-ended responses while the human operator can verify these codes and handles the responses not yet "learned" by the program.(47) Although this technology may have limited use during data collection, this may be a computer assisted technique which could benefit other survey management tasks like case assignments, questionnaire coding, and automatic call scheduling.

Conclusions

Government survey agencies have taken about 20 years to implement one new technology, CATI. Meanwhile, technology has advanced into many other areas such as computer assisted personal interviews, computerized self-administered questionnaires, geographic and communication technologies, voice technology, and artificial intelligence. This technology explosion means that survey agencies need to evaluate an ever increasing number of methods which may improve data collection and survey management. In addition to investigating new technologies, the associated organizational and methodological factors must be addressed so that all implications are considered before implementing advanced computer assisted survey methods. All the while, studies must continue to evaluate the effects of these factors on survey costs, timeliness, and data quality.

Acknowledgement

A note of gratitude goes to Bill Nicholls for material

presented in his report as referenced in [39].

References

1 American statistical Association, Proceedings of the Section on Survey Research Methods, Washington, D. C.: American Statistical Association, 1978. Experiences with CATI in a Large-Scale-Survey, by William L. Nicholls II, pp. 9-17.

2 American Statistical Association, Proceedinas of the Section on Survey Research Methods, Washington, D.C.: American Statistical Association, 1983. Measuring CATI Effects on Numerical Data, by Carol C. House and Betsy Morton, pp. 135-138.

3 American Statistical Association, Proceedings of the Section on Survey Research Methods, Washington, D.C.: American Statistical

Association, 1988. Development of a Computer Assisted Personal Interview For the National Health Interview Survey, by Stewart C. Rice Jr., Robert A. Wright and Ben Rowe.

4 American Statistical Association, Proceedings of the Section on Survey Research Methods, Washington, D.C.: American Statistical Association, 1989. Developing a Cost Model for Alternative Data Collection Methods; Mail, CATI, and TDE, by Richard Clayton and Louis Harrell.

5 Bass, Robert T., and Robert D. Tortora, "A Comparison of Centralized CATI Facilities for An Agricultural Labor Survey," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.

6 Berry, Sandra H. and Diane O'Rourke, "Administrative Designs for Centralized Telephone Survey Centers: Implications of the Transfer to CATI," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.

7 Birkett, N. J., "Computer-Aided Personal Interviewing: A New Technique for Data Collection in Epidemiologic Surveys," American Journal of Epidemiology, Vol. 127, No. 3, 1988t pp. 684-690.

8 Carpenter, Edwin H., "Software Tools for Data Collection: Micro-Assisted Interviewing," Social Science Computer Review, Fall 1988.

9 Catlin, Gary and Susan Ingram, "The Effects of CATI on Costs and Data Quality: A Comparison of CATI and Pater Methods in Centralized Interviewing," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.

10 deBie, Steven E., Inede A. L. Stoop and Katrinus L. M. deVries, CAI Software -- An Evaluation of Software for Computer Assisted Interviewing, Amsterdam, Association of Social Research Institutes, 1989.

11 Dictionary of Computing, New York, Oxford University Press, 1986.

12 Dillman, Don A. Mail and Telephone Surveys -- The Total Design Method. New York: John Wiley & Sons, 1978.

13 Dillman, Don A. and John Tarnai, "Administrative Issues in Mixed Mode Surveys," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.

14 Fink, James C., "CATI's First Decade: The Chilton Experience," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 153-168.

15 Freeman, Howard E., "Research Opportunities Related to CATI," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 143-152.

16 Groves, Robert M. and Nancy A. Mathiowetz, "Computer Assisted Telephone Interviewing: Effects on Interviewers and Respondents,

Public Opinion Quarterly, Vol. 48, 1984, pp. 356-369.

17 Groves, Robert M. and Lou J. Magilavy, "Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys," Public Opinion Quarterly, Vol. 50, 1984, pp. 251-266.

Groves, Robert M., "Implications of CATI," Sociologica1 Methods Research, Vol. 12, No. 2, 1983, pp. 199-215.

19 Groves, Robert M and William L. Nicholls II, "The Status of Computer-Assisted Telephone Interviewing: Part II -- Data Quality Issues," Journal of Official Statistics, Vol. 2, No. 2, 1986, pp. 117-134.

20 House, Carol C., "Questionnaire Design with Computer-Assisted Telephone Interviewing," Journal of 0fficial Statistics, Vol. 1, No. 2, 1985, pp. 209-219.

21 Lyberg, Lars, "Administration of Telephone Surveys," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York,

Wiley Press, 1988.

22 Nelson, R. O., B. L. Peyton and B. Z. Bortner, "Use of an On-Line Interactive System: Its Effects on the Speed, Accuracy, and Costs of Survey Results." Presented at the 18th Advertising Research Foundation Conference, New York City, 1972.

23 Netherlands Central Bureau of Statistics, Automation in Survey Processing, Select Report 4, Voorburg, Netherlands, Central Bureau of Statistics, 1987.

24 Nicholls II, W. L., "Computer-Assisted Telephone Interviewing: A General Introduction," in Robert M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.

Nicholls II, William L., "CATI Research and Development at the Census Bureau," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 191-197.

26 Nicholls II, William L. and R. M. Groves, "The Status of

Computer-Assisted Telephone Interviewing: Part I -- Introduction and Impact on Cost and Timeliness of Survey Data," Journal of Official Statistics, Vol. 2, No. 2, 1986, pp. 93-115.

27 Palit, Charles and Harry Sharp, "Microcomputer-Assisted Telephone Interviewing," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 169-189.

28 Pallett, D.S. (editor), Automation of Data Capture for the Census in the Year 2000. Final report to the U.S. Bureau of the Census from the National Bureau of Standards under the interagency agreement for fiscal year 1987.

29 Ponikowski, Chester, and Sue Meily, "Use of Touchtone Recognition Technology in Establishment Survey Data Collection." Preesented at the First Annual Field Technologies Conference, St. Petersburg,

Florida, 1988.

30 Schuman, Howard and Stanley Presser. Questions and Answers in Attitude Surveys -- Experiments on Ouestion Form, Wording, and Context. Orlando, FL: Academic Press, Inc., 1981.

31 Shanks, J. Merrill, "The Current Status of Computer-Assisted Telephone Interviewing," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp.1119-142.

32 Sharp, Harry, and Charles Palit, "Sample Administration with CATI: The Wisconsin Survey Research," Journal of Official Statistics, Vol. 4, No. 4, 1988, pp. 401-413.

33 Sudman, Seymour, "Survey Research and Technological Change," Sociological Methods & Research, Vol. 12, No. 2, 1983, pp. 217-230.

34 Swann, T. C., "Electronic Data Collection in the Petroleum Supply Reporting System." Presented at the American Statistical Association Committee on Energy Statistics, April 28-29, 1988.

35 Tortora, Robert D., "CATI in an Agricultural Statistical Agency,"

Journal of Official Statistics, Vol. 1, No. 2, 1986, pp. 301-314.

36 Tucker, Clyde, "Interviewer Effects in Telephone Surveys," Public

Opinion Quarterly, Vol. 47, 1984, pp. 84-95.

37 U.S. Department of Agriculture, A Comparison of CATI and NonCATI on a Nebraska Hoct Survey. ' Statistical Reporting Service Staff

on a Nebraska Hog Survey. Statistical Reporting Service Staff

Report No. 89, by Richard Coulter. Washington, D.C.: Statistical

Reporting Service, May, 1985.

39 U.S. Department of Agriculture, Computer Assisted Telephone

Interviewing on the Cattle Multiple Frame Survey. Statistical

Reporting Service Staff Report No. 82, by Carol C. House.

Washington, D.C.:" Statistical Reporting Service, October, 1984.

39 U.S. Department of Commerce, The Impact of High Technoloav on

Data Collection. CATI Research Report N. GEN-1, by William L.

Nicholls II. Washington, D.C.: U.S. Bureau of the Census,

February, 1989.

40 U. S. Department of Commerce, Proceedings of the Bureau of the Census First Annual Research Conference, Washington, D.C.: Bureau of the Census, 1985. Cost and Error Modeling for Large-Scale Surveys, by Robert M. Groves and James M. Lepkowski, pp. 330-357.

41 U.S. Department of Commerce, Proceedings of the Bureau of the Census Third Annual Research Conference, Washington, D.C.: Bureau of the Census, 1987. Use of Historical Data in a Current Interview Situation, by Bradley.V. Pafford and Dick Coulter, pp. 281-298.

42 U. S. Department of Commerce, Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, D.C.: Bureau of the Census, 1988. "Discussion" of papers by Rothschild and

Wilson and by Sebestik et al., by William L. Nicholls II, pp. 340-342.

43 U.S. Department of Commerce, Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, D.C.: Bureau of the Census, 1988. Initial Experiences with CAPI, by Jutta Sebestik, Harvey Zelon, Dale DeWitt, James M. O'Reilly and Kevin McGowan, pp. 357-365.

44 U.S. Department of Commerce, Proceedings of the Bureau of the Census Fourth Annual Research Conference, Washington, D.C.: Bureau of the Census, 1988. Nationwide Food Consumiption Survey Using Laptop Computers, by Beth B. Rothschild and Lucy B. Wilson, pp. 347-356.

45 U.S. Department of Commerce, Proceedings of the Bureau of the Census 1990 Annual Research Conference, Washington, D.C.: Bureau of the Census, 1990. Building Predictive Models of CAPI Acceptance in a Field Interviewing Staff, by Mick Couper, Robert M. Groves and Curtis A. Jacobs, forthcoming.

46 U.S. Department of Commerce, Proceedings of the Bureau of the Census 1990 Annual Research Conference, Washington, D.C.: Bureau of the Census, 1990. The Impact Of Microcomputers on Survey Processing at the Netherlands Central Bureau of Statistics, by Wouter J. Keller and Jelke G. Bethlehem, forthcoming.

47 U.S. Department of Commerce, Proceedings of the Bureau of the Census 1990 Annual Research Conference, Washington, D.C.: Bureau of the Census, 1990. Improvling Data Quality in National Surveys: Experience with Computer-assisted Methods in the National Post-secondary Student Aid Studieg, by James E. Smith and Carmen Vincent, forthcoming.

48 U.S. Department of Labor, Voice Recognition and Voice Response Applications for Data Collection in a Federal/State Establishment Survey. By Richard Clayton and Debbie Winter. Washington, D.C.: U.S. Bureau of Labor Statistics, 1989.

49 vanBastalaer, Alois, Frans Kessamakers and Dirk Sikkel, "Data Collection with Hand-Held Computers: Contributions to Questionnaire Design," Journal of Official Statistics, Vol. 4, No. 2, 1988, pp. 141-154.

50 Weeks, Michael F., "Call Scheduling with CATI:, Current Capabilities and Methods," in Robert M. Groves at al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988.

51 Werking, George, Alan Tupek and Richard Clayton, "CATI and Touchtone Self-Response Applications for Establishment Surveys," Journal of Official Statistics, Vol. 4, No. 4, 1988, pp. 349-362.

DISCUSSION

William L. Nicholls II

U. S. Bureau of the Census

Marc Tosiano's paper has a didactic purpose. He presents basic information about CATI and related topics as background for a more technical paper on computer assisted survey information collection (CASIC) to follow. Since his paper is primarily a condensation of summary articles on CATI and CAPI previously prepared by others, it contains much that is familiar and little that is original. Rather than add another layer of commentary to this well worked material, I will use the discussant's time to counterpose the tone of technological and methodological optimism which seems to characterize many papers of this conference with some historical reality. This also will be familiar material to some readers, since it is based on the same sources as Tosiano's paper.

CATI and its associated technologies provide many opportunities to improve the timeliness and quality of survey data, often at the same or lower cost per case (Catlin and Ingram, 1988; Nicholls and Groves 1986; Groves and Nicholls 1986). But those increasingly documented benefits have not necessarily prompted Federal data collection agencies to implement CATI expeditiously in their major surveys or in ways that optimize those benefits.

The first CATI survey was conducted by Chilton Research in 1971; and by 1980 CATI was in widespread use in commercial market research and in university survey research (Nicholls, 1988). But even those Federal agencies moving most quickly, such as NASS, will not fully implement CATI in the their major continuing surveys before 1992. That will be 21 years, or a full generation, after CATI was invented. For the Census Bureau's major household surveys, such as the Current Population Survey and the National Crime Survey, the earliest conceivable date for full CATI (and CAPI) implementation is 1994, but slippage, say to 1996, seems increasingly likely under current budgetary constraints. That would

represent a quarter century after the first CATI survey in the private sector. Federal agencies have introduced CATI more quickly into new and infrequently conducted surveys. But why has it taken so long to implement CATI for major, continuing Federal surveys?

There are many reasons. In the early 1970s, according to Dillman and Tarnai (1988), the managers of most Federal surveys regarded the telephone interview as a generally inferior data collection method and were reluctant to try it. Where a readiness for change was present, the technology often was lacking. CATI software was initially designed for market research and was not adequate for many government applications until enhanced by university organizations with government support in the late 1970s.

When U.S. government agencies began active internal development of CATI, around 1980, they often started with research programs to assess its effects on costs, data quality, and estimates. This research still continues, although both the Census Bureau and Statistics Canada produced major summaries of results in the late

1980s (U.S. Census Bureau, 1987; and Catlin and Ingram, 1988). The familiar delays of government planning, budgeting, and procurement also undoubtedly played a role in delaying CATI implementation.

CATI's extended incubation period in government also may be partly explained by its initial association with two related methodologies, random digit dialing (RDD) and centralized telephone interviewing, which also are topics of this session. Together, RDD, centralized telephone interviewing, and CATI are sometimes described as "modern telephone methods." Their joint evolution was described by Groves and Kahn in their influential 1979 volume Surveys by Telephone as one of the major developments in the history of survey methods, ranking with area probability sampling and the use of computers for survey analysis. By 1980, Berry and O'Rourke (1988), among others, have noted that modern telephone methods (RDD, centralized, and with CATI) had become the dominant survey methods in U.S. commercial market research and in university survey research centers. Government agencies were the exception.

"Modern telephone methods" did not transfer readily to govern-

ment data collection as a package. This is most apparent for random digit dialing, whose potential to reduce survey costs attracted major interest among government statisticians (Biemer al., 1985). The National Center for Health Statistics, the Census Bureau, and Statistics Canada all began their investigations of modern telephone methods with years of careful testing of random digit dialing (Marquis and Blass, 1985). But, as Drew, Choudhry, and Hunter (1988) have observed, RDD sampling methods are used in few government surveys conducted in the U.S. or elsewhere in the world. The omission of nontelephone households (about 7 percent of the U.S. total) and the typically higher refusal rates of cold contact telephone interviews have presented major barriers to the use of RDD in many or most government survey applications.

Random digit dialing remains a valuable sampling method for populations with high telephone subscribership such as Canada and Sweden) and for surveys which can tolerate its coverage and nonresponse problems. For some governmental statistical agencies, however, the early emphasis on RDD proved a diversion from what now appear to be more fruitful uses of CATI. Only when RDD was ruled

out as a sampling method for most U.S. government household surveys, which at the Census Bureau occurred around 1986, could plans to implement CATI in single-frame, mixed mode designs proceed. The somewhat faster adoption of CATI by establishment and agricultural surveys may be partly attributable to their traditional reliance on list frame samples. A change to RDD was not an issue.

The second major element of modern telephone methods which has not translated easily to government data collection is centralized telephone interviewing. In U.S. university and commercial market research, the shift from "dispersed" local, interviewers making calls from their own homes to "centralized" telephone interviewers calling from large national or regional offices was largely completed by the late 1970s. Government household surveys are one of the few major users of dispersed telephone interviewing persisting into the 1980s.

Mr. Tosiano's paper has reviewed the ways in which centralized telephone interviewing and CATI can be mutually supporting methodologies. Computer-assistance is easier to arrange for centralized interviewers who share the same hardware, programs, sample, and technical staff. At the same time, CATI encourages centralized interviewing to gain these efficiencies and to benefit from such large-staff CATI features as automatic call scheduling,, online supervision, and field report generation. Centralization contributes to standardized field procedures and interviewing quality control through easier recruitment, training, and supervision of interviewers, while CATI contributes to these same goals through tailored question wordings, computer controlled branching, and online editing. Supervisory audio-visual monitoring of interviewer performance, currently feasible only with centralized CATI interviewers, provides feedback ensuring that CATI quality enhancement features are appropriately used and that interviewers deviating from performance standards are identified and retrained when necessary.

"Centralization" has a different meaning for government establishment surveys than for government household surveys.

Because the establishment surveys typically began with mailed questionnaire methods, later supplemented with telephone prompting and interviews, they generally are conducted from offices. The choice typically is between national, regional, and state offices. The introduction of CATI strengthens the arguments for greater centralization. The Census Bureau's Business Division is perhaps unique among Federal agencies in withdrawing its Retail and Whole-sale Trade Surveys from a set of regional offices to centralize them in a national site before placing them on CATI. More commonly, existing organizational arrangements, Federal-State agreements, and formal or informal commitments to employees have resulted in continuation of state-based offices averaging about 10 interviewing stations per state but ranging from 2 to perhaps 30 stations (Nicholls 1988). In national private sector survey and market research organizations, CATI installations more typically reach 45-100 stations.

The introduction of CATI into mixed-mode personal-telephone household surveys presents even greater organizational problems. This is illustrated by the Census Bureau's plans to phase CATI into

the Current Population Survey (CPS) and the National Crime Survey

(NCS). Both surveys have a rotating panel design. The first visit to each sample address is by personal visit to identify ineligible housing units and to encourage household participation. The fifth CPS and NCS visits also are in person to re-establish personal contact with the household part way through the sequence of interviews. Other interviews are by telephone when possible and acceptable to the respondent and by personal visit otherwise. The same local interviewers traditionally conduct both the personal and telephone visit interviews, placing the telephone interview calls from their own homes.

When CATI is introduced into these surveys, no change is made in the initial visits to each sample address. These remain personal visit interviews since comparable response and panel retention rates have not been attainable with cold contact telephone interviews (Marquis and Blass, 1985). CATI replaces

dispersed telephone interviews from the local interviewers' homes in the second and later visits of the panel design. This field design has several potential benefits: (1) reduced field costs; (2) reduced interviewer recruitment problems in tight labor markets; and (3) possibly improved survey estimates. Nevertheless, the transition poses a number of design and organizational problems which require time and effort to resolve.

The first is developing appropriate methods for rapid but controlled transfer of individual case records between personal visit and CATI interview modes. When the first visit personal interview is complete, household enumeration data and field records must be data entered into computer files for second and later interviews by CATI. Case records also move from CATI to the local interviewers for CPS and NCS fifth visits and for personal followup of households unreachable by CATI.

The second transition problem is the temporarily reduced efficiency of the sample designs. Both the CPS and NCS employ cluster samples chosen initially to minimize costs for interviewing

assignments containing both personal visit and dispersed telephone interviews. When the dispersed telephone interviews are removed to CATI, the remaining personal visit cases may no longer constitute acceptable or efficient field assignments. Since the CPS and NCS samples are based on the decennial census, they are efficiently revised only once a decade.

The third problem in moving dispersed telephone interviews to CATI is the need to reduce the field staff while increasing the CATI staff. For the CPS, the Census Bureau's largest current survey, the transition will be based initially on field interviewer attrition and has been constrained by the rate at which attrition occurs.

The fourth and final transition problem is finding a sufficient volume of work for the CATI interviewers. The CPS

conducts its interviews in the third week of each month and the NCS

in the first week with some carryover into the second. Centralized CATI interviewing is restricted to even fewer days per month to permit field followup of cases unreachable by telephone. These two surveys will provide the CATI staff with relatively few days of employment per month.

Of the four transition problems, only the first derives from the CATI technology. Case transfers between dispersed local interviewers and centralized CATI interviewers are complicated by the move between paper-and-pencil records and computer files. The problems of field sampling efficiency, field staff phase-down, and insufficient work at the CATI facilities arise from the centralization of previously dispersed interviews. They would be the same whether the central facility used CATI or paper-and-pencil methods.

The most difficult problems of implementing CATI in government agencies appear to derive from the organizational issues CATI typically raises about centralized vs. decentralized interviewing.

References

Biemer, P., D.W. Chapman, and C. Alexander, "Some Research Issues in Random-Digit Dialing Sampling and Estimation," Proceedings of the Bureau of the Census First Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1985, pp. 71-86.

Berry, S. H. and D. O'Rourke, "Administrative Designs. for Centralized Telephone Survey Centers: Implications of the Transfer to CATI," in R. M. Groves et al. (editors) Telephone Survey Methodology, New York, Wiley Press, 1988, pp. 457-474.

Catlin, G. and S. Ingram, "The Effect of CATI on Cost and Data Quality: A comparison of CATI and Paper Methods in Centralized Interviewing," in R. M. Groves et al. (editors) Telephone Survey Methodology, New York, Wiley Press, 1988, pp. 437-452.

Dillman, D. A. and J. Tarnai, "Administrative Issues in Mixed Mode Surveys," in R. M. Groves et al. (editors), Telephone Survey Methodology, New York, Wiley Press, 1988, pp. 509-528.

Drew, J. D., G. H. Choudhry, and L. A. Hunter, "Nonresponse Issues in Government Telephone Surveys," in R. M. Groves et al. (editors) Telenhone Survey Methodology, New York, Wiley Press, 1988, pp. 233-246.

Groves, R. M. and R. L. Kahn, Surveys by Telephone, New York, Academic Press, 1979.

Groves, R. M. and W. L. Nicholls II, "The Status of Computer-Assisted Telephone Interviewing: Part II -- Data Quality Issues," Journal of Official Statistics, Vol. 2, No. 2, 19860 pp.,117-134.

Marquis, K. and R. Blass, "Nonsampling Error Considerations in the Design and Operation of Telephone Surveys," Proceedings of the Bureau of the Census First Annual Research Conference, Washington, U.S. Department of Commerce, Bureau of the Census, 1985, pp. 301-329.

Nicholls II, W. L., "The Impact of High Technology on Survey Data Collection," CATI Research Report GEN-1, U.S. Department of Commerce, Bureau of the Census, February 1989.

Nicholls II, W.L. and Groves, R.M., "The Status of Computer-Assisted Telephone Interviewing: Part I -- Introduction and Impact on Cost and Timeliness of Survey Data," Journal of Official Statistics, Vol. 2, No. 2, 1986, pp. 93-115.

U.S. Census Bureau, Evaluation of CATI Data Ouality and Costs in the Current Population Survey, CATI Research Report No. CPS-2 of the Computer-Assisted Interviewing Central Planning Committee, CATI Research and Analysis Subcommittee, September 1988.

DISCUSSION

James T. Massey

National Center for Health Statistics

The paper by Leyla Mohadjer and David Morganstein enumerates and provides a brief overview of almost all of the key methodological issues related to telephone surveys. The concept of total survey design mentioned in the first section of the paper is an excellent way to compare and summarize the advantages and disadvantages of telephone surveys versus other modes of data

collection. The total survey design concept was never fully developed to compare the different modes of data collection. Most of this paper focused on the operational and sample design efficiencies of telephone surveys to improve data quality.

The advantages and disadvantages of telephone surveys given by Mohadjer and Morganstein are listed below along with several additional ones:

Advantages of Telephone Surveys

- Lower cost

- Better quality control and supervision of interviewers

- Better access to some hard to reach persons

- Smaller design effects

- smaller interviewer effects

- Cost effective method to sample rare population (use of Donnally tape with characteristics of persons in prefix area)

- Use of CATI to control flow of sample, interview, edits, and processing

- Local area surveys from central location

- Better use of bilingual interviewers

Disadvantages of Telephone Surveys

- Lack of visual aids

- No group interviews

- Noncoverage of persons without telephones

- Lower response rates

- Cost of dual frame surveys

- Cost of CATI (relative to other telephone surveys)

Now I would like to turn my attention to where we are in the development and use of telephone surveys and some areas that still need research.

I see six reasons for the emergence of telephone surveys:

1) Better coverage

2) Development of CATI which lead to development of CAPI

3) Development of better RDD methods

4) Higher costs to face-to-face surveys

5) Slow dea h of the myth of the length of a telephone
interview

6) Recognition of data quality equal to face-to-face surveys

Considerable progress has been made over the past 15 years in almost every aspect of telephone surveys including data quality. There are, however, several areas where progress has been limited and more research is needed. These are listed below.

1) Techniques to improve response rates: While response rates have improved, there is still much that could be

done to adapt procedures in the face-to-face surveys to telephone surveys. I just reviewed a paper that used several inducement techniques to dramatically improve telephone survey response rates in another country.

2) Validation of data collected by telephone versus face-to-face surveys: Most comparative studies have assumed that higher levels of reporting is better. For some types of data this assumption is questionable and additional statistical validation studies are needed.

3) Research on the collection of sensitive data and other, specific types gf informatign: We should take full advantage of one of the key features of telephone interviewing, the autonomy and anonymity of the interview. Some research has been done that showed sensitive data and questions have socially desirable

responses are obtained better over the telephone. There is some recent unpublished data that indicates that smoking habits, crimes, and unemployment may have higher reporting over the telephone. These results should be published and validated.

4) Research on difficult questions and questions with multiple response: Questions requiring flashcards and scaled responses are still more problematic over the telephone. CATI does offer a way to randomize the order of responses.

5) Research on better and cheaper ways to correct for noncoverage.

Finally, I would like to make two other observations. In 1984

when the OMB Working Paper 12 on Telephone Data Collection was published, telephone surveys were primarily used in the Federal government to conduct follow-up surveys and follow-up interviews. Most initial contact surveys by telephones used list frames. This is still the case today for almost all of the large government surveys, although greater use of telephone interviewing is being made.

For those of you who are new to the study of telephone survey, I recommend you start with the book Telephone Survey Methodology. It has several state-of-the-art review papers and has an extensive bibliography. The paper Owen Thornberry and I wrote contains as many reference tables on telephone coverage as Bob Groves would allow us to include. I hope many of you will continue to conduct research on telephone surveys and extend our knowledge of this very valuable data collection method.

1 For a copy of the latest version, write to Dr. Daniel Kasprzyk, Chief, SIPP Research and Coordination Staff, Office of the Director, Bureau of the Census, Washington DC 20233.

2 By Fritz Scheuren, Director, Statistics of Income Division (R:S), Internal Revenue Service. Based, in part, on a Discussion of "Rolling Samples and Censuses," by Leslie Kish, to appear in the June 1990 issue of Survey Methodology. The views expressed in this paper are those of the author and do not necessarily represent the position of the Internal Revenue Service.

3 This is a summary of a longer paper that was prepared for the Seminar on Quality of Federal Data.

4 A recent bibliography of the staff papers and reports prepared on the use of administrative records for social data in the Small Area and Administrative Data Division was recently completed. (Statistics Canada, 1990)

5 Unmarried persons who (a) declare themselves to be single, (b) are under the age of 30, (c) reside with their parents and (d) file a tax return are defined to be "filing children".

6 To minimize the T1FF data processing costs, most of the T1FF data in this paper are based on samples.

7 In processing the 1986 tax file, a somewhat earlier file was used than in other years. As a result, the coverage was lower than in other years. Had this not occurred, the coverage

in 1986 would have been higher than 93.7%.

8 The Tl does contain some information on dependent children, namely, relationship to taxfiler and birthdate. This information is not, however, captured.

9 The taxfiling rate for the 65+ population increased from 60% in 1985 to 75% in 1987.

10 The SCF is an annual supplement to the Canadian Labour Force Survey. The SCF is similar to the March supplement to the Current Population Survey (CPS) in the United States.

11 This table was adapted from Vigder and Leyes (1989).

12 These remarks are attributable to the author and do not necessarily represent the views of the Census Bureau.

13 This paper is a condensation of Survey Coverage, Statistical Policy Working Paper 17. Authors are listed in the second

paragraph. The views expressed are those of the authors and do not necessarily reflect those of their agencies.

Reports Available in the Statistical Policy Working Paper series

1. Report on Statistics for Allocation of Funds (Available through NTIS Document Sales, PB86-211521/AS)
2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques (NTIS Document Sales, PB86-211539/AS)
3. An Error Profile: Employment as Measured by the Current Population Survey (NTIS Document Sales PB86-214269/AS)
4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (NTIS Document Sales, PB86-211547/AS)
5. Report on Exact and Statistical Matching Techniques (NTIS Document Sales, PB86-215829/AS)

6. Report on Statistical Uses of Administrative Records (KTIS Document Sales, PB86-214285/AS)

7. An Interagency Review of time-Series Revision Policies (NTIS Document Sales, PB86-232451/AS)

8. Statistical Interagency Agreements (NTIS Documents Sales, PB$6-230570/AS)

9. Contracting for Surveys (NTIS Documents Sales, PB83-233148)

10. Approaches to Developing Questionnaires (NTIS Document Sales, PB84-105055/AS)

11. A Review of Industry Coding Systems (NTIS Document Sales, PB84-135276)

12. The Role of Telephone Data Collection in Federal Statistics (NTIS Document Sales, PB85-105971)

13. Federal Longitudinal Surveys (NTIS Documents Sales, PB86-139730)

14. Workshop on Statistical trees of Microcomputers in Federal Agencies (NTIS Document Sales, PB87-166393)

15. Quality in Establishment Surveys (NTIS Document Sales, PB88-232921)

16. A Comparative Study of Reporting Units in Selected Employer Data Systems (NTIS Document Sales, PB-90-205238)

17. Survey Coverage (NTIS Document Sales, PB90-205246)

18. Data Editing in Federal Statistical Agencies (NTIS Document Sales, PB90-205253)

19. Computer Assisted Survey Information Collection (NTIS Document Sales, PB90-205261)

20. Seminar on the Quality of Federal Data (NTIS Document Sales, PB91-142414)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161 (703) 487-4650

Statistical Policy Working Paper 20 - Seminar on Quality of Federal Data - Part 2 of 3

Statistical Policy

Working Paper 20

Seminar on Quality of Federal Data

Part 2 of 3

Federal Committee on Statistical Methodology

Statistical Policy Office

Office of Information and Regulatory Affairs

Office of Management and Budget

March 1991

MEMBERS OF THE FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY

(February 1991)

Maria E. Gonzalez, Chair

Office of Management and Budget

Yvonne M. Bishop

Energy Information Administration

Daniel Kasprzyk

Bureau of the Census

Daniel Melnick

Warren L. Buckler
Social Security Administration

Charles E. Caudill
National Agricultural
Statistics Service

Cynthia Z.F. Clark
National Agricultural
Statistics Service

Zahava D. Doering
Smithsonian Institution

Robert M. Groves
Bureau of the Census

Roger A. Herriot
National Center for
Education Statistics

National Science Foundation

Robert P. Parker
Bureau of Economic Analysis

David A. Pierce
Federal Reserve Board

Thomas J. Plewes
Bureau of Labor Statistics

Wesley L. Schaible
Bureau of Labor Statistics

Fritz J. Scheuren
Internal Revenue Service

Monroe G. Sirken
National Center for
Health Statistics

C. Terry Ireland
National Computer Security Center

Robert D. Tortora
Bureau of the Census

Charles D. Jones
Bureau of the Census

PREFACE

In 1975, the Office of Management and Budget (OMB) organized the Federal Committee on Statistical Methodology. Comprised of individuals selected by OMB for their expertise and interest in statistical methods, the committee has during the past 15 years. determined areas that merit investigation and discussion, and overseen the work of subcommittees organized to study particular issues. Since 1978, 19 Statistical Policy Working Papers have been published under the auspices of the Committee.

On May 23-24, 1990, the Council of Professional Associations on

Federal Statistics (COPAFS) hosted a "Seminar on the Quality of Federal Data." Developed to capitalize on work undertaken during the past dozen years by the Federal Committee on statistical Methodology and its subcommittees, the seminar focused on a variety of topics that have been explored thus far in the Statistical Policy Working Paper series. The subjects covered at the seminar included:

- Survey Quality Profiles
- Paradigm Shifts Using Administrative Records
- Survey Coverage Evaluation
- Telephone Data Collection
- Data Editing
- Computer Assisted Statistical Surveys
- Quality in Business Surveys
- Cognitive Laboratories
- Employer Reporting Unit Match Study
- Approaches to Developing Questionnaires
- Statistical Disclosure-Avoidance
- Federal Longitudinal Surveys

Each of these topics was presented in a two-hour session that featured formal papers and discussion, followed by informal dialogue among all speakers and attendees.

Statistical Policy Working Paper 20, published in three parts, presents the proceedings of the "Seminar on the Quality of Federal Data." In addition to providing the papers and formal discussions from each of the twelve sessions, this working paper includes Robert M. Groves' keynote address, "Towards Quality in a Working Paper Series on Quality," and comments by Stephen E. Fienberg, Margaret E. Martin, and Hermann Habermann at the closing session, "Towards an Agenda for the Future."

We are indebted to all of our colleagues who assisted in organizing the seminar, and to the many individuals who not only presented papers and discussions but also prepared these materials for publication. A special thanks is due to Terry Ireland and his staff for their work in assembling this working paper.

Table of Contents

Wednesday, May 23, 1990

Part 1

KEYNOTE ADDRESS

TOWARDS QUALITY IN A WORKING PAPER SERIES ON QUALITY. 3

Robert M. Groves, The University of Michigan and U. S.

Bureau of the Census

Session 1 - SURVEY QUALITY PROFILES

THE SIPP QUALITY PROFILE. 19

Thomas B. Jabine, Statistical Consultant

INITIAL REPORT ON THE QUALITY OF AGRICULTURAL SURVEY PROGRAM. 29

George A. Hanuschak, National Agricultural Statistics

Service

DISCUSSION. 40

Barbara A. Bailar, American Statistical Association

DISCUSSION. 46

Nancy A. Mathiowetz, U. S. Bureau of the Census

Session 2 - PARADIGM SHIFTS USING ADMINISTRATIVE

RECORDS

PARADIGM SHIFTS: ADMINISTRATIVE RECORDS AND CENSUS-TAKING. . . 53

Fritz Scheuren, Internal Revenue Service

AN ADMINISTRATIVE RECORD PARADIGM: A CANADIAN EXPERIENCE . . . 66

John Leyes, Statistics Canada

DISCUSSION. 77

Gerald Gates, U.S. Bureau of the Census

DISCUSSION. 83

Edward J. Spar, Market Statistics

Session 3 - SURVEY COVERAGE EVALUATION

CONTROL MEASUREMENT, AND IMPROVEMENT OF SURVEY COVERAGE . . .87

Gary M. Shapiro, U. S. Bureau of the Census; Raymond R. Bosecker, National Agricultural Statistics Service

QUALITY OF SURVEY FRAMES. 100

Judith T. Lessler, Research Triangle Institute

DISCUSSION. 108

Fritz Scheuren, Internal Revenue service

DISCUSSION. 114

Joseph Waksberg, Westat, Inc.

Session 4 - TELEPHONE DATA COLLECTION

QUALITY IMPROVEMENT IN TELEPHONE SURVEYS. 123

Leyla Mohadjer, David Morganstein, Westat, Inc.

COMPUTER ASSISTED SURVEY TECHNOLOGIES IN GOVERNMENT:

AN OVERVIEW. 137

Marc Tosiano, National Agricultural Statistics Service

DISCUSSION 155

William L. Nicholls II, U. S. Bureau of the Census

DISCUSSION. 161
.161

James T. Massey, National Center for Health Statistics

Part 2

Session 5 - DATA EDITING

OVERVIEW OF DATA EDITING IN FEDERAL STATISTICAL AGENCIES .167

David A. Pierce, Federal Reserve Board

EDITING SOFTWARE (An excerpt from Chapter IV of Working-

Paper 18). 173

Mark Pierzchala, National Agricultural Statistics

Service

RESEARCH ON EDITING. 180

Yahia Ahmed, Internal Revenue Service

DISCUSSION. 184

Charles E. Caudill, National Agricultural Statistics Service

DISCUSSION. 186

Richard Bolstein, George Mason University

Session 6 - COMPUTER ASSISTED STATISTICAL SURVEYS

OVERVIEW OF COMPUTER ASSISTED SURVEY INFORMATION COLLECTION. 191

Richard L. Clayton, U. S. Bureau of Labor Statistics

A COMPARISON BETWEEN CATI AND CAPI. 197

Martin Baum, National Center for Health Statistics

COMPUTER ASSISTED SELF INTERVIEWING. 202

Ralph Gillmann, Energy Information Administration

COMPUTER ASSISTED Self INTERVIEWING: RIGS AND PEDRO,
TWO EXAMPLES 205
Ann M. Ducca, Energy Information Administration

DATA COLLECTION. 209
Cathy Mazur, National Agricultural Statistics Service

DISCUSSION. 212
Robert N. Tinari, U. S. Bureau of the Census

DISCUSSION. 216
David Morganstein, Westat, Inc.

Thursday, May 24, 1990

Session 7 - QUALITY IN BUSINESS SURVEYS

IMPROVING ESTABLISHMENT SURVEYS AT THE BUREAU OF LABOR

STATISTICS 221

.Brian MacDonald, Alan R. Tupek, U. S.Bureau of Labor

Statistics

A REVIEW OF NONSAMPLING ERRORS IN FEDERAL ESTABLISHMENT

SURVEYS WITH SOME AGRIBUSINESS EXAMPLES. 232

Ron Fecso, National Agricultural Statistics Service

DISCUSSION. 243

David A. Binder, Statistics Canada

DISCUSSION. 247

Charles D. Cowan, Opinion Research Corporation

Session 8 - COGNITIVE LABORATORIES

THE BUREAU OF LABOR STATISTICS' COLLECTION PROCEDURES RESEARCH LABORATORY: ACCOMPLISHMENTS AND FUTURE DIRECTIONS . .253

Cathryn S. Dippo, Douglas Herrmann, U. S. Bureau of Labor Statistics

THE ROLE OF A COGNITIVE LABORATORY IN A STATISTICAL AGENCY. . 268

Monroe G. Sirken, National Center for Health Statistics

DISCUSSION. 278

Elizabeth Martin, U. S. Bureau of the Census

DISCUSSION. 281

Murray Aborn, National Science Foundation (retired)

Part 3

Session 9 - EMPLOYER REPORTING UNIT MATCH

STUDY

INTERAGENCY AGREEMENTS FOR MICRODATA ACCESS:

THE ERUMS EXPERIENCE. 291

Thomas B. Petska, Internal Revenue Service; Lois

Alexander, Social Security Administration

SAMPLE SELECTION AND MATCHING PROCEDURES USED IN ERUMS. . . 301

John Pinkos, Kenneth LeVasseur, Marlene Einstein,

U. S. Bureau of Labor Statistics; Joel Packman, Social

Security Administration

RESULTS, FINDINGS, AND RECOMMENDATIONS OF THE ERUMS PROJECT. 309 .309

Vern Renshaw, Bureau of Economic Analysis; Tom Jabine,

Statistical Consultant

DISCUSSION. 318

W. Joel Richardson, Charles A. Waite, U. S. Bureau of the Census

DISCUSSION. 324

Thomas J. Plewes, U. S. Bureau of Labor Statistics

Session 10 - APPROACHES TO DEVELOPING QUESTIONAIRES

TOOLS FOR USE IN DEVELOPING QUESTIONS AND TESTING QUESTIONNAIRES. 331

Theresa J. DeMaio, U. S. Bureau of the Census

TECHNIQUES FOR EVALUATING THE QUESTIONNAIRE DRAFT. 340

Deborah H. Bercini, National Center for Health Statistics

DESIGNING QUESTIONNAIRES FOR CATI IN A MIXED MODE ENVIRONMENT. 349

Gemma Furno, U. S. Bureau of the Census

DISCUSSION 360

Carol C. House, National Agricultural Statistics Service

Session 11 - STATISTICAL DISCLOSURE - AVOIDANCE

DISCLOSURE AVOIDANCE PRACTICES AT THE CENSUS BUREAU. 367

Brian Greenberg, U. S. Bureau of the Census

THE MICRODATA RELEASE PROGRAM OF THE NATIONAL CENTER

FOR HEALTH STATISTICS 377

Robert H. Mugge, National Center for Health Statistics

(retired)

DISCUSSION. 385

George Duncan, Carnegie Mellon University

Session 12 - FEDERAL LONGITUDINAL SURVEYS

FEDERAL LONGITUDINAL SURVEYS 393

Daniel Kasprzyk, U. S. Bureau of the Census; Curtis Jacobs, U. S. Bureau of Labor Statistics

THE ADVANTAGES AND DISADVANTAGES OF LONGITUDINAL SURVEYS. . ..407

Robert W. Pearson, Social Science Research Council

LONGITUDINAL ANALYSIS OF FEDERAL SURVEY DATA. 425

Patricia Ruggles Joint Economic Committee

DISCUSSION. 438

Michael Brick, Westat, Inc.

DISCUSSION. 447

Marilyn E. Manser, U. S. Bureau of Labor Statistics

TOWARDS AN AGENDA FOR THE FUTURE

Stephen E. Fienberg, Carnegie Mellon University 455

Margaret E. Martin. 462

Hermann Habermann, Office of Management and Budget. 465

Part 2

Session 5

DATA EDITING

OVERVIEW OF DATA EDITING IN FEDERAL STATISTICAL AGENCIES

David A. Pierce

Federal Reserve Board

Abstract

This paper is the first of three in the session on Data Editing presenting highlights of the report "Data Editing in

Federal Statistical Agencies", Statistical Policy Working Paper 18, OMB, prepared by the Subcommittee on Data Editing in Federal Statistical Agencies, FCSM. Included in this paper are a listing of the Subcommittee members, a discussion of its mission statement from the FCSM, definition and concepts of data editing, the major areas investigated and the methods used to do so, the development of case studies, and the Subcommittee's recommendations for data editing in Federal statistical agencies. The paper highlights the findings from a survey of current data editing practices which was conducted by the Subcommittee.

1. Introduction

The Subcommittee on Data Editing in Federal Statistical Agencies was established by the Federal Committee on Statistical Methodology (FCSM) in November 1988 to document, profile, and discuss the topic of data editing in Federal censuses and surveys. The Subcommittee consisted of the following individuals:

George Hanuschak, National Agricultural Statistics Service,

Chair

Yahia Ahmed, Internal Revenue Service

Laura Bauer, Federal Reserve Board

Charles Day, Internal Revenue Service

Maria Gonzalez, Office of Management and Budget

Brian Greenberg, Bureau of the Census

Anne Hafner, National Center for Education Statistics

Gerry Hendershot, National Center for Health Statistics

Rita Hohenbrink, National Agricultural Statistics Service

Renee Miller, Energy Information Administration

Tom Petkunas, Bureau of the Census

David Pierce, Federal Reserve Board

Mark Pierzchala, National Agricultural Statistics Service

Marybeth Tschetter, Bureau of Labor Statistics

Paula Weir, Energy information Administration

A key aim of this effort was to further the awareness within agencies of each other's data editing practices, as well as of the state of the art of data editing, and thus to promote improvements in data quality throughout Federal statistical agencies. To further these goals, the Subcommittee was given a "charge", or mission statement, of

> determining how data editing is currently being done in Federal agencies, recognizing areas that may need attention, and, if appropriate, recommending any potential improvements for the editing process.

Among the many items investigated by the Subcommittee were the role of subject matter specialists; hardware, software, and the data base environment; new technologies of data collection and editing, such as CATI and CAPI; current research efforts in the various agencies; and some recently developed editing systems such as at

the Census Bureau and Statistics Canada.

In fulfilling its mission the Subcommittee followed a number of paths, including developing a questionnaire on survey editing practices, assembling several case studies of editing practices, investigating alternative editing systems and software, exploring research needs and practices, and compiling an annotated bibliography of literature on editing. The result of the Subcommittee's work is its report (1990), organized into 5 main chapters with several supporting appendices as follows:

Chapters	Appendices
I. Executive Summary	A. Questionnaire Responses
II. Background	B. Case Studies
III. Current Editing Practices	C. Software Functions checklist
IV. Editing Software	D. Annotated Bibliography
V. Research on Editing	E. Glossary of Terms

After discussing some general topics pertaining to editing and to

the Subcommittee's work, this paper summarizes some of the main results of a questionnaire on Current Editing Practices, designed, administered and compiled by the Subcommittee. The two papers immediately following address, respectively, the subjects of software developments and recent research findings in editing.

2. Data Editing--Definition and Concepts

The subcommittee first addressed the definition of data editing. While no universal definition of survey data editing exists, the following working definition was developed:

> Procedures designed and used for detecting erroneous and/or questionable survey data, with the goal of correcting (manually or electronically) as much of the erroneous data as possible (not necessarily all of the questioned data), usually prior to data imputation and summary procedures.

Thus data editing can be seen as a data quality improvement tool by which erroneous or highly suspect data are found and (if necessary) corrected. We have focused primarily on editing rather than imputation in our work, though in practice the boundary between these is not absolute.

3. Current-Editing Practices

To obtain a profile of current editing practices, in the various Federal statistical agencies, the subcommittee developed an editing questionnaire, which was completed for 117 Federal censuses and surveys representing 14 different Federal agencies. These 117 surveys were selected by subcommittee members, and thus they were not a scientific sample of all Federal surveys; however the

Subcommittee felt that the 117 surveys represented a broad coverage of agencies and types of surveys or censuses that would present different editing situations.

The Subcommittee members primarily involved with the questionnaire and editing profile were Charles Day, Yahia Ahmed, George Hanuschak, Rita Hohenbrink and Renee Miller.

The questionnaire that was designed was a six-page document containing general questions about the particular survey as well as specific questions on editing. The report contains a complete listing of the questions asked, along with a tally of the results obtained for the 117 surveys, and should serve as a useful reference for the current (1990) state of data editing practice. A few of the major results follow.

Regarding general characteristics of the surveys, about three-fourths of the surveys are actually sample surveys, and the remaining one-fourth censuses. A wide range of frequencies of collection are represented, from daily to quinquennial. About one-

fourth are completed by individuals, and three-fourths by establishments. While traditional means of data collection such as mail, personal and telephone interviews were most common, a small

proportion of the surveys used CATI, and some were administrative records.

Turning to editing, while the idea that there's no such thing as a free lunch seems to be as true of data editing as it is of anything else, there was wide variation in the actual cost of editing as a percent of total survey cost. The median editing cost for the surveys was more than one-third of the total cost of the survey. One of the interesting findings was that surveys of individuals had lower relative editing costs than surveys of establishments.

The questionnaire also elicited information on when in the survey process the editing occurs. For about two-thirds of the 117

surveys, most of the data editing takes place after data entry. Editing at the time of data entry is on the increase but not yet common.

Subject matter analysts play a large and important role in data editing. In about three-fourths of the surveys, subject matter analysts review all unusual or large cases. Only seven of the surveys had little or no intervention by subject-matter specialists. In this regard, we found that surveys of establishments had heavier involvement from subject-matter specialists than surveys of individuals; and this could also be related to the, finding, mentioned above, of lower editing costs in individual than in establishment surveys.

The degree of automation in data editing varies considerably among the surveys in our study. In about three-fifths of the surveys, automated edit checking is done, but error correction is performed by clerks or analysts. In about one-third of the cases, only unusual situations are referred to analysts. Only 3% of the surveys were totally automated, though all but 1% had at least some

automation.

There are different types of edits that are applied to surveys. Almost all the surveys in our study use validation editing, which detects inconsistent data within a record. About five-sixths also use macro editing, where aggregated data are examined. The majority of surveys use other types of edits as well, such as range edits, edits using historical data, ratio edits, some of which may overlap. Additional information is also utilized in editing many of the surveys, such as comparisons with other surveys, comparison to a value estimated by regression analysis, or the use of interquartile measures.

Satisfaction with the current editing system varied widely. About half the respondents were satisfied with their current editing systems, and another one-fourth felt only minor changes were needed. The remaining one-fourth thought major changes were desired, with 5% of those being in favor of a complete overhaul.

Among those desiring improvements, those most frequently mentioned were:

- an on-line system for data editing,
- the use of prior periods' data to test the current period,
- more statistical edits,
- more sophisticated validation and macro editing,
- an audit trail,
- more automation, particularly automated error correction,
- user-friendlier systems,
- incorporation of imputation into the error package,
- evaluation of effects of data editing,
- reduction of the number of edit flags to follow up,
- incorporation of information on auxiliary variables,
- greater use of Expert Systems, and
- multivariate editing.

An Audit trail, or a complete record of the original and corrected data, the edits failed and any other relevant information, is very helpful in monitoring and improving the editing process. The

importance of an evaluation of the effects of editing on the data, and our current lack of knowledge of such effects, have also been noted by Bailar (1990).

4. Case Studies

In addition to the breadth of valuable information obtained from the questionnaire, the Subcommittee also felt that an examination of a relatively few surveys in greater depth would shed light on the complexity of the different editing situations in operation. Therefore several case studies are described, some in two-paragraph summary format and others in greater detail. These comprise Appendix B of the report. Anne Hafner and Yahia Ahmed had primary responsibility for preparation of the Case Studies.

5. Recommendations

The report lists a number of recommendations for future data editing practice, some general and some specific. Many of them fall into the following general categories.

The quality of an agency's existing editing practices and technology should be examined in the light of possible improvements or alternatives, with respect to such criteria as cost efficiency, timeliness, statistical defensibility, and accuracy.

Important recent developments in data processing, such as new microcomputers, workstations, local area networks, data base software, and mainframe linkages, should be

examined for their possible incorporation into the survey editing process.

Agencies should stay in communication with each other and with other professionals regarding their research in editing, particularly the development and implementation of new editing procedures and related methodologies such

as data base technologies and expert systems.

References

Bailar, Barbara (1990), "Discussion of 'Survey Quality Profiles'", Seminar on the Quality of Federal Data, May 22, 1990, COPAFS. This Proceedings.

Groves, Robert (1990), "Towards Quality in a Working Paper Series on Quality", Keynote Address, Seminar on the Quality of Federal Data, May 22, 1990, COPAFS This Proceedings.

Hanuschak, George, Yahia Ahmed, Laura Bauer, Charles Day, Maria Gon-zalez, Brian Greenberg, Anne Hafner, Gerry Hendershot, Rita Hohenbrink, Renee Miller, Tom Petkunas, David Pierce, Mark Pierzchala, Marybeth Tschetter and Paula Weir (1990), Data Editing in Federal Statistical Agencies, Statistical Policy Working Paper 18, Statistical Policy Office, Office of Management and Budget, Washington, DC.

EDITING SOFTWARE

(An excerpt from Chapter IV of Working Paper 18)

Mark Pierzchala

National Agricultural Statistics Service

A. Introduction

For most surveys, large parts of the editing process are carried out through the use of computer systems. The task of the Software Subgroup has been to investigate software that in some way

incorporates new methodologies, has new ways of presenting data, operates in recently developed hardware environments, or integrates editing with other functions. In order to fulfill this charge, the Subgroup has evaluated or been given demonstrations of new editing software. In addition, the Subgroup has developed an editing software evaluation checklist that appears in Appendix C of Statistical Policy Working Paper 18. This checklist contains possible functions and attributes of editing software, which would be useful for an organization to use when evaluating editing software.

Extremely technical jargon can be associated with new editing systems; and new approaches to editing may not be familiar to the reader. The purpose of section B is to explain these approaches and their associated terminology as well as to discuss briefly the role of editing in assuring data quality.

A distinction must be made between generalized systems and software meant for one or a few surveys. The former is meant to be used for a variety of surveys. Usually there is an institutional

commitment to spend staff time and money over several years to develop the system. It is hoped that the investment will be more than recaptured after the system is developed through the reduction in resources spent on editing itself and in the elimination of duplication of effort in preparing editing programs. Some software programs have been developed that address specific problems in a particular survey. While the ideas inherent in this software may be of general interest, it may not be possible to apply the software directly to other surveys. Section C of Chapter IV of Working Paper 18 describes three generalized systems in some detail, and then briefly describes other systems and software. These three systems have been used or evaluated by Subgroup members in their own surveys.

New and exciting statistical methodology is also improving the editing process. This includes developments in detecting outliers, aggregate level data editing, imputation strategy, and statistical quality control of the process itself. The implementation of these activities, however, requires that the techniques be encoded into a computer program or system.

B. Software Improving Quality and Productivity

Reasons for the Development of New Editing Software

Traditional editing systems do not fully utilize the talents or expertise of subject matter specialists. Much of their time may be spent in dealing with unimportant or spurious error signals and in coping with system shortcomings. As a result, the specialist has less time to deal with important problems. In addition, editing systems may be able to give feedback on the survey itself. For example, a pattern of edit failures may suggest misunderstandings by the respondent or interviewer. If this is recognized, then the expertise of the specialist may then be used to improve the survey itself.

Labor costs are a large part of the editing costs and are either steady or increasing, whereas the cost of computing is decreasing. In order to justify the heavy reliance on people in

editing, their productivity will have to be improved through the use of more powerful tools. However, even if productivity is improved, different people may do different things in similar situations. If so, this makes the process less repeatable (reproducible) and more subject to criticism. When work is done on paper, it is hard to track, and it is impossible to estimate the effect of editing actions on estimates. Finally, some tasks are beyond the capability of human editors. For example, it may be impossible for a person to maintain the multivariate frequency structure of the data when making changes.

These reasons and several others are commonly given as explanations for the increased use of computer software to improve the editing process. It is in the reconciliation of these two goals, (the increased use of computers for some tasks and the more intelligent use of human expertise), that the major challenge in software development lies. There will always be a role for people, but it will be modified. One positive feature of new editing software is that it can often improve the quality of the editing process and productivity at the same time.

Ways That Productivity Can Be Improved

One way to improve productivity is to break the constraints imposed by computer systems themselves. The use of mainframe systems for editing data is widespread. In some cases, however, an editor may not use the system directly. For example, error signals may be presented on paper printouts, and changes entered by data typists. Processing costs may dictate that editing jobs are run at low priority, overnight, or even less frequently. The effect of the changes made by the editor may not be immediately

known: thus, paper forms may be filed, taken from files, and re-filed several times.

The proliferation of microcomputers promises to eliminate many

of these bottlenecks, while at the same time it creates some challenges in the process. The editor will have direct access to the computer, and will be able to prioritize its use. Once the microcomputer is acquired, user fees are eliminated, thus resource-intensive programs such as interactive editing can be employed, provided the microcomputers are fast enough. Moving from a centralized environment (i. e., the mainframe) to a decentralized environment (i.e., microcomputers) will present challenges of control and consistency. In processing a large survey on two or more microcomputers, communications will be necessary. This will best be done by connecting them into a Local Area Network (LAN).

New systems may reduce or eliminate some editing tasks. For example, where data are edited in batch and error signals are presented on printouts, a manual edit of the questionnaires before the machine edit may be a practical necessity. Editing data and error messages on a printout can be a hard, unsatisfactory chore because of the volume of paper and the static and sometimes incomplete presentation of data. The purpose of the manual edit in this situation is to reduce the number of machine-generated error

signals. In an interactive environment, information can be efficiently presented and immediately processed. The penalty associated with machine-generated signals is greatly reduced. As a result, the preliminary manual edit may be eliminated. In addition, questionnaires are handled only once, further reducing filing and data entry tasks.

Productivity may be increased by reducing the need for editing after data are collected. Instruments for Computer Assisted Telephone Interviewing (CATI), Computer Assisted Personal Interviewing (CAPI), and on-site. data entry and editing programs are gaining wider use. Routing instructions are automatically followed, and other edit failures are verified at the time of the interview. There may still be many error signals from suspicious edits, however, the analyst has more confidence in the data and is more likely to let them pass.

There are two major ways that productivity can be improved in the programming of the editing instruments. First is to provide a system that will handle all, or an important class, of the agency's editing needs. In this way the applications programmer need not

worry about systems details. For example, in an interactive system, the programmer does not have to worry about how and where to flag edit failures as it is already provided. The programmer only codes the edit specification itself. In addition, the end-user has to learn only one system when editing different surveys. Second is the elimination of multiple specification and programming of variables and edits. For example, if data are

collected by CATI, and edited with another system, then essentially the same edits will be programmed twice, possibly by two sets of people. If the system integrates several functions, e.g., data entry, data editing, and computer assisted data collection, then one program may be able to handle all of these tasks. This integration would also reduce time spent on data conversion from one system to another.

Systems That Take Editing and Imputation Actions

Some edit and imputation systems take actions usually reserved for people. They choose fields to be changed and then change them. The human element is not removed, rather this expertise is incorporated into the system. One way to incorporate expertise is to use the edits themselves to define a feasible region. This is the approach outlined in a famous article by Fellegi and Holt (1976). Edits that are explicitly written are used to generate implied edits. For example, if $100 < x / y < 200$, and $3 < y / z < 4$, are explicit edits, then an implied edit obtained algebraically is $300 < x / z < 800$. Once all implied edits are generated, the set of complete edits is defined as the union of the explicit and implied edits. This complete set of edits is then used to determine a set of fields to be changed for every possible edit failure. This is called error localization. An essential aspect to this method is that changes are made to as few fields as possible, or alternatively to the least reliable set of fields which are determined by weights given to each field.

The analyst is given an opportunity to evaluate the explicit

edits. This is done through the inspection of the implied edits and extremal records (the most extreme records that can pass through the edits without causing an edit failure). In inspecting the implied edits, it may be determined if the data are being constrained in an unintended way. In inspecting extremal records, the analyst is presented with combinations of the most extreme values possible that can pass the edits. The human editor has several ways to inject expertise into this kind of a system: (1) the specification of the edits; (2) the inspection of implied edits and extremal records and then the re-specification of edits; (3) the weighting of variables according to their relative reliability.

There are some constraints in systems that allow the computer to take editing actions. Fellegi and Holt systems cannot handle certain kinds of edits, notably nonlinear and conditional edits. Also algorithms that can handle categorical data cannot handle continuous data and vice versa. Within these constraints (and others), most edits, can be handled. For surveys with continuous data, a considerable amount of human attention may still be

necessary, either before the system is applied to data or after.

Another way that computers can take editing actions is by modeling human behavior. This is the "expert system" approach. For example, if typically maize yields average 100 bushels per acre, and the value 1,000 is entered, then the most likely correction is to assume that an extra zero was typed. The computer can be programmed to substitute 100 for 1,000 directly and then to re-edit the data.

Ways That Data Quality Can Be Improved or Maintained

It is not clear that editing done after data collection can always improve the quality of data by reducing non-sampling errors. An organization may not have the time or budget to recontact many of the respondents or may refrain from recontacts in order to reduce respondent burden. Additionally, there may be cognitive

errors or systematic errors that an edit system cannot detect. Often, all that can be done is to maintain the quality of the data as they are collected. To use the maize yield example again, if the edit program detects 1,000 bushels per acre, and sets the value to 100 bushels per acre, then the edit program has only prevented the data from getting worse. Suppose the true value was really 103 bushels per acre. The edit and imputation program could not get the value closer to the truth in this case. Detecting outliers is usually not the only problem. The proper action to take after detection is the more difficult problem. One of the main reasons that Computer Assisted Data Collection is employed is that data are corrected at the time of collection.

There are a few ways that an editing system may be able to improve data quality. A system that captures raw data, keeps track of changes, and provides well conceived reports, may provide feedback on the performance of the survey. This information can be used, to improve the survey in the future. To take another agricultural example, farmers often harvest corn for silage (the whole plant is harvested, chopped into small pieces, and blown into

a silo). Production of silage is requested in tons. Farmers often do not know their silage production in tons. Instead, the farmer will give the size (diameter and height) of all silos containing silage. In the office, silo sizes are converted into tons of production. If this conversion takes place before data are entered, then there is no indication from the machine edit of the extent of this reporting problem.

Another way that editing software can improve the quality of the data is to reduce the opportunity cost of editing. The time spent on editing leaves less time for other tasks, such as persuading people to participate, checking overlap of respondents between multiple frames, and research on cognitive errors.

Ways That Quality of the Editing Process Can Be Defended or

There is a difference between data quality and the quality of the editing process itself. To refer once again to the maize yield example, a good quality process will have detected the transcription error. A poor quality process might have let it pass. Although neither process will have improved data quality, the good quality process would have prevented their deterioration from the transcription error. Editing and imputation have the potential to distort data as well as to maintain their quality. This distortion may affect the levels of estimates and the univariate and multivariate distributions. A high quality process will attempt to minimize distortions. For example, in Fellegi and Holt systems, changes to the data will be made to the fewest fields possible and in a way such that distributions are maintained.

A survey organization should be able to show that the editing process is not abusing the data. For editing after data collection, this may be done by capturing raw (unedited) data and keeping track of changes and the reasons for change. This is

called an audit trail. Given this record keeping, it will be possible to estimate the impact of editing and imputation on expansions and on distributions. It will also be possible to determine the editor effect on the estimates. In traditional batch mode editing on paper printouts, it is not unusual for two or more specialists to edit the same record. For, example, one may edit the questionnaire before data entry while another may edit the record after the machine edit. In this case, it is impossible to assign responsibility for an editing action. In an on-line mode one person handles a record until it is done. Thus all changes can be traced to a person. For editing at the time of data collection, (e.g., in CATI), it may be necessary to conduct an experiment to see if either the mode of collection, or the edits employed, will lead to changes in the data.

A high quality editing process will have other features as well. For example, the process should be repeatable, in time and in space. This means that the same data passed through the same process in two different locations, or twice in one location, will look (nearly) the same. The process will have recognizable criteria for determining when editing is done. It will detect

real errors without generating too many spurious error signals. The system should be easy to program in and have an easy user interface. It should promote the integration of survey functions such as micro- and macro-editing. Changes made by people should be on-line (interactive) and traceable. Database connections will allow for quick, and easy access to historical and sampling frame data. An editing system should be able to take actions of minor impact without human intervention. It should be able to accommodate new advances in statistical editing methodology.

Finally, quality can be promoted by providing statistically defensible methods and software modules to the user.

Acknowledgements

Other members of the Editing Software Working Group for Working Paper 18 were Tom Petkunas, Bureau of the Census, Gerry Hendershot, National Center for Health Statistics, Charles Day,

Internal Revenue Service, Marybeth Tschetter, Bureau of Labor Statistics, and Rita Hohenbrink, National Agricultural Statistics Service.

RESEARCH ON EDITING

Yahia Ahmed

Internal Revenue Service

Introduction

This paper is one of three papers presented in a session organized to present topics from the Statistical Policy Working Paper 18, "Data Editing in Federal Statistical Agencies." The Subcommittee on Data Editing in Federal Statistical Agencies was established by the Federal Committee on Statistical Methodology to document, profile and discuss data editing practices in Federal surveys. To effectively accomplish its mission, the subcommittee was I divided into four major groups: Editing Profile, Case Studies, Editing Software, and Editing Research.

The purpose of this paper is to present briefly the goals, findings and recommendations of the Editing Research Group. A more detailed description of editing research is provided in Chapter V of the Working Paper.

The goals of the Editing Research Group were to identify areas in which improvements to edit systems would prove most useful, to describe recent and current research activities designed to enhance edit capabilities, to make recommendation for future research an to develop an annotated bibliography on editing.

Areas Which Need Improvement,

The Editing Research Group used two sources of information to identify areas which need improvement. The first source was the editing profile questionnaire which was administered to managers, of 117 Federal surveys covering 14 different agencies. This questionnaire included questions about edit movements. One question asked was "For future applications, what would you like your edit system to do that it doesn't do now?" The second source was discussions with those responsible for edit tasks within a number of Federal agencies. The following areas emerged as priorities:

0 More on-line edit capabilities

0 Better ways to detect potentially erroneous responses

0 More sophisticated and extensive macro-editing

0 Evaluation of the effect of data editing.

Areas of Edit Research

Much editing research has been conducted in national statistical offices around the world. It is these organizations, which conduct huge and complicated surveys, that have the most to be gained from developing new systems and techniques. They also have the resources upon which to draw for this development.

One area of current research interest is that of "on-line edit capabilities". BLAISE, SPEER, and PEDRO discussed in the preceding paper are examples of such research activities.

A second area of active research is in the detection of potentially erroneous responses. The method most commonly used is to employ explicit edit rules. For example, edit rules may require that:

1) The ratio of two fields lie between prescribed bounds,

2) various linear-inequalities and/or equalities hold, or

3) the current response be within some range of a predicted value based on a time series or other models.

Edit rules and parameters are highly survey specific. A related area of editing research is the design of edit rules and the development of methods for obtaining sensitive parameters.

In order to make sure that all errors are flagged, often many unimportant error flags are generated. These extra flags not only take time to examine but also distract the reviewer from important problems. These extra flags are generated because of the way that the error limits are set. A related area of research focuses on developing statistical editing techniques to reduce the-number of error flags, while at the same time, ensuring that not many errors escape detection. Several research studies in which different

statistical techniques (such as clustering, exponential smoothing and Tukey's biweight) to detect potentially erroneous responses or to set error bounds are described in the working paper.

In contrast to the rule-driven method f or the detection of potentially erroneous response combinations within a record, one alternative procedure is to analyze the distribution of questionnaire response. Records which do not conform to the observed distribution are then targeted as outliers and are selected for review. Although there has been research interest in this method, no application of these multivariate methods was found.

Recommendations

The most important recommendation is that agencies recognize the value of editing research and place in high priority on devoting resources to their own research, to monitoring developments in data editing at other agencies and elsewhere and to implement improvements.

Often innovations in editing methods made by survey staff are viewed as enhancements to processing for that particular survey and little thought is given to the broader applicability of methods developed. Accordingly, survey staff do not prepare discussion of new methods for publication. We encourage survey staff to take the time to describe their work and publish them in order to share their experiences with others who may be working under similar conditions. It is often in such articles that methods which may be applicable to more than one survey are first introduced and described.

The survey on editing practices indicated that there was little analysis of the effect of editing on the estimates that were produced. Considering that the cost of editing is significant for

most surveys, this is clearly an area in which more work is required. A related issue is the need to attempt to determine when to edit and not to edit.

Clearly, all the errors are not going to be found and we should not attempt to find them all. Therefore, there is a need to design guidelines for determining what is an acceptable level of editing.

Another neglected research area in this country concerns the editing of data at the time they are keyed from mail responses. This area is usually discussed in the setting of quality control; however, it is an area that can benefit from further research from the perspective of data editing.

Annotated Bibliography

It is quite difficult to provide a complete assessment of current research activities in the area of editing because so much

of the research, progress, and innovations are described only in specific documentation. However the group was able to identify 86 references which describe research efforts over the past years. Appendix D of the working paper contains the annotated bibliography The annotations are brief and are only intended to give a very general idea of the paper's content. The appendix provides a valuable source of information on the editing literature. In addition it includes papers which describe the underlying methods, the software, proposed uses, and possible

advantages of three generalized editing software systems -- GEIS, BLAISE and SPEER.

Acknowledgements

Other members of the Editing Research Group for Working Paper 18 were Laura Bauer, Federal Reserve Board, Brian Greenberg, Bureau

of the Census, Renee Miller, Energy Information Administration, David Pierce, Federal Reserve Board, Paula Weir, Energy Information Administration.

DISCUSSION

Charles E. Caudill

National Agricultural Statistics Service

As Administrator of a Federal-State Cooperative Statistical

Agency, I am quite impressed with the information contained in OMB Statistical Policy Working Paper No. 18 on Data Editing in Federal Statistical Agencies. The working paper thoroughly, documents many existing editing practices, generalized editing software developments and provides a detailed software evaluation protocol. In addition, it covers current research activities on editing, provides an annotated bibliography and has a good executive summary including recommendations.

I believe that this report, if read and seriously considered by federal survey managers and administrators, can have a substantial effect on improving productivity. Thus, "precious" resources could be freed up to more formally address nonsampling errors, quality control, and total survey error models, measurements and structures. In my opinion, if there was ever a report that survey administrators should take seriously, this is it.

There are several more detailed comments and observations that I have about working paper number 18. The data on the costs of editing was intriguing. My observation is that there may be an

upward bias in the data, and some non-editing cost may have been included. However, even if this is the case, there obviously is still plenty of room for productivity gains in the editing process. With the proliferation of personal computer networks and data base software, there is substantial potential to improve the productivity of editing systems by being on-line and providing the editor with immediate screen feedback and re-editing of their proposed changes.

Recent computer processing technology advances also make the use of audit trails more available for more users. Inexpensive audit trails provide the capability to analyze and conduct research on the effects of editing on the estimators and also on the overall performance of the survey as well.

The detailed checklist of edit software system features in Appendix C of working paper 18 will be beneficial to both the development of new systems and maintenance and evaluation of existing systems. The annotated bibliography of articles and papers on editing presented in Appendix D will be valuable for

researchers and system developers as a substantial source of literature and information.

Working paper 18 certainly demonstrated that current data editing practices are labor intensive. Many remain mainframe and batch oriented, with multiple passes of the data. Also, I think that there may be a tendency to stay with existing systems too long.

My final comments are on total quality management of surveys. As an Administrator, one of my major concerns is with the quality of the final products and reports that the Agency delivers to the public. Thus, if the editing process can be made more efficient, without degrading accuracy, then that adds to the potential of using the saved resources on other important areas of the survey process. Total quality management techniques applied to surveys

are useful tools in efficiently identifying the most important potential sources of survey error.

DISCUSSION,

Richard Bolstein

George Mason University

The serious impact that erroneous survey data can have on results, the fact that the number of errors tend to increase with the size and complexity of the survey, and the relatively large proportion of survey costs currently required to edit and correct data, make the need for new and improved methods of data editing imperative. To this end, the authors have done a laudable job in researching methods currently used, presenting several case studies, testing and discussing the advantages and disadvantages of some current and developing editing software, and providing a synopsis of current research.

A working definition of editing was clearly necessary in this

study, since, among other things, in order to estimate costs of editing, a fairly rigorous definition of the scope of editing was required. The working definition used by the authors, namely, "procedure(s) designed and used for detecting erroneous and/or questionable survey data with the goal of correcting as much of the erroneous data as possible, usually prior to data imputation and summary procedures" is quite suitable for this purpose. We should keep in mind, however, that while it feels comfortable to clean up erroneous data prior to imputation for missing data, in practice the two are often intertwined.

The paper states that the cost of editing was available for 40% of the 117 surveys in the sample, and cost estimates were possible for an additional 40%. It was reported that between 75% and 80% of these surveys had editing costs of at least 20% of total costs. It is not too meaningful to compare the relative costs of editing across all types of surveys however, since one would naturally expect that these costs would be higher in less expensive surveys (such as mail or administrative records) than in expensive surveys (such as personal interview, surveys of institutions), as found by the authors. Thus, it would be more informative if the

relative cost figures cited above were reported by survey type. Another factor that can account for a large percentage of editing costs is the presence of a relatively large number of questions requiring open-ended responses and subsequent coding of the responses. But although the distribution of the relative cost of editing may vary considerably, there is no doubt that editing is costly and methods to reduce this cost and improve data quality are much needed.

Finally, no discussion of the costs of editing is complete without determining what percentage is due to bad data that should not have occurred but for inadequate interviewer training, poor supervision and quality control of interviewers, and simple common

sense errors. For these are errors which should not have occurred and should be deducted from the cost of editing in the estimates of the surveys above, since they are likely to have varied

considerably.

Although elimination of such unnecessary errors was not part of the project of the three authors, it seems appropriate in a discussion of improving data editing procedures to mention ways in which the need for editing can be reduced. To illustrate an example of a common sense error that should be eliminated, in a certain survey, the sponsor of which I will not name, fishermen are interviewed and their catch is weighed and measured. The interviewer is supposed to record weight in kilograms, but the scale used shows weight in both pounds and kilograms. As expected, frequent errors occur. The obvious solution is to use a scale that only shows kilograms, but when I suggested this to the survey firm, the response was "no one makes such a scale". When I then suggested taping over the side of the scale showing pounds, the reply was "but the fishermen want to know what their fish weigh in English". Finally, I suggested taping over the kilogram side of the scale, have the interviewer record the weight in pounds, and have the data entry program convert it to kilograms. The response to this suggestion I am sure you have all heard before: "well,

that's the way we're used to doing it". There are numerous other examples of course (for example, in some surveys interviewers are required to record the hour in military time).

The most promising methods to reduce editing costs and improve data quality (after elimination of the unnecessary errors) are found in interactive data entry software and in general editing software systems. These methods seem appropriate for large, complex surveys, or surveys which are repeated. For small one-time surveys the cost of purchasing, learning, and programming the software will most likely outweigh the savings, as this is even true with CATI. But this is generally not the case with surveys gathering Federal Data. The three generalized editing software systems studied in detail by Mark Pierzchala seem very promising, especially BLAISE because of its generality and ability to handle both categorical and continuous data. GEIS and SPEER are specific to economic type surveys.

To what extent can graphics or other theoretical tools be used in editing systems? The STAR WARS software described uses graphics

to compare edited values with the originals, but not to detect outliers. The parallel coordinate system for graphic displays of high-dimensional data [see Miller and Wegman (1989), Wegman (1990)] may be used to detect outliers. Yahia Ahmed noted that analysis of the multivariate distribution of questionnaire responses to flag records that don't conform to the distribution as outliers has been infrequently used, no doubt due to its complexity. I believe that graphical methods for detecting outliers will meet with more acceptance than the multivariate analysis approach has but it would

not be cheap (time-wise) and probably would be best used as a final check rather than at the front-end of the editing task.

Finally, I have two recommendations. In view of the increasing abundance of software we will see in the future, we should construct a standard set of test data sets for evaluating present and future software editing systems. Secondly, a one or two-day demonstration seminar of some of these systems would be well received.

References

Miller, J.J. and Wegman, E.J. (1989), "Construction of line densities for parallel coordinate plots", Technical Report No. 53, Center for Computational Statistics, George Mason University.

Wegman, E.J. (1990), "Hyperdimensional data analysis using parallel coordinates", Journal of the American Statistical Association, to appear.

Session 6

COMPUTER ASSISTED STATISTICAL SURVEYS

OVERVIEW OF COMPUTER ASSISTED SURVEY INFORMATION COLLECTION

Richard L. Clayton

U. S. Bureau of Labor Statistics

This section provides a summary of Working Paper 19 on Computer Assisted Survey Information Collection (CASIC). For additional information, we encourage you to see this document.

The power of rapid calculating has been applied to virtually every phase of the survey process, including sample design and selection, and estimation. The most important implication of these applications is that survey practitioners are allowed to consider a growing range of techniques which were not affordable prior to

the availability of inexpensive and fast calculating capability.

The field of computer assisted collection applications may be the area of greatest and most rapid change in survey methods. This field includes the rapidly expanding variety of applications based on the availability of powerful and inexpensive computers. Most familiar of the new techniques are CATI and CAPI. However, a variety of other collection methods are being developed across the Federal government's statistical agencies, including Touchtone Data Entry, Prepared Data Entry and more recently, voice Recognition Entry.

High quality published data begins with collecting high quality data from our respondents. Much of survey processing addresses, and compensates for, weaknesses in the quality of the collected data and the data we do not collect. Those methods which capture data quickly and accurately should be developed which allow respondents to answer our questions accurately and quickly. With this in mind, we provided the results of research and development activities using new technological features throughout the Federal

government seeking new data collection methods, and in modifying the old, to improve the quality of data collection.

For the purposes of this report, we defined computer assisted survey information collection methods as those using computers as a major feature in the collection of data from respondents, and in transmitting of data to other sites for post-collection processing.

Goal: The overall goal of Working Paper 19 was to provide information on new data collection methods to challenge Federal survey managers to reconsider their operations in light of recent changes in survey methods available, or made attainable through changing technology to reassess their methods of accomplishing the common goal of providing the critical information to the public which is accurate, timely and relevant. We hope that by sharing information and experiences, that others may gain and forward the overall effectiveness of governmental activities.

Objectives: The primary objective is to describe emerging

methods of interactive electronic data collection, the potential benefits, and current examples of its use in Federal surveys. In describing current uses and tests, a secondary objective is to pose questions about the implications of use of computer assisted methods and try to suggest some answers. These questions involve such factors as quality, costs, and respondent reaction to. computerized surveys.

Scope: The survey operations included in this report includes all of the activities and tasks from the transmittal of the questionnaire, conduct of the interview, data entry, editing and followup for nonresponse or edit reconciliation.

The last major survey operation to benefit from automation is data collection. Computers were first applied to collection using mainframes to control certain aspects of telephone collection, and Computer Assisted Telephone Interviewing (CATI) was born. The first applications of CATI stimulated new research worldwide evaluating the impact on of CATI on the survey error profile and costs. CATI is now used to assist interviewers in all collection

activities, including scheduling calls, controlling detailed interview branching, editing and reconciliation, providing much greater control over the collection process and reducing many sources of error. At the same time, a tremendous amount of information it captured by the computer providing additional insight into the data collection process.

The ongoing advances in computer technology, and particularly the advent of microcomputers, continue to offer additional opportunities for improving the quality of published data. The first portable computers were quickly pressed into service to duplicate the advantages of CAT! in a personal visit environment. Thus, Computer Assisted Personal Interviewing (CAPI) was launched from the work in CATI.

While CATI and CAPI represent advances for surveys requiring interviewers, microcomputers are now finding important roles in self-administered questionnaires, where interviewers are not needed.

Prepared Data Entry (PDE), developed by the Energy Information Administration, allows respondents which have a compatible microcomputer or terminal to access and complete the questionnaire directly on their screen.

Touchtone Data Entry (TDE), developed at the Bureau of Labor Statistics, allows respondents to call a toll-free telephone number. Questions posed by a computer are answered using the keypad of their touchtone telephone. The machine repeats the answers for verification with the respondent which are stored in a database. TDE systems are now commonplace for bank transfers, and

telephone call routing, as examples. We have just applied existing technology to the data collection process.

As an extension of this approach, techniques have been developed more recently allowing respondents to answer the questions by speaking directly into the telephone. The incoming

sounds are matched to known patterns recognizing the digits and the words "yes" and "no". Voice Recognition Entry (VRE), as this is known, is not the distant future. The Bureau of Labor Statistics is currently conducting live tests where this method is being warmly received by respondents as natural and convenient.

Both TDE and VRE offer inexpensive data collection where the respondents initiate the calls, enter and verify the data. Refinements to procedures will now focus on minimizing nonresponse prompting activities.

Respondent Burden: For many respondents, the use of automated methods can actually reduce the collection burden placed on them. For example, use of Prepared Data Entry, where respondents interact with computer screens, provides a single set of step-by-step procedures with on-line editing to prevent inconsistent or incorrect reporting, thus reducing the need for expensive and troublesome recontacts. Also, these methods have, in some cases, substantially reduced the time taken to provide complex data for large establishments. Similar methods may be applied to other surveys covering large establishments where the one-time costs of

data conversion to a standard format would be cost-effective, especially in repeated surveys.

Ouality: Automated collection allows for improved control yielding reduced error from several sources including errors caused by the respondent, the interviewer, and post collection processes such as key entry error. The instant status capabilities of CATI, for example, provide stronger intervention features for nonresponse prompting, reducing nonresponse error.

In deciding which collection method to use, quality can become a relative concept that is affected by a tradeoff between cost and benefit. The choice of a data collection method is usually based on a combination of performance and cost factors determining affordable quality. For traditional collection methods, these factors and the decision-making process are fairly well known. Now, these new methods discussed in Working Paper 19 expand the array of potential collection tools and challenge the survey designer to reevaluate old cost/performance assumptions.

Costs: The data collection process is composed of a few major activities, including transmitting and receiving the questionnaire, data entry, editing and nonresponse prompting. The labor and nonlabor costs will vary depending on the method used. For example, under mail collection virtually each action is conducted

manually and postage is the dominant nonlabor cost. By contrast, CATI operations can minimize postage costs reduces many of the expensive mail handling operations. However, CATI adds new costs in the form of telephone line charges and computers (including Systems design and ongoing maintenance). Self-response methods, such as TDE, VRE and PDE collection, reduce postage, the manual mail operations and the labor involved in CATI interview activities, but may still require edit reconciliation and nonresponse followup.

Thus, the factors of production, and the composition of each those inputs vary greatly among the existing and newer techniques.

Many factors can change in a short period. Only a few years ago, automation costs were driven by the scarcity of mainframe hardware capacity. Now, the costs of automation are driven by the labor involved in developing specialized systems dominates automation costs. Portable and desktop microcomputers were not widely available at the beginning on this decade. Now, microcomputers are widely available, very inexpensive and extremely powerful.

Old assumptions about costs need to be reevaluated. Labor and postage costs have risen steadily in recent years, while capital costs, such as microcomputers and telephone services have been declining.

The decision on which collection mode to use, or which combination, will depend on the particular survey application and the existing cost structure. However, it is important to view such investments over the long-term as the relative costs of each of the inputs do not remain constant over time. Survey managers should periodically review old assumptions in light of new technology and project operating costs over the reasonable foreseeable future in

deciding not to investigate new methods.

Users: Automated data collection includes three major groups of people: the respondents, the interviewers and the designers and developers of the system and procedures for collection. This report covers the essential factors involved in successfully including the requirements of each group.

Respondents: The respondent must be considered the primary user of any survey vehicle, whether automated or not, and all aspects of the response environment must be developed with the respondent in mind. The cooperation of the respondent is the single most critical factor in survey operations. Respondents must be treated with the greatest care. We must consider our respondents as a Customer, after all, if our survey vehicle doesn't "sell", if the questionnaire is not successful in getting an accurate response, we will have no input for the rest of our production process.

Even one-time surveys must strive to leave the respondent with the feeling of contribution and importance, and most of all, a willingness to participate in other surveys in the future if called on. Thus, our primary job is to develop techniques which allow the respondent to complete the survey completely and accurately and with a minimum level of burden.

The use of these collection methods, while bringing improvements in the quality of collected data, has entailed other challenges. These automated collection methods are made possible through the close interaction of subject matter experts, statisticians, and computer scientists. To effectively use these methods, each of these groups learned the basic tenants of the others. This close relationship will only continue to grow, with advances in each field aiding advances in the others.

Interviewers: The second most important user is the interviewer. The systems provided to assist in the interview process must be easy to use, must work infallibly and must actually

provide improvements in his or her work environment. Interviewers must feel as they are the most valuable feature in the interview, that the machine is merely a tool to expedite and simplify their work. This is not always an easy task.

Survey Practitioners: We are the third major group of users. The decisions made early in the development process will carry over into the ongoing use and maintenance of the system.

Systems designers face difficult choices, such as building customized systems from scratch versus linking standardized "off the shelf" routines or commercial, packages. The inevitable limitations would have to be traded off against reduced maintenance and lower start up costs.

Automated collection methods can also improve data quality. All of the methods discussed could be designed to include on-line editing to prevent impossible and inconsistent entries. Some of these methods, such as TDE and VR, improve data quality by verifying recorded data with the respondent.

These are potential improvements. The final impact of quality lies in the up front planning and execution. This place responsibility for clearly defining and controlling the collection environment directly with the survey designer.

Future: The future application of these techniques is limited only by our creativity and initiative of program managers and planners. The "case studies" serve to illustrate the options available, and will surely raise many more questions for further investigation.

We hope that the discussion of technological advances generates discussion and stimulates creative, new applications to the whole range of governmental information collection activities.

In addition to the methods described here there are other advances in, technology which hold potential for vastly changing data collection. Integrated Services Digital Network (ISDN) is a powerful network system which will provide simultaneous transmission of sound, video and data. The result could be a change in the way some surveys are conducted offering all of the benefits of personal interviewing with the lower costs of telephone interviewing.

You have heard a several different collection methods described and discussed which are currently available. And you can see that the pace of change will accelerate and match changes in technology. So what does the future hold?

You have to ask yourself how your survey operations will be conducted in 5 or perhaps 10 years. In doing so, ask yourself how things were done 5 or 10 years ago. What sorts of things have happened and what were their implications?

A COMPARISON BETWEEN CATI AND CAPI

Martin Baum

National Center for Health Statistics

Introduction

I will describe for you some of the critical factors one must consider when deciding whether to conduct a survey by either CATI or CAPI. I also will try to indicate the similarities and differences between these to methods of survey data collection automation.

Definition

Let me first define each of the methods. Computer Assisted Telephone Interviewing (CATI) is a computer assisted survey process which uses the telephone for voice communications between the interviewer and the respondent. Computer Assisted Personal Interviewing (CAPI) is a personal interview usually conducted at the home or business of the respondent using a portable computer.

Rationale

The rationale for the development and for your use of these methods are based primarily on reasons of improved data quality and improved timeliness of data release. Cost is a factor, but in our experience, it has been a break-even situation; the cost of automating has equaled the savings. This result has been due primarily to the high cost of software development.

Factors

The following are critical factors that must be considered in addition to those of improved data quality and timeliness, and cost when deciding whether to use CATI or CAPI for your survey data collection. I will discuss each of these factors in some detail.

Hardware CATI

Initially CATI was developed as a mainframe application but as computer technology changed, CATI moved to the mini computer and then to a networked micro computer application. The investment in hardware has steadily decreased without any lost of capability. Telephone technology, which impacts telephone availability is important to the CATI application - no phone no respondent.

Hardware CAPI

The most important computer hardware criteria for a CAPI application are generally quite different from those that would be

critical to most other applications. The major reason is the role that environmental conditions play in the selection of CAPI hardware. The fact that CAPI is a personal interview situation, usually taking place in or at the home of the respondent, dictates a number of possible circumstances under which the interview will be conducted.

For example, screen visibility becomes a paramount criterion because of the environmental conditions. Interviews will take place under all types of lighting conditions; outside in bright sunlight, twilight, and normal light, and inside under lamp light, fluoresce light, and bear bulb.

Weight is especially critical because of the variety of environmental conditions. Interviewers may be conducting the survey in an urban setting where the computer will be carried up and down the stairs of apartment houses; or in a suburban setting where the computer is carried many blocks; or in a rural setting where the computer is carried long distances from car to house. In any of these conditions, the computer is moved in and out of a car many times. This situation is further compounded by the fact that

the interviewer must also carry considerable paper e.g. back-up paper questionnaires in case the computer fails, letters of explanation, introduction, and thank you. Carrying all of this weight in and out of cars and up and down steps all day is no easy job, particularly if the computer and back up battery weighs 10 plus lbs. and the paper weighs an additional 5 lbs. or more.

For a household type survey, the interviewers are generally reluctant to ask for the respondent's permission to use power for the computer because of fear of possibly losing the interview. Also, surveys frequently are conducted outside of the house where no power is available. Many of our surveys can last as long as 2-4 hours. Consequently, battery life it critical.

Environmental conditions often impact the ergonomics of the hardware. Consider a survey interview conducted where the computer must be placed on the interviewer's lap. This situation would be quite difficult if the computer were either top heavy when open or the interviewer was small and the computer's depth long. Balancing would be a problem. Also consider the door step

interview with a 10 lb. clam shell design computer.

Software

Now let's discuss the most costly factor in the CATI/CAPI decision - software. There are four components to the CATI/CAPI

software: Questionnaire, Case Management, Output Reporting, and Authoring System.

The questionnaire component refers to the software that places each question in the survey on the computer screen in the proper sequence with the appropriate information (i.e. prompts) and allows the entry of an answer or answers to the question with edits on those answers such as; range, specific values, consistency with another question's answer. This software should also contain on screen help and if necessary, rostering.

The case management component is the software that allows the interviewer to keep track of the status of the survey interview; that is, is the interview complete?; if the interview is not complete, what has been completed and what is the next question to be asked?; is the interview a partial interview or is the interview to be completed later?; what sections of the survey are mandatory?; and in some instances, interviewer assignments. In the case of CATI, case management software also would provide the sample selection and dialing of the phone number.

The output reporting component is often either overlooked or given minimal consideration. This is a big mistake. Collection of the data is not very useful if the data cannot be easily accessed for analysis. Output reports can be categorized as either survey questionnaire statistics or management statistics. The level of detail and complexity can vary significantly. Survey questionnaire reporting can be as little as the ability to place the data into specific analysis software file format e.g. SAS or can include actual analyses.

Management statistics can be extremely useful for the conduct of the survey data collection. For example, data can be automatically collected on the time to complete a section of the questionnaire by interviewer. This information could provide insights for training and/or question rewrite.

The authoring system allows a non-computer programmer e.g. a survey questionnaire designer, to create the questionnaire while simultaneously and automatically generating the questionnaire software component. It has been our experience that this is the most difficult component to develop. Although there are a number of such systems that are available, none of these systems has met all of our requirements for the type of complex survey we conduct e.g. NHIS. The authoring system should be extremely user-friendly and be able to handle a large number of question types.

Data Transmission

In the case of CATI, the data is automatically transmitted to a central point for either uploading to larger computer or further processing e.g. analysis.

In the case of CAPI, the data collection is dispersed generally over a wide geographic area. The two primary methods for data transmission have been mailed floppy disk or telecommunications. For data that is needed in one day or later, floppy disk has been adequate. Telecommunications, however, adds a new dimension - Two way communications. Not only can data be transmitted to a central point, but instructions for the interviewers, for example, could be transmitted from the central point to the field. The major problem with the telecommunications method has been consistent quality of the communication lines. Cost can also be a barrier.

Interviewer Training

The level and amount of training needed depends, to large extent, on the level of user-friendliness of the software. Our experience has shown that the type of training is different for either a CATI or CAPI conducted survey than for the pencil and paper conducted survey. In the paper and pencil conducted survey, training is focussed on almost entirely on the content of the questionnaire, management of the questionnaire, and the proper question sequencing. It would not be unusual to have an accompanying instruction manual 3-4 inches thick that would have to learned by each interviewer. Whereas, in the CATI or CAPI conducted survey, training included both questionnaire content and the care and use of the computer. The major focus, being the computer not the content because the computer software can handle most of the problems the interviewer needs to worried about in the pencil and paper conducted survey, such as; probes, question sequencing, completeness.

There is one major difference between CATI and CAPI that

impacts on the training: the level of interviewer anxiety. CATI is conducted at a central location where supervision and help are readily available. CAPI, on the other hand, is conducted in the field where no supervision or help is readily available. Therefore, CAPI training must try to provide the interviewers with sufficient confidence in the software and hardware to cope with this lack of help. One method that has proven effective is to emphasize hands-on practice. Interviewers are encouraged to take home their computer and practice interviews with anyone they can get prior to going into the field. In addition, interviewers are given their computer prior to the training so they can have some familiarity with them. CAPI interviewers must be able to cope with

problem occurrences. Consequently, training must concentrate on such situations.

Future Technology

Impending technological advances can have a profound impact on these automation methods; particularly CAPI. Changes in hardware such as; an "etch-a-sketch" microcomputer and an inexpensive long-life, light-weight battery would open new possibilities for the CAPI conducted survey. Use of a light-weight computer, under 5 lbs,no key board, with light pen hand-written entry would allow door step surveys as well as reduce training efforts. The "etch-a-sketch" computer has been introduced by one vendor and several other are about to announce. The long-life light-weight inexpensive battery, although not currently announced or available, when available will produce much faster and larger light-weight computers. Thus allowing larger and more complex surveys to be automated.

The development of an generalized authoring system software would open up the use of CATI and CAPI to the quick-turn-around type survey. Survey questionnaires could be designed and implemented quickly and easily. Staff productivity would also increase significantly because computer programming efforts to automate each survey questionnaire would be reduced to a minimum.

The survey designer, in effect, would be programming the survey while designing the questionnaire.

COMPUTER ASSISTED SELF INTERVIEWING

Ralph Gillmann

Energy Information Administration

The phrase "computer assisted self interviewing" (CASI) covers all survey methods in which respondents access computers. These methods include "computerized self administered

questionnaires" (CSAQ) and "prepared data entry" (PDE) where the respondent fills out a computerized version of the survey instrument. Also included are methods where the respondent uses a telephone to access a computer: "touch tone data entry" (TDE) and "voice recognition data entry" (VRE).

Let's step back for a moment and look at different ways that computers can be used in interviews:

The top line represents direct interaction between an

interviewer and a respondent. The left line represents the interviewer accessing a computer such as in CATI and CAPI which were previously discussed. CASI methods are illustrated by the lower right triangle. The diagonal represents respondents accessing an agency computer as in TDE and VRE. The right line represents respondents accessing their own computers as in PDE. With the personal computer (PC) becoming ubiquitous, at least in establishments, respondents usually have access to a computer.

The bottom represents computer to computer interaction for data transmission. The missing diagonal would represent the activities of hackers and spies.

Next, let's compare manual and computer assisted methods:

Some methods are part manual and part computer assisted. For instance, CATI and CAPI combine a personal interview with an electronic survey instrument. One survey which uses all of the computer assisted methods is the Petroleum Electronic Data Reporting option (PEDRO) in use at the Energy Information Administration. In general, the manual methods are slower and more prone to processing errors. Labor and postage costs are also rising faster than the operational expenses of computer assisted methods.

For transmission of the data to the collecting agency, paper copies can be sent via facsimile machines (fax). This method is

faster than the mail but doesn't eliminate the need to key in the data. If the data are in electronic form, a diskette with the data can be mailed in. This is useful if security and authenticity are a particular concern. Transmission time may be saved by sending the data over the telephone network or using "electronic mail" over a computer network. (Note that it's becoming harder to tell telephone and computer networks apart.)

The use of an electronic mail service is feasible now and likely to be more important in the future. This method allows a third party to handle the support for telephone lines, security, and temporary storage. Respondents only need to have a terminal which operates over ordinary telephone lines if the survey instrument resides with the electronic mail service in the form of an electronic questionnaire. Security can be provided by passwords and data encryption. The survey agency can retrieve the data at its convenience.

Finally, CASI offers several quality improvements:

Increased timeliness of the data (especially important in monthly and weekly surveys)

Fewer follow-up calls to respondents (because many, if not all, data edits can be done immediately)

Reduced respondent burden (fewer persons are needed to fill out an electronic form)

Lower costs (at least in cases where labor and postage make up a large part of the costs)

COMPUTER ASSISTED SELF INTERVIEWING:

RIGS AND PEDRO, TWO EXAMPLES

Ann M. Ducca

Energy Information Administration

I am going to talk about two systems that the Energy Information Administration has for reporting data using personal computers (PC's). One system is a mail submission of a PC diskette, and the other uses telecommunications between the respondent's PC and our mainframe computer.

The first example is the Reserves Information Gathering System, known as RIGS. It is a system for reporting data on domestic oil and natural gas reserves on PC diskettes. The data are collected by the EIA in its annual survey of oil and natural gas well operators. Reporting to this survey is mandatory.

Briefly, this survey is a stratified sample survey with the stratification being done according to the amount of production of oil and natural gas. Respondents in the first strata, representing the largest amounts of production and having the most data to report, are eligible to report using RIGS. They will also continue to have the option of reporting on paper forms. The EIA cannot require an electronic form of submission. RIGS first became operational for the reporting of 1988 data. We anticipate that 25-30 percent of the 1989 reserves information will be reported using the RIGS system.

The EIA sends PC diskettes containing the RIGS processing software by mail to respondents. A user's guide is also provided. The respondents install RIGS onto their PC's and use it to enter data.

The basic hardware requirement is an IBM compatible PC with at least 360K of random access memory, and two floppy disk drives or one floppy and one hard disk drive. A printer should also be

attached to the system so that a hard copy can be printed. Version 2.0 or higher of MS DOS is also required. The IBM PC compatible computer was chosen because of its wide availability.

The software for RIGS was originally written in dBASE III, a PC database management system. dBASE III programs can only be executed using the dBASE III software, that is, stand-alone programs cannot be created. Since the EIA did not want to purchase and provide the dBASE III software for every respondent, Clipper, a linkage compiler, was used to compile dBASE III into object code to make it a portable system. The licensing agreement with Clipper permits run-time programs created by it to be operated outside the agency. Thus, the respondents are provided with an executable load module, not programs. Licensing agreements must be carefully

reviewed before planning to use software products outside an agency.

An advantage of a load module is that respondents cannot directly or inadvertently change the programs. Also, there is no cost to the respondents since the RIGS software was developed by the government.

Using the RIGS software, the respondents enter data directly on their PC. The data entry screens for RIGS are formatted like the data collection form. There may be some benefits to exploring other formats which take advantage of options available to automated collection, such as question sequencing.

There is also the option of sending an ASCII file to the RIGS system so that data already available in an automated form at the respondent site can be submitted without re-keying. The RIGS User's Guide gives the instructions and record layout requirements for downloading ASCII files.

Respondents are required to submit to us by mail a diskette containing a copy of the cover page and the data. They must also return a paper copy of the cover page with the signature of the

certifying official.

Because the survey is an annual one, it was decided that telecommunications with the EIA mainframe computer was not needed, and that the mail submission would be sufficient. Since the data in the RIGS system are proprietary, it was also decided that respondents would not be provided with their previous year's data because of the risk of sending confidential data to the wrong respondent.

Preliminary edits such as range checks are performed as the data are entered into the RIGS system. If the system detects an incorrect entry, the bell sounds and a message appears across the top of the data entry screen. The message will prompt the user for a response. Help screens are available to assist the user, and help is also available by telephone on a toll-free number. For data that have been downloaded into RIGS, an edit report is produced afterwards. A respondent may then use the RIGS edit function to correct the errors.

Final edits, such at comparisons with previous year's reports,

are made after the data are returned to the EIA. These edits are performed on our mainframe system. When questionable data are identified, a quality control analyst contacts the respondent by telephone and changes are made by the EIA.

Respondents also have the option to make notes in a footnote. These notes may be helpful in explaining data that appear to be questionable.

The second example is the Petroleum Electronic Data Reporting Option (PEDRO). It gathers monthly data for petroleum supplies from petroleum companies. The respondents eligible to use PEDRO participate in 7 monthly surveys. They include refineries, storage facilities, pipelines, importers, and extraction facilities. Reporting to these surveys is also mandatory. But again, the EIA cannot require an electronic form of submission.

The participation in PEDRO varies among the 7 surveys. The

market share represented by reports to PEDRO ranges from 25 to 90 percent of the total volume for a survey.

The main difference between the PEDRO and RIGS systems is that PEDRO uses telecommunications to transmit data directly to the EIA mainframe computer. PEDRO users need an IBM compatible PC with a hard disk and a floppy drive, and a modem. As with the RIGS system, respondents are provided with an executable load module at no cost. PEDRO also requires the Arbiter communications software which is licensed only for use with the EIA. Arbiter was selected because it satisfied our security needs. The EIA supplies the respondents with Arbiter.

The basic methods of entering data to PEDRO are the same as those with RIGS -- keying on the PC or sending an ASCII file to the PEDRO system. However, data submission in PEDRO is done by telecommunications directly to our mainframe, rather than by mailing diskettes. Since these are proprietary data, PEDRO submissions are encrypted. The transmissions are time-stamped to replicate a postmark. The respondents must use passwords to

transmit data, and the password, rather than a written signature, serves as the certification of the validity of the data.

All edits in the PEDRO system appear on the respondent's PC. Since there is a direct link to our mainframe, all data needed for editing comparisons, for example prior month's data, are available on-line. Preliminary edits are performed before respondents transmit. any data. Final edits are performed after the link to the EIA mainframe and transmitted back to the user.

The EIA is very pleased with the RIGS and PEDRO reporting systems. We believe that we are getting data faster and more accurately from these systems, and are encouraged by the increase in interest in using them.

IMAGE
NOT AVAILABLE

Cathy Mazur

National Agricultural Statistics Service

In this session, I will first mention several factors to consider when deciding on a mode of data collection. Then I will spend a few minutes comparing the modes of data collection that have been discussed.

The primary factors in choosing a method of data collection for a given survey are (as previously ;mentioned) the available time frame, the desired quality, and the cost of resources. It is unusual to have all three of these in abundance. Therefore, tradeoffs must be considered.

Several other factors to consider which relate to survey design and operation are whether the survey is mandatory or voluntary, whether a onetime or ongoing survey is to be implemented, whether households or businesses are sampled, whether

the data will be collected; in a centralized or decentralized manner, whether networking of computers will be done, the sample size, and the complexity of the questionnaire.

The remaining factors to consider in automated data collection refer to the characteristics of the technology. First is the speed of the hardware and data transmission over the phone lines. Next is the size of the computer's memory, and the system's weight (as in CAPI). Portability is a concern to data collection when different hardware and/or software is to be used (as in Prepared Data Entry (PDE). The type of display is important in some modes (as in CAPI). The mode of data entry can be through the keyboard, a pushbotton phone, or using one's voice. Data verification depends on the desire for quality, the complexity of the data, and other factors. The database generation is also an important step (as was discussed by Martin Baum). It refers to integrating the data with other survey processes (label generating, data summaries). Hardware is selected based on cost, the amount of time available, the data quality desired, and the background of the staff that will operate the machines. Lastly, training is important in any survey, the amount of which depends on the

technology chosen.

The priorities that are given to these factors and the relationships between them, help to decide which technology to use. All combine data collection with data entry, and most add editing at the time of data collection. This reduces the time component and increases the quality component. Also, mixed modes of data collection are possible in a survey.

First, (as a means of comparison), a mail or manual survey would require a fairly long time to send out personal enumerators or to send and receive questionnaires through the mail. The amount of editing is very limited as data entry and editing is done after all the data is collected and the interview is completed. The cost is fairly high if personal interviews are done, and nonresponse may also be high if questionnaires are mailed out.

CATI is used because it collects data quickly and accurately. The cost component (which is fairly high), comes from the hardware, software, training, and support factors (such as phone charges). One cost component which is eliminated is the travel expense. One suggestion is that CATI improves the cost benefit. The respondent, however, must have a phone. Other benefits are that it is useful in complex survey environments, can provide information on call scheduling successes/failures, and can be used for non-response follow up.

CAPI also has fairly high costs, but it provides accurate data with a tendency for higher response rates (which may be a problem in CATI), and saves on the separate keyentry time. The largest cost component is due to travel (with some in hardware and software support costs). The weight, battery life, and screen visibility are important issues to CAPI.

As to computer-assisted interviewing, 3 data collection modes are discussed -- Prepared Data Entry (PDE), Touchtone Data Entry

(TDE) and Voice Recognition Entry (VRE). PDE provides faster and more accurate data, for an average cost. Costs are incurred in software development and support areas. This mode requires the availability of a PC (usually by establishments), and two issues are data security and data integration (as different PC's are used).

TDE allows respondents to call and answer questions posed by a computer using the keypad of their touchtone telephone. VRE also allows respondents to call and answer questions posed by a computer, but the respondent answers by speaking directly into the telephone, and a computer system translates the incoming sounds into text. TDE and VRE offer low cost alternatives in a short data collection time, but editing is more limited. In both, surveys tend to be shorter and simpler, non-response prompts are used, and respondent acceptance is a concern. TDE requires access to a touchtone phone and service, where VRE can use any phone. The Bureau of Labor Statistics collects data monthly for the Current Employment Statistics Program using mail, CATI, TDE, and VRE. The VRE system recognizes any American English-speaking person with

continuous speech of the numbers 0-9, yes, and no.

These are not simple issues, and there are no clear cut answers. The definitions and importance of the factors must be

agreed upon. This comparison only represents the current state of technology, much will change with future development.

Lastly, I hope this session has made you more aware of the possibilities, the issues, and what to consider when choosing a data collection method.

DISCUSSION

Robert N. Tinari

U. S. Bureau of the Census

I want to begin my remarks today by noting that this paper is a very thorough treatment of the issues surrounding automated survey collection methodologies.

I am impressed with the organization of the paper and the thoroughness of discussion of the many considerations that go into selecting, designing, and implementing these types of data collection systems. The subcommittee is to be commended for the excellent job they have done in bringing together in one document a tremendous amount of information that I think will be extremely

useful to those considering alternative data collection methodologies.

Based oh my experience as a program manager responsible for the initial development and implementation of CATI on the National Crime Survey, there are several issues raised in the paper that I believe need more emphasis.

The first issue I want to discuss has to do with organization and its affect on CATI/CAPI development and implementation.

In its conclusion, the committee notes that increased reliance on software development has important implications for hiring and training skilled survey designers. It also states that previously distinct boundaries between occupational groups will-continuously blur and disappear and survey design will likely be increasingly accomplished through teams of skilled workers from different occupations.

Based upon my experience, I believe that this is an accurate

assessment. Obtaining the maximum benefit from the these data collection methodologies requires that a fully integrated system be developed and this, in turn, requires the concerted effort and collaboration of programmers, survey design experts, statisticians, field staff, program managers, and survey sponsors.

However, the level of cooperation and communication necessary to successfully design and implement CATI/CAPI may be very difficult to achieve in a large, hierarchical organization. Staffs tend to be highly specialized and not experienced in projects requiring a multi-disciplined approach.

From my own experience working on one of the first CATI applications at the Census Bureau, we had a very difficult time organizing the right team with the right experience necessary to get the project underway and in keeping the lines of communication

open among the various divisions involved to implement it

successfully.

We learned a lot from that process and have come a long way. A recent example is a cooperative effort between the Economic Area and the Demographic Area in successfully developing and implementing a CATI system for the Survey of Manufacturing Technology. The Industry Division was responsible for conducting the survey and wanted to use CATI for nonresponse followup of manufacturing plants. The division lacked the experience to develop the questionnaire on CATI. Demographic Surveys Division offered to help with the authoring, Industry assisted with testing and Field Division worked on interviewer training and data collection. The survey was carried out on time, within budget, and with high quality. This is a good example of what can be accomplished by individuals working together from the various divisions and sharing their expertise to get the job done.

Poor organization and control can have a very serious impact on the cost and time of development and the quality of the final product. I believe that what is needed to successfully design and implement automated data collection methodologies is:

0 commitment and full support from upper-level management.

0 a full-time, dedicated staff - no part-time work along with other projects.

open lines of communication with clear assignment of responsibility/accountability.

0 designate a project coordinator/facilitator

0 breaking down of traditional barriers between survey statisticians, mathematicians, survey designers, programmers, and field staff in order to work effectively.

0 ongoing commitment and organizational change to adapt to needs of the new data collection methodology. Especially important if you are using mixed mode such as personal visit (paper) and centralized telephone (CATI).

0 reduced layers of bureaucracy.

0 empowerment of the team to get the job done.

We must think of new ways of organizing ourselves to be more flexible and effective in designing and implementing new technologies. In addition, there must be more sharing of

information among the various statistical agencies on approaches and experiences in the area of organization.

The second issue has to do with interviewer acceptance of new technologies like CATI and CAPI. The paper points out the importance of involving the user in the design process. I do not think this point can be over-emphasized.

In the rush to develop survey instruments on tight time

schedules or in deciding which portable machines to use for CAPI applications, we the developers and/or program managers, take it upon ourselves to decide what is best for the interviewers and may not actively involve them in the decision or development process. This can be a big mistake.

If the interviewers are not comfortable with the interface, if it is slow, clumsy or awkward to use, "not natural" feeling, not helpful, etc., the survey is in serious trouble. If the interviewers have no say in the design and for any reason should decide that the system is not helping them to get the job done better, then you face an uphill struggle to gain their acceptance, and in some instances, the system may never be fully accepted. Interviewers may work to defeat the system, morale may suffer, respondent cooperation may suffer, turnover rates will increase, quality will suffer, and costs will escalate.

In addition, if you are contemplating switching from a personal visit environment to CATI, you must consider the effect on the interviewer staff out in the field. Field interviewers will be

concerned about losing their jobs and quality may suffer during the transition to CATI. How the Field interviewers will be treated and possible impact on data quality during the transition period should most definitely be taken into account. For example, in planning the transition of cases from personal visit to CATI for the National Crime Survey we used attrition among interviewing staff and hard to enumerate areas for conversion to CATI. By using this approach, CATI was viewed as positive tool by Field staff. This plan helped to gain acceptance of CATI.

The third and final area I want to discuss has to do with the need for adequate testing and evaluation of these new methodologies.

Before implementing any survey operation, it is good practice to allow enough time for adequate testing and evaluation of the instrument and the data collection and processing system. This is especially crucial for automated data collection systems. Complex questionnaires (those with complex branching or edits)need to be thoroughly tested and evaluated before they are introduced on a production basis.

While the automated data collection systems provide us with the ability to field much more complex questionnaires than we could using conventional paper forms, they also pose additional challenges related to testing. Aside from the obvious problems that may surface during interviewing, if the instrument is not adequately tested, there may be logic errors hidden in the instrument that go undetected or aren't found until after the data collection phase is complete.

In addition, when changes are introduced to the questionnaire, (even minor ones), thorough testing should be conducted again to insure that other questions or skip patterns have not been affected.

In the paper, the committee discusses the possible application

of expert systems in questionnaire development. I would suggest that perhaps some application could be found for these systems to testing and evaluating as well. There is definitely a need for more systematic and thorough methods for checking out the questionnaire. In addition, attention must be paid to testing the case management, call scheduling, training, data transmission, and processing systems before the survey is fielded.

This is not something that only needs to be done before, a survey is fielded. It should be an ongoing effort to evaluate how well the system is functioning. It should allow for feedback for continuous improvement/refinement such as monitoring, observation, debriefing interviewers/respondents.

I want to thank the organizers for giving me the opportunity to share my views on this important topic. I think the committee has made an important contribution by bringing together in one document many of the issues facing project managers in deciding whether or not to adopt these technologies. I hope that the document will be treated as a dynamic one that will be expanded as

we gain more experience with the various aspects of these data collection methodologies.

DISCUSSION

David Morganstein

Westat, Inc.

I thank Terry Ireland for organizing this intriguing session and I would like to express my appreciation to the speakers for the

work they have done in their examination of new methods for assisting in the processor conducting government surveys. It is a pleasure to be given this opportunity to participate in the session as a discussant.

The job description for a discussant might be:

- To agree with the speakers comments,
- To point out errors or omissions,
- To suggest areas of new research, or
- To do something completely different that they'd like to do!

I think I will try a little of all four of these objectives.

There is a great heed for new approaches to gaining cooperation as the respondent population is increasingly bombarded with requests for survey participation. The initial 1990 Census experience indicates the level of difficulty surveyors can anticipate.

According to our speakers, their "Primary job is to develop ... computer related techniques which allow the respondent to answer the survey completely and accurately". The emphasis on the respondent's cooperation is very appropriate. There is a potential trap of having the software developed by software experts who have little knowledge of or interest in the respondent/interviewer who must use the system. At a minimum, a part of the system designer team should be practitioners of long standing who understand the process. There may be good reason to have this leader-of the team be such a practitioner.

I was concerned by the following statement found in the paper, "Interviewers must believe that Computer Assistance will improve their effectiveness. They need to be convinced that the computer is simply a tool to expedite and simplify their work. This sounds a bit like psychological behavior modification. Such verbal persuasion should be unnecessary. In fact, the users WILL believe and be convinced IF the system actually DOES this! You can be sure that no amount of argumentation will insure the interviewer is support if the system is awkward, difficult to use and makes their

work harder.

The focus of the paper was primarily on the technology. It said little about comparison studies which measure the accuracy/reliability of CASIC responses as compared to more traditional methods. For example, an L84 paper by Waterton & Duffy in the International Statistical Review indicated self-reported alcohol consumption rates that were significantly higher when obtained via CASI than previously measured by interviewer. Perhaps there have not been enough such studies, however, there is a need for them.

The paper pointed out the importance of a good authoring system to CAPI but didn't say the same for CATI. I believe it is true in that environment as well.

Quality Measures (Human Interface discussion) are very important and are needed if we are to evaluate the efficacy of these new approaches. The authors also mentioned an evaluation by 'user' (interviewer), something I agree is important as it speaks to the committees 'primary job' mentioned earlier.

I found the Appendix 3 examples a useful reference for contacts. The authors would perform a valuable service if they would include names and phone numbers for all contacts.

These approaches conform to the modern concept of quality. Reduced variability is designed into the system. They reduce the potential for 'creative interviewing' in which undesired variation is introduced by the interviewer during the interview process.

While I have not worked with CASI, it would appear that it could suffer from a potential loss of control by the survey operator. It could be subject to 'creative respondents' who are intrigued by technology or who seek to befuddle the survey operators. Care must be taken to insure that this does not occur.

The survey instrument's logic/design still depends upon the human mind. Techniques for encoding it into a CATI/CAPI/CACI system need to be better understood. An unrealized advantage of these methods is that they force the designer to better understand the instrument/flow earlier in the process. The designer can't rely upon last minute training/role plays with the interviewers to clarify muddy logic or instrument flow.

I would like to close my comments on the value of these high tech methods for assisting in survey operations with the following short essay on the beauty of the abacus written by Robert Fulghum.

Essay taken from All I Needed to Know I Learned in Kindergarten, Robert Fulghum.

Session 7

QUALITY IN BUSINESS SURVEYS

IMPROVING ESTABLISHMENT SURVEYS AT THE BUREAU OF LABOR STATISTICS

Brian MacDonald

Alan R. Tupek

U. S. Bureau of Labor Statistics,

Introduction

The report on "Quality in Establishment Surveys" (see Statistical Policy Working Paper 15, 1988) concluded that there were few commonly accepted approaches to the design, collection, estimation and analysis of establishment surveys. In contrast to household surveys, there was little standardization of methodological approaches across establishment surveys. The report classified potential sources of errors in establishment surveys and examined the range of practices which are used to improve and measure quality.

Each Federal agency which collects statistical data from

establishments develops their own frame of business establishments. These frames are of varying quality, which greatly affects the methodology for surveys and contributes to the divergence of methodology across establishment surveys.

This paper first provides a summary of the design considerations for establishment surveys as discussed in Statistical Policy Working Paper 15. This paper then describes the efforts at the Bureau of Labor Statistics (BLS) for improving their business establishment list, the effect of these improvements on BLS surveys, and the potential impact on other statistical agencies.

Design Considerations for Establishment Surveys

Establishment populations differ from household populations in several ways (see Statistical Policy Working Paper 15). These dissimilarities result in frame development, sample design, and estimation approaches which are in some areas markedly different from Approaches for household surveys. Among the major

distinctions between establishment and household populations and frames are:

1. Establishments come from skewed populations wherein units do hot contribute equally (or nearly equally) to characteristic totals, as is the case for households; and

2. Accuracy of frame information about individual population units is crucial to sample design and estimation for establishment surveys, while for household surveys the

accuracy of frame characteristics concerning individual units is not as critical to the sample design.

Establishment surveys are characterized by the skewed nature of the establishment population (see, for example, Table 1). A few large firms commonly dominate the estimates for most of the

characteristics of interest. This is especially true for characteristics tabulated within an industry. Small firms may be numerous, but often have little impaction survey estimates of level although they may be more critical to estimates of change over time or for measuring characteristics related to new businesses. This distribution has a major impact on both the frame development and maintenance and on the sample designs used for establishment surveys.

SOURCE: U.S. Bureau of Labor Statistics

List frames are widely used in establishment surveys conducted by the Federal government. The use of list frames for establishment surveys arose from the availability of administrative records on businesses compiled mainly for tax purposes. However, because these administrative record files are not normally developed for statistical purposes, they often need refinement before being used as sampling frames for surveys of businesses. Extensive resources are spent on maintaining the list frames since a significant source of nonsampling error may be due to inadequacies in the frame.

Establishment list frames typically are characterized by detailed establishment identification information, periodic updating of this information, and multiple sources for the information. The data on the frame are required for sample design, sample selection, identification of sample units, and estimation.

The primary source of administrative records for a frame may have shortcomings which require the identification information to be supplemented using other sources of information. This may include using identification information from the surveys themselves. Supplemental filet, including the use of area frames, may also be required to overcome coverage problems in the primary source. Duplication of sampling units may also be a problem associated with the use of list frames. Refinement of the frame includes efforts to unduplicate units prior to sampling.

The individual establishment information on the frame is critical to the effectiveness of the sample design and estimation for the survey. Maintaining a frame over time is complicated by the dynamic nature of the establishment community. Changes in ownership, mergers, buyouts, and internal reorganizations make frame maintenance a real challenge. Matching and maintaining unit integrity over time provides the opportunity for consistent unit identification in the numerous periodic surveys conducted by the Federal Government.

New establishments must be added to the frame. However, it is often difficult to differentiate, using administrative records, new establishments from formerly existing establishments that have changed their name or corporate identity. It is also difficult to link businesses over time when there have been ownership or other changes. Each survey may have different requirements as to handling of new establishments and changes in existing establishments. The timeliness of adding new establishments to the frame and reflecting them in the sample is also a problem. The lag time between formation of new establishments and selecting them into the sample may be anywhere from several months to several years. While new establishments may have little impact on estimates of level, in some instances they may dominate estimates of change.

The Business Establishment List Improvement Project

In May 1987, the Economic Policy Council issued a report that noted five areas in national economic statistics where improvements

were needed. One of these areas dealt with the business lists used by the three major Federal statistical agencies to conduct their surveys. One of their recommendations was that the Bureau of Labor Statistics and the National Agricultural Statistical Service of the Department of Agriculture be designated as the central Federal government agencies for the collection of nonagricultural and agricultural, respectively, business identification information.

In addition, the Economic Policy Council recommended that efforts be initiated to revise the statutes that prohibit the sharing of survey data collected by the Census Bureau with other specified Federal statistical agencies. The main purpose of the Economic Policy Council recommendations was to have a single, high- quality source of business data available to selected Federal statistical agencies in order to increase the quality and comparability of national economic statistics.

Shortly thereafter, the Office of Management and Budget (OMB)

requested that the BLS develop a proposal to assume this role. The issue of devoting resources to developing a central frame is not unique to the fragmented U.S. statistical system. Statistics Canada is in the process of developing a central frame for its business establishment surveys (see Colledge and Lussier 1981).

For the BLS universe file to sufficiently serve as the primary frame for statistical survey sampling by Federal statistical agencies, the BLS recognized that modifications to its existing file were necessary. The most critical need was to improve the information available about employers engaged in multiple operations within a State. The Business Establishment List (BEL) Improvement Project was initiated to do this. Its primary purpose is to create an establishment (i.e. worksite) based register of units with full identification information on United States' businesses. At present, data for multi worksite employers in the BLS register are available mostly at a higher level of aggregation.

The data for the current BLS universe file come primarily from administrative records collected by State Employment Security Agencies (SESAs) as part of the administration of the Federal/State

Unemployment Insurance (UI) System.

All employers covered by unemployment insurance are required to file quarterly UI Contributions Reports with the SESAs for each of their UI accounts. On these forms, employers report the number of full and part-time workers, employed during the pay period including the 12th of each month in the quarter and the total payroll for the quarter. This reporting is mandatory for single location employers as well as those engaged in multiple operations in the State.

Data collection and classification procedures for multi-unit employers differ from those for single units. For multi-unit employers, the statistical branch of the SESA is responsible for the direct collection and review of monthly employment and quarterly wages at the reporting unit (county by industry) level of detail. A multi-unit employer is defined as an employer who has more than one industrial activity (four-digit SIC) and/or county location covered by the same UI account and meets, the following criteria. To quality as a multi-unit employer the employer must

have 50 or more employees in the sum of their secondary industries

or counties. The primary industry or county is defined as the industry or county that has the greatest number of employees.

Under the BEL Improvement Project (see Searson and Pinkos 1990), this threshold is being lowered from 50 employees to 10 employees with the States being responsible for collecting employment, wage and identifying data at the worksite level. Thus, more detailed business identification information will be available for small multi-establishment employers.

Multi-unit employers that do not meet the above criteria are treated as if they were single-unit employers for data collection and recordkeeping purposes. These small multi-unit employers who are engaged in multiple industrial activities within one county are assigned industry codes based on their primary activity (that is, the activity providing the most shipments or sales). Conversely,

those in one industry with several locations are given a county code based on the location employing a majority of all the employees.

Large multi-unit employers are treated differently than single units as they are requested to file a quarterly statistical supplement form in addition to the Contributions Report. On the SESAs' current forms, large multi-unit employers report monthly employment, quarterly wages, industry and location information for each reporting unit. These supplements are used to maintain separate identification and characteristic records on the individual reporting units to ensure correct geographical and industrial totals are maintained.

As part of the BEL Improvement Project, the BLS is replacing the 53 individually-designed State forms with a standardized statistical supplement form. The name of the form is being changed to the Multiple Worksite Report. Each quarter, the employer will be requested to verify the identifying information (trade name, description of the establishment, and physical location address)

for each establishment (worksite) that will be computer printed on the new Multiple Worksite Report. In addition, the employer will be requested to provide the monthly employment and total wages for each worksite for that quarter. By using a standardized form, the reporting burden on many large employers, especially those engaged in multiple economic activities at various locations across numerous States, should be reduced. States will accept listings and floppy diskettes of this information in lieu of the form. In addition, the BLS is investigating the central collection of multiple worksite employers data from major multi-establishment employers. The Multiple Worksite Report form will be used in all States to collect data by establishment (worksite) beginning with data for the first quarter of 1991. Some twenty-one States, however, are switching to a State version of the new form with data, collected for the first quarter of 1990.

As a result of these efforts at worksite reporting, we expect the number of units on the frame to increase from approximately six

million to slightly more than seven million. Because the UI system still serves as the basis for the worksite based frame, both the scope as well as the data on employment and wages on the new frame will be identical to that on the old frame, only the level of disaggregation will be different.

Implications of BEL on BLS Surveys

Several features of the BEL Improvement Project will affect the design of BLS sample surveys (see Plewes 1989). These include:

- reporting unit number for each worksite of multi-unit companies;

- better identification information, including multiple unit multiple addresses, worksite descriptions and telephone numbers;

- better linking of data over time through the use of

reporting unit number for worksites within multi-unit UI account numbers. Also, UI accounts will be linked through the use of predecessor and successor codes for ownership changes such as buyouts, mergers, etc;

- more data items for each unit, such as initial date of tax liability, date of establishing a new worksite, and comment codes for explaining unusual employment changes;

- quarterly data, historical files, and response history files to track the surveys for which a worksite has been selected and whether they have responded;

- linking of units within enterprises or corporations, across UI accounts; and

- improved standard industrial classification (SIC) refiling process, in order to identify new multi-'worksite reporters in addition to updating SIC codes on a 3-year cycle.

The effect of these BEL improvements on four areas of survey design will be examined. These include sample frame development, sample design, data collection, and estimation. Implications for the short-term, during the period in which the survey program will transition into the improved system, as well as the long-term will be discussed. The transitional period implications are usually related to problems in maintaining consistency of survey estimates while BEL improvements are implemented. The long-term implications

are usually related to improvements that can be made to survey designs by reexamining survey design objectives.

Over the years, each BLS survey has developed activities for creating their sampling frame from the old Universe Maintenance System, which BLS will change. These unique activities for each survey focus on specific survey requirements as well as limitations

of the list. For example, BLS surveys which attempt to maximize sample overlap over time must match frame units from one time period to another. The BEL improvements will affect the matching operation, due to the shift to worksite reporting. During the transition period, the surveys may need to reexamine the need to maximize sample overlap. If they maintain this objective, then less sample overlap is likely, and much of the operation will need to be done manually. However, in the long-term the use of reporting unit numbers, and predecessor and successor codes should greatly facilitate the automated matching operation. Other BLS surveys use supplemental frames to survey populations not entirely covered by the BEL. These populations may include railroads; federal, state and local government; religious organizations; and seasonal industries. BEL improvements will allow many surveys to reexamine the need for supplemental frames, especially for state and local governments, and seasonal industries.

Several other long-term benefits for sample frame development are possible through BEL improvements. The availability of quarterly data can be used by some surveys for creating their sample frame. The identification of new businesses on the BEL can

be used as a stratification variable for surveys.

Although BLS does not now do so, the new list will enable survey operators to conduct surveys of enterprises or companies. This will bring about reconsideration of the scope of the surveys. All surveys will need to modify their control file systems to handle additional data items on the BEL.

At this stage of the planning process, certain obvious changes have been identified for each survey. The following three examples illustrate the types of operational modifications which are planned.

First, the survey which is used to develop the Producer Price Index (PPI) must use lst quarter data for measures of size. The BEL improvements will allow PPI to use more current quarterly data, or other quarters for seasonal industries. This is expected to improve the coverage of some industries, and to increase the sample design efficiency.

Second, an annual survey which measures occupational industries and illnesses supplements the BEL with a frame of the 500 largest companies in the United States,

including all of their subsidiaries. Currently, this supplemental frame is developed specifically for this survey. The BEL improvement plan will provide adequate organizational relationships for large companies, so that the separate operation will be terminated.

Third, a monthly survey of employers, which measures employment and average hourly earnings, lags in measuring the affect of new businesses. A sampling strategy is being developed for this survey, which will bring in a sample of new businesses each month, once the BEL improvements are introduced.

Greater flexibility in sample designs will be possible with the introduction of BEL improvements. Separate strata for seasonal or volatile firms can be considered. Stratification by age of firm may be appropriate for some surveys. Surveys designed to produce local area estimates can use worksite locations for stratification. Surveys may want to stratify by multi-reporters versus single reporters, or by enterprise size. The survey response history can be used to avoid overlap between surveys and to spread respondent burden.

During the transition period for BEL improvements, there will be some loss in sample design efficiency. The use of current data to develop sample designs for surveys conducted during the transition period will be somewhat inappropriate. In the long-term, sample design efficiencies will be possible through the use of new design variables and more homogeneity within size classes.

Surveys with size cutoffs will need to reevaluate the survey scope or target population. Some BLS surveys cover only large establishments. For example, most of the occupational wage surveys

cover only, establishments with 50 or more employees. The BEL improvements will shift units between size classes. In general, the sampling unit will shift from a county-wide report to a worksite report. Maintaining a 50 or more employee size cutoff will artificially move units in or out-of-scope of the survey and decrease employment coverage. The effect on wage estimates will need to be examined, and decisions made on how to maintain consistency over time.

Surveys designed to measure change can use the linking of data over time to improve on the efficiency of the sample design through sample overlap. Samples for surveys conducted three or more years apart are now independently selected. With historical relations maintained over time, samples could be selected which improve upon estimates of change, possibly using composite estimation.

The new features of BEL will be most beneficial during the data collection phase. Because of better address information, especially physical location addresses and telephone numbers,

response rates are expected to increase for mail and telephone surveys, since one of the primary reasons for low response rates is failure to reach the correct respondent. Additionally, better address information will result in a decrease in data collection time and effort, such as reduction in telephone and mail follow-up of nonrespondents.

The breakdown of the multi-establishment companies that presently report on a consolidated basis (e.g., county-wide) into establishment or worksite level reporting will affect all BLS surveys. Surveys will need to make special reporting arrangements with these companies to provide data on a worksite basis. Recent cognitive research conducted by Statistics Canada shows that respondents who are in the survey on a regular basis report data in the same manner from one time period to another and usually do not take into account changes to the survey instrument or procedures. The worksite information should reduce the reporting error due to failure to identify the selected sample unit.

The impact of BEL during the estimation process for BLS surveys will vary significantly by survey type and estimation procedures used. An area of survey estimation that will be affected by BEL is benchmarking. Benchmarking is a process that accounts for changes that occur during the time lapse between the reference date of the sampling frame and the date of data collection. In other words, it accounts for births, or those units which have come into existence since the sampling frame was created. This is accomplished by multiplying the sample estimates of totals by the benchmark factor at the estimating cell level, usually SIC or size class within an SIC. For BLS surveys, the benchmark factor is calculated at the estimating cell level as the rator of the reference period employment (benchmark employment) to the weighted employment from the sample.

Surveys that benchmark at the size class level would be most affected because of the change in the distribution of units across size classes due to worksite level reporting. For example, size class benchmarks for a survey that measures occupational employment statistics (OES) by industry may be inappropriate during the

transition period. A possible solution for all surveys which benchmark by size class is to benchmark at the industry level during the transition period.

With the new business registry, population data for benchmarking employment will be available for all 12 months. This additional information may be utilized by the Current Employment Statistics (CES) Survey, which is a monthly survey of about 300,000 establishments that measures employment at National and State levels by industry, to benchmark the employment data quarterly and thereby better analyze the components of error by time period.

Central Agency Status

When the OMB issues the directive naming BLS as the central agency charged with maintaining a list for nonagricultural

businesses, several actions will have to be undertaken before extracts from the BLS list can be made available to other Federal statistical agencies for use in surveys.

First, BLS will have to conduct a series of negotiations with the State Employment Security Agencies to gain their agreement to waive or modify existing State confidentiality rules and regulations that would currently not allow widespread use of the state provided UI data. We expect that most SESAs will readily welcome the sharing for statistical purposes of these data. There have recently been examples where most, if not all, State agencies authorized this type of data sharing, but on a much more limited basis. In those few States where current State law might prohibit the sharing with other Federal statistical agencies, we will propose modifications to the State Unemployment Insurance laws to allow the sharing and work with the state agencies to seek passage of the heeded legislation.

Similarly, there will have to be certain actions taken both by BLS and those Federal statistical agencies authorized by OMB to have access to the BLS list before the sharing can begin. BLS will

have to develop formal procedures for use of the file by other agencies. These procedures will include such obvious items as security measures for the data, assurances that the confidential data will be used for statistical purposes only, agreements on feeding back 'corrections' or updates to the file, access rules and techniques (the BLS list is maintained at the NIH computer facility) and arrangements made to cover marginal Operating costs for providing the data. A possible solution to the question of providing for satisfactory computer security may be for the using agency to have conducted an application security review for its own sensitive Automated Information System in compliance with the requirements of OMB circular A-130.

Summary

A central agency charged with maintaining a list of nonagricultural businesses provides an opportunity for improving business establishment surveys conducted by the Federal Government. However, the key to its success will rest with the ability of all the agencies involved to provide clear and concise requirements to

the central agency, and to weigh the costs of improvements to the central list against the benefits to survey operations and data quality.

References

Colledge, M. and Lussier, R. (1987), "A Generalized Methodology for Economic Surveys" in Proceedings of the Business and Economic Section of the American Statistical Association Annual Meetings, pp. 131-149.

Plewes, T. (1989), "Improving the Business Establishment List: Survey Design Implications" in Proceedings of the Fourth International Roundtable on Business Survey Frames, Newport, Gwent, United Kingdom: Available through the U.S. Department of Labor,

Bureau of Labor Statistics, in press.

Searson, M., and Pinkos, J. (1990), "The Bureau of Labor Statistics' Business Establishment List Improvement Project" in Proceedings of the Sixth Annual Research Conference, Washington, D.C.: U.S. Department of Commerce, Bureau of the Census, in press.

Statistical Policy Working Paper 15 (1988), "Quality in Establishment Surveys", U.S. Office of Management and Budget.

ESTABLISHMENT SURVEYS WITH SOME AGRIBUSINESS EXAMPLES

Ron Fecso

National Agricultural Statistics Service

Working Paper 15 (WP-15), "Quality in Establishment Surveys," addresses the accuracy of establishment surveys. Although WP-15 concentrates on accuracy, we need to recognize that accuracy is only a part of the total quality picture. Remember the importance of other aspects of quality and their interaction with accuracy concepts. The definition of survey quality is the totality of features and characteristics of a survey that bears upon its ability to satisfy a given need. Sometimes these ideas are referred to as "fitness for use." Discussions of quality usually address how well something is made. We must also address the true needs of the product or service as well as productivity issues such as increased output and unit cost. Continued pressure on budgets and demands for increased statistical output are quality aspects which may be occupying major portions of our time. Thus, a model for survey quality needs four elements: accuracy, timeliness,

relevance and resources.

The intent of this paper is to provide a glimpse of the nonsampling error treatment from WP-15 and several examples of the treatment of nonsampling errors in agricultural surveys. I hope that I can persuade the audience to study working paper 15 in more detail after seeing this commercial.

Many sources of error are possible in establishment surveys. While there are several good ways to organize the presentation of these errors, WP-15 chose two main groupings: design and estimation, and methods and operations. The latter group contains the nonsampling errors which are highlighted here.

Nonsampling Errors

Errors which arise during the specifications for and the conduct of establishment surveys are called nonsampling errors. Commonly known examples of nonsampling errors include incomplete

sampling frames, nonresponse and keypunching errors. The variety of nonsampling error sources and results from studies of these sources lead survey researchers to believe that nonsampling errors may often far exceed sampling error. There are three objectives found in the chapter on nonsampling errors in WP-15. The objectives are to outline major categories of nonsampling errors in establishment surveys, to identify some of the diverse sources of error in each category, and to provide insight into strategies to detect, measure, and control these errors. The error categories

discussed are specification, coverage, response, nonresponse, and processing errors.

WP-15 defines each of these error groups, gives examples, identifies major sources of the error, describes methods to control and measure the errors, and profiles the control and measurement techniques used in the major establishment surveys of the Federal

Government (9 agencies and 55 surveys). (The presentation contained some detail about response error treatment and examples of WP-15's graphics since most of the audience had not seen WP-15. These materials are not reproduced here.)

Although several good references are available concerning nonsampling errors in surveys of individuals (for example United Nations, 1982), WP-15 is the first detailed treatment for Federal establishment surveys. The need for this separate treatment arises because establishment surveys differ from surveys of individuals by typically seeking hard data for which records are available. This characteristic both simplifies the collection and complicates the interpretation of the data. The collection is simplified when hard data on record can be used, rather than relying on the memory, opinions, or interpretations of the respondents. These differences present complications when establishing the concepts and definitions to be used in the surveys. Special care must be taken to consider carefully the establishments' recordkeeping systems, definitions, and data availability to avoid introducing specification error into the data.

Establishment surveys, which commonly use list frames, are subject to errors such as duplication, overcoverage of out-of-scope and out-of-business units, under coverage of business births, and misclassification of units. The availability of records affects the structure of the response and nonresponse errors as well as the methods to measure and control them. The treatment of processing errors differs the least from other types of surveys.

SOME HIGHLIGHTS OF WP-15

WP-15, unfortunately, makes no specific recommendations. Yet, the profile of nonsampling error practices used in 55 Federal establishment surveys by nine agencies provides considerable insight into the state of quality in these surveys. This commercial for the paper will present a few of the highlights.

0 No single measurement of specification error is used in a large majority of the surveys profiled.

0 Relatively little is done to measure specification error.

0 Few direct measures of list coverage error were reported as regularly used.

0 Outside of the calculation of edit failure rates, little response error measurement is done.

0 Although follow up procedures for large units are common, very little is done to directly measure nonresponse error.

0 Cognitive studies are rare.

0 Questionnaire pretesting was not widely used on a regular basis.

0 Relatively few nonsampling error measurements are

published.

0 There is relatively little information about processing errors.

WP-15 contains considerably more detail on good practices which are currently in use as well as those practices which are lacking in use and need examinations WP-15 states in an overview that "Nevertheless, the tenor of the findings can be depicted as recommending more work to improve and document the quality of surveys... a need to focus additional attention, and resources, on the general improvement and documentation of survey practices."

A Reinterview Study from Agribusiness

An example of measuring response error in an establishment surveys is next. The results presented are from a reinterview study which measured the bias of Computer Assisted Telephone Interviewing (CATI) methods on a National Agricultural Statistics Service (NASS)

survey.(Fecso and Pafford) As part of its estimating program, the NASS publishes quarterly estimates of crop acreage, intentions to plant, actual plantings, harvested acreage, stocks of grains, and livestock numbers. The source of these estimates is a multi-purpose, multi-frame survey.

Because of the detailed nature of acreage, stocks and livestock inventory items, the NASS had relied primarily on personal interviews to get the most accurate answers from the farm population. For example, on-farm grain stocks data, extremely important because of their effect on commodity trading, is a collection problem because farmers may store these grains in multiple bins on property they own and/or rent. In addition farmers often have multiple operating arrangements involving their own grains, those of landlords, and those where formal and informal partnerships exist.

Recently, NASS has expanded the use of telephoning, including CATI to collect these data. The primary reasons for change are

inadequate budget and the need to reduce the time between initial data collection and publication. We suspected difficulty in using the telephone to collect some of these quarterly survey data. Obtaining accurate responses is difficult because of the detailed nature of these data and the centralized (state) telephone interviewers often lacking farm experience and familiarity with farm terms. The reinterview study is our first attempt to measure response errors.

You cat find the use of reinterview methods in the literature for measurement of simple response variance (Bailar, 1968; O'Muircheartaigh, 1986) and correlated response variance (Groves and Magilavy, 1986), for example. This response error study focused on measurement of the bias by treating the final reconciled response between the CATI and independent personal reinterview response as the "truth." To obtain truth measures, experienced supervisory field enumerators reinterviewed approximately 1,000 farm operations for the December 1986 Agricultural Survey. The following tables contain the results for the grain stocks items

(corn and soybean stocks).

Table I indicates that the difference in the CATI and final reconciled responses, "the bias," was significant for all but one item (soybean stocks in Indiana). The direction of the bias indicates that the CATI data collection mode tends to underestimate stocks of corn and soybeans.

The process of reconciliation identified the reasons for differences. A summary given in Table 2 indicates that an overwhelming percent of differences (41.1%) could be related to definitional problems (bias related discrepancies), and riot those of simple response variance (random fluctuation). Definitional discrepancies contributed almost half of the large bias. About two-thirds of the definitional discrepancies had a relative difference (the reconciled response minus the CATI response divided by the CATI response) more than 25% or less than -25%. In contrast, the differences due to rounding and estimating contributed less than 10% of the overall bias. Almost all of the rounding and estimating relative differences were between -25% and

25%.

TABLE I. Estimates of Bias in CATI Collected Responses

* Indicates the CATI and final reconciled response were significantly different at a=.05.

These results suggest that we can reduce the bias in the survey estimates generated from the CATI telephone sample using a revised questionnaire design, improved training, or a shift in mode of data collection back to personal interviews. considering the constraints of time and budget, the change to additional personal interviews is unlikely. Thus, the alternative is to use reinterview techniques to monitor this bias over time to determine whether the bias has been reduced through improvement in questionnaires or training. If large discrepancies continue, the estimates for grain stocks can be adjusted for bias through a continuing reinterview program. If the bias stabilizes, even at zero, periodic reinterview studies can validate a "constant" bias adjustment used in interim periods.,

NASS conducts crop yield surveys in states which are major, producers of field crops. The survey data are used to forecast expected yield and production during the growing season and to estimate these values at harvest.

Briefly, the survey design can be described as a multiple step sampling procedure. Samples are drawn from an area frame to estimate acreage for harvest, followed by subsampling of fields and small plots to make measurements related to yield per acre. Detailed information on the area frame design is available in Fecso, Tortora and Vogel. More detail on the crop yield surveys, called objective yield (OY) surveys,, is in Matthews (1985), Reiser, Fecso and Taylor (1987), and Francisco, Fuller and Fecso (1987).

Several control procedures existed for the OY surveys. Supervisory enumerators visited the plots (approximately a 10 percent subsample which included the first sample visited by each enumerator). The field office survey statistician occasionally visited plots. Data are hand and computer edited. Finally, periodic validation surveys, covering a subset of crops and states in a given year, were conducted to measure the overall bias of the survey estimate in the domain studied.

These control procedures had shortcomings. For example, visits by the supervisory enumerator served mostly as a retraining

system; the data was not used to improve the estimates or to estimate biases. Budget and staff reductions reduced the number of field visits by survey managers. Edits have been changing. New computer edits and some areas creating individualized recording forms have resulted in estimates which may differ from those based on the old editing procedures. Finally, the expensive and administratively burdensome validation survey received increased questioning.

The validation survey had one major goal -- to measure the differences between the objective yield crop cutting and the farmer's harvest. The validation surveys had clearly shown that the difference between the OY crop cutting and farmer's harvest is not equal to zero. These studies found differences by crop, year, and state. Since the validation surveys have answered the major question for which they were designed, we asked what purpose would they have in the future?

Our main consideration remained the assessment of the bias. Several concepts needed attention. Was the overall bias consistent

over the years? Our data is a time series, especially when considered by the users; thus, knowledge of bias-included level change is important. Are the sources of bias changing? Are there large enough bias changes to deserve extra concern? Are there any needs for procedural changes to reduce specific bias sources, or do we only need to monitor the overall level of bias? Finally, if we use overall bias measures to adjust survey values, are the biases within a specified tolerance?

NASS currently conducts a redesigned validation survey for soybean OY. This survey is done in all states in the OY sample program. This design removed some unpopular aspects of the old validation surveys, including the concentration of work in one or two states and the variable workload resulting from changing states each year. Our goal was to verify the approximate 6% bias adjustment suggested by the historic series of studies. The current approach differs from prior studies. We now combine sources of error rather than trying to measure specific components. Thus, the results provide a basis for adjusting the survey for the many

small sources of error found in prior studies. These errors
These errors

included: incorrectly measured row widths, field counts differing from lab counts, time lag bias due to the enumeration differing by several days from actual harvest, new planting patterns causing enumeration and imputation difficulties, enumerator fatigue errors, and plot location biases.

The rational for the redesign begins with our estimator of state yield, the mean of the sample field yields, which is basically unbiased, except that we do not have the true field yield, Y; but a sampled value, y. This estimate can be modeled as follows:

Three years of data from the validation survey have produced the following results:

Year	Estimated Bias in Bushels	Estimated Standard Error	Bias as percent of the Estimate

1987	2.2	.9	5.8
1988	2.3	.8	7.6
1989	3.2	.9	8.7

Thus, the studies validated the 6% adjustment of the survey data as reasonable. Future research can determine the optimal use of the validation survey for adjustment. We also need to assess the implicit missing at random assumptions. We can get some ideas on the reasonableness of the assumption using farmers reported yields to measure group differences. We need the assumption that the biases measured by the validation survey are uncorrelated with the action of obtaining elevator yields. This assumption is reasonable, but should be tested occasionally. With the redesigned validation survey we have two of the three estimates (the OY yield estimate, the validation survey estimates of OY bias, and a nonresponse bias estimate). These are the estimates, of the major error components which are necessary to assess the accuracy of the between-year of yield estimates.

Conclusion

Although the level of nonsampling error in establishment surveys was not directly measured in WP-15, nonuse of control and measurement techniques should not be interpreted as a lack of errors. Is it time for us to regain the balance between the importance which we put on the elements of survey quality and our actual practice? For too many years, emphasis in most government agencies has been on timeliness and resources (usually shrinking). It's time to shift more effort to relevance and accuracy issues. We might help ourselves by training users in survey quality concepts so they can help us prioritize our efforts and maybe lead the effort to secure more funding. Our easiest beginning in this road to quality could start merely by publishing more of what we do know about the errors.

Increased interest in organized quality efforts such as total u

quality management philosophies is promising. organizations need

to ask questions such as:

1. What measure(s) does top management use to quantify survey or organizational effectiveness? (Is it the same as the data users?)

2. How are these measures used to manage and plan' for the long run?

Agencies need to assess their training needs. We will face at least some shortage of new hires with the survey research skills necessary. Some predict that the shortage will be acute and go beyond survey skills to general quantitative skills. Will agencies respond with creativity in developing staffing and training plans? We should do more to address this problem now.

Finally, WP-15, actually all the working papers, needs to be

more widely read. (Only a small percentage of the audience at the presentation had seen WP-15.) Agencies and users can benefit by identifying errors which were not previously considered and/or techniques which could be used. I caution against being overwhelmed with the quantity of errors displayed it WP-15. Don't worry that you can't eliminate or measure them all at once. I doubt that you have all these errors. Yet, don't be complacent. To improve survey quality you need a strategy. The strategy should define a systematic approach to the improvement and measurement of the effects of existing error sources as well as proposed changes in the survey process. Be flexible as you move along with the strategy, enjoying small successes as they come and avoiding the expectation of overnight miricles.

References

Bailar, B.A., (1968) "Recent Research in Reinterview Procedures," JASA 63:41-63.

Fecso, Ron, (1986) "Sample Survey Quality: Issues and Examples from an Agricultural Survey," Proceedings of The Section on Survey Research Methods, American Statistical Association.

Fecso, R., R.D. Tortora and F.A. Vogel, "Sampling Frames for Agriculture in the United States," J. of official Statistics, Vol. 2, No. 3, pp. 279-292, 1986.

Fecso, Ron and Brad Pafford, "Response Errors in Establishment Surveys with an Example From Agribusiness Survey," Proceedings of the Section on Survey Research Methods, ASA, 1988.

Francisco, C., W.A. Fuller and R. Fecso, "Statistical Properties of Crop Production Estimators," Survey Methodology Vol. 13, No. 1, June 1987, pp. 45-62.

Groves, Robert M. and Lou J. Magilavy, (1986) "Measuring and Explaining Interviewer Effects in Centralized Telephone Surveys," Public Opinion Quarterly, Vol. 50:251-266.

Matthews, R. V. , "An overview of the 1985 Corn, Cotton, Soybean, and Wheat Objective Yield Surveys," USDA, Stat. Rept. Ser., Staff Report. Nov. 1985.

Office of Management and Budget, Ouality in Establishment Surveys, Statistical Policy Working Paper 15, Washington, D.C., 1988.

O'Muircheartaighl Coln A., (1986) "Correlates of Reinterview Response inconsistency in the Current Population Survey." Second Annual Research Conference, Bureau of the Census, March 23-26, 1985 in Reston, Va.

Pafford, Brad, (1988) "Use of Reinterview Techniques for Quality Assurance: The Measurement of Response Error in the Collection of December 1987 Quarterly Grain Stocks Data Using CATI," National Agricultural Statistics Service, Research Report, USDA.

Reiser, M., R. Fecso and K. Taylor, "A Nested Error Model for the Objective Yield Survey," Proc. of Section on Survey Research Methods, ASA, 1987.

United Nations, National Household Survey Caipability Proctram, Nonsampling Errors in Household Surveys, New York, 1982.

DISCUSSION

David A. Binder

Statistics Canada

I would like to thank the organizers for inviting me as a discussant at this important session on Quality in Business Surveys. Prior to these meetings, I reviewed once again the Statistical Policy Working Paper 15, "Quality in Establishment Surveys", and I would highly recommend it be read by both novices and experienced survey statisticians who deal with the design or analysis of business surveys.

One clear fact which comes out of Working Paper 15 is that there are many issues and methods which are common to most federal business surveys. Certain issues faced in business surveys are more difficult than in social and demographic surveys. Part of this is due to the complex and dynamic structures within which the business community, operates. When designing and conducting such surveys, it is important to keep in mind the operational realities of the business world.

Since there are many commonalities among business surveys,

statistical agencies should pool their knowledge and expertise to take advantage of their combined experience. For example, there are sufficiently many common practices for sampling, data collection, editing, estimation and dissemination of the results, that certain, standards and guidelines could be developed among the agencies. Sharing information and expertise is a worthwhile objective which meetings such as this can help accomplish. Whereas legalities of data sharing pose some obstacles at present, hopefully these can be overcome in the longer term.

There are, of course, many aspects to improving the quality of business surveys, including frame issues and non-sampling errors. The development of general purpose business frames can lead to sophisticated and expensive systems, especially with respect to development and maintenance. This is because, a general purpose frame should reflect the realities of the operating structures in the business world and there must also be user-friendly interfaces with such a frame. In practice, there is of ten a gap between conceptual frameworks and actual application.

Quality of the Frame

An important area of concern in the quality of business surveys is the quality of the frame itself. Survey quality will depend on the quality of the frame information as well as the ease of accessibility to the frame data. Frames can never be perfect. Some of the sources of error are:

- undercoverage, especially for births
- overcoverage, especially due to duplication and inclusion of out-of-scope units
- misclassification of industry code, employment size, other size measures, etc.
- identification of appropriate reporting units (collection entities) which reflects the operating structure of the

business

It is important to include in the development of a frame a Program to measure the quality of the frame information. This is particularly true when the frame will be used by a variety of users other than the developers themselves. Examples of quality measures are:

- site of the backlog for SIC classification

- distribution of lag times for births and other updates to the frame

- errors resulting from cutoffs for multi-unit employers

- duplication

- matching errors

If the frame is to contain the most up to date information, there should be some facility for incorporating and verifying

feedback from the surveys themselves. This can lead to complications, where the information being derived from one survey may affect other surveys (e.g. a change in the relationships among multi-unit employers).

Structure of the Frame

If it is anticipated that the Business Establishment Listing (BEL) of the Bureau of Labor Statistics will be used by other agencies conducting business surveys, it should be noted that many of their needs cannot be met within the framework being discussed here. The administrative world does not always correspond to the business world. A listing which is useful for employment and related labor characteristics may not be suitable for surveys of economic production and other special characteristics

The structure of the BEL for multi-unit employers needs some clarification. Whereas the worksite may be able to report employment data, it may not be able to report on profit and loss or

balance sheet data. Different reporting units (collection entities) may need to be identified for different surveys. It

cannot be assumed that the respondent will necessarily conform to your concepts.

At Statistics Canada, we have developed a hierarchical structure of statistical entities for the larger businesses. These are (i) the enterprise, where a full set of consolidated financial statements are available, (ii) the company which can report on profit and loss and other balance sheet items, (iii) the establishment, which can report on such items as value of output, cost of intermediate inputs, inventories, number of employees, and salaries and wages, (iv) the location, which can report sales and number of employees. This recognizes the relationship between the business world and the statistical needs for economic surveys. However, it is a complex structure to maintain.

Retrieval Systems

Not only are frame maintenance procedures resource intensive, but effective retrieval systems can be quite complex. and expensive to develop. Quality improvements to business surveys through better quality frames can only be realized if the frame information is easily obtained both cross-sectionally and through time. Examples of some of the needs which are expressed by users of frame information are:

- linking of data through time

- historical files

- response histories

- linking of data within enterprises

- identification of seasonal and volatile firms

- having sufficient structure to roll up to enterprise and

track changes in structure over time

- survey feedback (and verification)

- requirements for estimation (regression, ratio, composite, benchmarking, poststratification)

Other Frame Considerations

The needs of the frame will change depending upon the survey frequency and the reference periods. For example, the units considered in-scope could vary according to whether the survey is monthly, quarterly or annual.

Even with all the complexities I have mentioned regarding the development and maintenance of business frameso I would strongly encourage such development, with any deficiencies explicitly laid out. One of the uses of a high quality frame is the ability to perform analyses of business demographics, showing behaviour of

births, deaths, mergers and amalgamations, which is an important side benefit.

Total Survey Error

As was pointed during the session, improving frame quality is only one of the many mechanisms to meet the overall objective of controlling survey errors. Development of survey quality profiles has been mentioned as an important tool to monitor, control and manage surveys.

Response errors should be a particularly important concern to the survey-taker. However, response errors are often due to the survey instrument itself, rather than the respondent. Recent experiences with cognitive methods have proven useful here. Often there are trade-offs between ideal concepts And the respondents' ability to respond accurately. For example, when asking a farm operator about value of equipment on land which he operates, he may prefer to report on equipment which he owns but which may be

situated on another farm, rather than including equipment which is owned by someone else, but which is situated on his land. This creates difficulties for the survey-taker who is trying to avoid coverage errors. These are not easy problems to overcome, but the first step in all these endeavors is to recognize the problem and possibly measure its impact. Without special studies, it would be difficult to assess the relative merits of coverage error on the, one hand and response error on the other.

In general, we need to concentrate on methods to synthesize all the errors into, an overall measure of survey quality. This would allow informed decisions to be made regarding the relative merits of improving one survey process over another. If such a model existed, we could answer some common concerns such as the relative contribution of edit and imputation to the reduction in total survey error and whether simpler methods could achieve comparable results.

One possibility would be to use develop a microdata simulation database which incorporate as many of the known errors as possible.

This database would consist of microdata which look like the real population. Various models for response and nonresponse errors could be simulated and then the data would be processed using existing or proposed methods. Since the original "true" data are known, we could assess the relative impacts of improving survey coverage versus using an Alternative estimator versus adding more edits to the survey process, for example.

DISCUSSION

Charles D. Cowan

Opinion Research Corporation

What These Papers Have in Common

If there is a single message that comes through in both the papers being discussed, it is that:

Avoidance and/or Control is the Best Approach in Dealing with Nonsampling Error.

Quality is something that one builds into surveys and continues to monitor. While one cannot completely avoid problems in surveys, it is markedly better to avoid or control a problem than it is to attempt to make an a posteriori correction to fix the problem. Such a fix usually is based on a much smaller amount of information collected from a supplemental sample or survey and adds variance to the original survey estimates. It is also usually the case that a fix introduced at the end of a survey only takes care of one problem and is not very cost efficient.

In their paper, Tupek and MacDonald describe a process of expanding a sampling frame for business surveys that addresses several different sources of nonsampling error. Their work with the sampling frame deals with coverage issues, timing issues, definitional problems in the surveys, estimation, use of administrative records for weighting and variance reduction, and other aspects of the conduct of business surveys. Their approach

is to improve the basic materials used for surveys to encourage more efficiency and accuracy at later stages.

Pecso in his paper describes a process of measuring and controlling as many aspects as possible of incidence of nonsampling error. He also supports the idea that nonsampling error it best dealt with by avoidance, but is also realistic in suggesting that a catalog of problems is useful for two primary purposes: planning future surveys and providing documentation for users of the current effort. This control process can be used to ensure that the data produced in a survey are of the best quality given the constraint that control is imposed as part of the process, since many types of nonsampling errors cannot be totally avoided.

Specific Quality Issues for Business and Establishment Surveys

As one reads and compares these papers, one is reminded of the fact that business and establishment surveys, are different household surveys in several key ways:

1) The availability of attributes on the frame and the use of this frame information at the unit level differs from what can be done in household surveys,

2) The surveys themselves make extensive use of records as a basis for reporting, and

3) The data to be collected in business and establishment surveys has a multilevel nature, meaning that information about the businesses is hierarchical and we are interested in the information at each level (e.g., Sears Headquarters, regional offices, distribution centers, and individual stores).

These factors are crucial to the design of business and establishment surveys. Use of information on the frame for design

and use of records in collection makes it possible to improve the quality of these types of surveys relative to household surveys, but, this is counterbalanced to an extent by the complications introduced by the multilevel nature of the data to be collected.

Tupek and MacDonald note in their paper that for the surveys they conduct that establishments come from skewed populations, and having this information on the frame makes it possible to design a survey that is much more efficient, especially for multiple characteristics to be measured simultaneously. However, reliance on this information in the frame makes the accuracy of frame information crucial at the individual unit level for both sampling and estimation purposes. Their project on frame expansion and improvements has an impact in several areas. The first is sample frame development, so that more business and establishments are represented. This is broader than a coverage issue, since coverage is usually viewed, as a problem that pervades an extant frame. Tupek and MacDonald address coverage issues in this way, but also include whole segments of the business population previously excluded from the frame.

A second area impacted by. the frame expansion and improvements project on which they report is the actual design of the sample, where the sample can be optimized for making different types of estimates using information available on the frame. A third area impacted by the frame expansion and improvements. is in data collection, and the final area is in estimation. Tupek and MacDonald point out that the new frame encourages the conduct of new longitudinal surveys, the selection of sample at the unit of analysis (instead of collecting the information by proxy or sampling down to the unit of analysis after starting at a higher level in the hierarchy), improvement in response rates because of higher eligibility rates, savings in terms of time and effort expended on the survey, and improvement in weighting and ratio estimation procedures.

Fecso takes a different approach to dealing with nonsampling error. He catalogs sources of nonsampling error, and his approach

is to detect, measure, and control the nonsampling error. Many of the sources of nonsampling error he lists are common to both household and business surveys, but with business surveys he has a variety of records, including past survey collections, available for detection and measurement of nonsampling error.

A primary concern for the use of records is the accuracy of the data in the records, since the records themselves could be in error. Although not mentioned in the paper, some of the most interesting work in health care surveys is modeling of nonsampling error when hospital records and information based on patient recall don't match and either is potentially wrong. The same is true for business surveys -- accuracy in the records systems is crucial for detection and measurement of nonsampling error as part of a quality management system for a survey. Another factor related to accuracy is the consistency of definitions used by different respondents. If the data are accurate but based on different definitions, then there is a problem in how the data might be used for detection and measurement of nonsampling error.

Concerns with Business and Establishment Surveys Not Covered

While both papers are excellent in the way they cover in depth quality issues facing business and establishment surveys, they both miss some salient points peculiar to these types of surveys. The first was mentioned earlier, namely that businesses are hierarchical, which leads to some difficult questions regarding who reports in these surveys, and how the various businesses relate to one another (i.e., at what level do we define the unit of analysis?). In terms of how units relate an example was given earlier for Sears, which owns not only Sears Retail, but also has Allstate Insurance, a mailing service, regional offices, catalog stores, and local retail stores. Are we interested in these surveys in getting reports from the lowest level in this chain? How does Sears headquarters report exactly -- for itself as an establishment with a certain number of employees, or does it include all employees and sales at all locations? If there is confusion in reporting rules for a survey, we could wind up with severe overcounting or undercounting of activities and personnel.

Another issue has to do with the reporting of activities within a firm. In reporting mailing activities, for example, each firm and each location of a firm will have some activities to report. To whom do we speak in the firm to get a complete picture? There are separate operating units within firms, each with a manager knowledgeable about his own unit's activities. And there are sometimes other units that assist in terms of technical or operational support. Do we talk to managers in both or all offices

or units, or is there a central source that can answer all questions knowledgeably and without duplication?

There ate two final concerns we have regarding quality in business and establishment surveys. One has to do with the process of improving and expanding the frame for a business survey, which usually translates into adding smaller firms. These firms are more

likely to be related to other members of the population, and they are more prone to movement in and out of the population (births and deaths). Because of these factors, they add a certain amount of instability to the estimation process. This may be good or bad -- on the one hand we have a more realistic representation of the population of businesses when we include more firms, but on the other hand for certain types of statistics we may be adding more variation without a real gain in forecasting or descriptive accuracy. This problem could be labeled: "messiness at the edge".

The other problem not addressed in either paper, and of particular concern in the Fecso paper, is that a large, well conceived and executed survey might not benefit from a Nonresponse/Nonsampling Error Correction that it estimated from a small onetime experiment. While in theory the idea of implementing research studies to monitor the quality of ongoing surveys is laudable and should enhance the quality of the surveys, implementation for Federal surveys often falls a bit short, with a simple, one-time study implemented to measure a particular problem. A small scale, high variance research study should be viewed as just that, and not a vehicle for making corrections to a

multimillion dollar effort. If the nonsampling error problem is sufficient to justify such an effort, and the nonsampling error cannot be dealt with as part of the design, then sufficient resources should be devoted to measurement and control to take care of the problem. Essentially, the problem becomes one of design again, with focus on the Proper allocation of resources between the survey and the experiment to fix the survey.

Conclusions

Both papers were excellent summaries of the state of the art for measuring and maintaining, quality in Federal surveys of businesses and establishments. Researchers involved in the design of either business or household surveys would benefit from studying and implementing the principles found in either paper.

Session 8

COGNITIVE LABORATORIES

THE BUREAU OF LABOR STATISTICS' COLLECTION PROCEDURES, RESEARCH LABORATORY: ACCOMPLISHMENTS AND FUTURE DIRECTIONS

Cathryn S. Dippo

Douglas Herrmann

U. S. Bureau of Labor Statistics

I. Introduction

The accomplishments of the Cognitive Aspects of Survey Methodology movement (Jabine, et al. 1984) have clearly been substantial. This is especially true in Washington, where three Federal agencies (Bureau of the Census, Bureau of Labor Statistics (BLS), and the National Center for Health Statistics) have established laboratories.

Consider the scope of BLS' survey research programs. Most of

the sampling units from which data are collected by or for BLS are establishments. While approximately 60,000 households are questioned about labor force participation each month in the Current Population Survey (CPS), 340,000 establishments are being asked to report their payroll employment each month in the Current Employment Statistics Survey. More than 200,000 price quotes are being collected each month from establishments in the Consumer Price, Producer Price, and International Price Index programs. Moreover, much of the data are currently being collected by mail, without person-to-person interaction. In the future, more and more of the data will be collected with computer assistance, and the human-machine interface will take on added importance. Furthermore, in most establishment surveys, the needed data can be directly observed (e.g., consumer prices) or exist in records rather than in the memories of the respondents. Even in household surveys, many respondents are being asked to recall not only autobiographical events, but also information that exists in household records and information about other members of their household.

Thus, the mission of the Bureau requires the BLS laboratory to

consider more than just questionnaires to be used with personal visit interviewing in the context of a household survey about autobiographical events. The Bureau acknowledged this fact when selecting the name for its laboratory -- the Collection Procedures Research Laboratory (CPRL) -- which was established in 1988. The basic goal of the CPRL is to improve through interdisciplinary research the quality of data collected and published by BLS. As originally envisioned, all forms of oral and written communication used in the collection and processing of survey data are appropriate subjects for investigation, as are all aspects of data collection, including mode, manuals, and interviewer training.

The CPRL's staff includes cognitive psychologists, social psychologists, sociologists, and a psychological anthropologist. For most of their projects, they work closely with the economists or program specialists responsible for defining the concepts to be

measured by the Bureau's survey programs. To augment staff resources, the CPRL has labor hour contracts with the Institute for Social Research at the University of Michigan and Westat, Inc. The laboratory also does work under contract for other Federal agencies such as the Internal Revenue Service.

Although the CPRL has only existed for two years, its research program has been both broad and prolific. In section II, some accomplishments of the CPRL, are reviewed. The discussion is organized within the framework of an information processing model. In section III, some directions for future research are described. The success of focusing on the cognitive system suggests that focusing on other behavioral systems may produce further gains in data quality through improved survey theory and practice. Moreover, the success of using laboratory techniques for investigating the data collection processes used in sample surveys leads us to believe the techniques can be useful in improving other aspects of survey design.

II. Accomplishments to date

The CPRL has integrated the cognitive approach into the Bureauls survey research program to good effect in many ways. Primarily, the laboratory has changed how data collection research is conducted at BLS. Not only has the research conducted to date affected our understanding of the survey process, but the fact of its existence has heightened awareness throughout BLS of the need for a better understanding of all aspects of the data collection process (Norwood and Dippo in press).

Some results of the CPRL's research efforts are presented here within the framework of an information processing model (Cannell et al. 1989; Tourangeau 1984) that has four distinct stages: comprehension, retrieval, judgment, and communication. As applied to respondents, these stages refer to the comprehension of a question, retrieval of pertinent information, judgment about the accuracy of the information retrieved, and communication about this information within social and other restrictions imposed by the survey situation. As applied to interviewers, these stages may

refer to comprehension of the question, retrieval of appropriate ways to say the question aloud, judgment about whether the respondent has understood the question, and communication to ensure the question has been understood (such as by rereading it) or, if the question has apparently been understood, to indicate that another question is about to be presented.

A. Comprehension

Question comprehension clearly requires that the terms making up a question be correctly understood. The accuracy of term comprehension has been shown by many psycholinguistic investigations to differ in certain ways.

Multiple meanings of terms: A term may lead some respondents to answer inappropriately because it may convey a meaning different from that intended by the designer. Research at BLS has

accordingly attempted to identify terms with several meanings that are not made explicit by the phrasing of questions and might be likely to produce misinterpretations. Since the issue of employment is of personal significance to most people, questions about employment status are likely to predispose respondents (especially the unemployed or those with insecure employment) to be influenced by social desirability when answering the CPS (DeMaio 1984; Edwards, Levine, and Allen 1989). The misinterpretation of employment status terms may easily occur in a survey such as the CPS (Martin 1987).

Accordingly, respondents I interpretations of two key terms on the CPS concerning unemployment status, "on layoff" and "looking for work," have been examined. The CPS definition of unemployment refers to persons who were not employed during the survey week, were available for work, and had made specific efforts to find employment sometime during the prior four weeks. Persons who are waiting to be recalled to a job from which they have been laid off need not be looking for work to be classified as unemployed. As expected, research demonstrates that these terms are sometimes

misinterpreted by laboratory respondents to the CPS. Similar research into the effects of multiple meanings of terms has also beet conducted for several sections of the Consumer Expenditure (CE) Interview Survey, including the sections on medical care, home purchase, and trip expenditures (Miller and Downes-LeGuin 1989). Since our results indicated that people interpret "payments" in different ways, the section on medical care expenditures has since been modified to avoid misinterpretations of this term.

Diverse Meanings: Diversity of term meaning also may impair comprehension. For example, in a recent pilot survey of business establishments, respondents were asked to report all "nonwage cash payments" paid to employees during the calendar year. BLS defined the payments to include bonuses and awards, lump-sum, cash profit sharing, and severance payments, and nonregular commissions, but since this technical term probably was not too familiar to respondents, the meanings of "nonwage cash payments" can be expected to vary across respondents. When the interpretations of this term by respondents were investigated, it was found that respondents interpreted "nonwage cash payments" in a diverse

fashion. Some interpreted it too broadly to include payments in kind, such as a new car (Boehm 1988), and some too narrowly to

include only cash And not cashable checks (Phipps 1990). Another group of respondents who had made such payments simply checked they had made no payments because of a lack of understanding of what the term included. Respondent exclusion and nonreporting of payments were more serious comprehension errors than inclusion of inappropriate payments, contributing to underreporting.

Format Properties: When respondents complete a survey form received in the mail, the format of the instrument may play a crucial role in the respondents' comprehension. If the format does not make it- clear what parts of the instructions are essential, respondents may overlook these parts and respond inappropriately. For example, in the Nonwage Cash Payments Pilot Survey (Phipps 1990), instructions, definitions, and examples were on the back of a one-page questionnaire, for which two different layouts were

,used. one layout required respondents first to provide an annual nonwage cash Payment total and an annual payroll total, then answer a set of yes/no questions asking if they made specific types of nonwage cash payments. The second layout placed the set of yes/no questions first, with the payments and payroll totals requested at the bottom of the page. Reporters receiving the second layout were much less likely to provide the annual payroll total, stating in retrospective interviews that they overlooked it or did not understand they were to provide it. Thus, the layout of the second form, combined with a lack of instruction, caused an entire section of the form to be overlooked. As expected, the format of a survey played an important role in the respondents I comprehension of survey items.

The types of cues used on a self-administered form like an expenditure diary also can affect comprehension. In developing a diary for recording clothing expenditures, alternative cueing levels were tested in a laboratory. Results indicated that a shorter diary with multiple pages that repeated the general cues, e.g., buying clothes, was more effective than a longer, more structured version with specific cues. Respondents were better at

clarifying the domain of purchases to be recorded with the general cues than with the specific cues, i.e., the specific cues led them to restrict their comprehension of listed items more narrowly than intended.

B. Retrieval

Most Federal surveys require respondents to retrieve information about factual or autobiographical events. Faced with the need to control data collection costs, the time period for which the events Art to be recalled is often long. For example, the reference period for the CE Interview Survey is three months. In the CPS, respondents may be asked questions about last week, the last four weeks,, or the last time they worked, which could require recall for a long period of time. (For further discussion of

memory retrieval errors in CE and CPS, see Dippo 1989 and Mullin

1990).

Cues: Often a situation is inadequate in the cues it presents for retrieval. Alternatively, when enough appropriate cues are brought forth, a person can retrieve the previously "forgotten" memory. while some information is probably lost from memory due to diseases and environmental influences (such as alcohol), cues clearly play an important role in retrieval. Accordingly, several investigations have attempted to increase response, accuracy on surveys by providing additional cues to retrieval, e.g., Lessler, et al. (1989). Still, it is important to recognize that some cues can be misleading and ensure that a respondent does not retrieve the appropriate information. Cues facilitate only when they correctly direct retrieval.

In the Nonwage Cash Payments Pilot Survey, underrepotting was investigated by presenting cues to facilitate retrieval. When respondents (company representatives) were given specific cues pertaining to bonus and award payments, recall of such payments was 11 percent higher thin without cues (Phipps 1990). Also, in the CE

Diary Survey, cues with varying levels of generality have been tested. For example, general cues included "beef (ground, roasts, steaks, briskets, etc.)" and specific cues included "ground beef, chuck roast, round roast, other roast, round steak, sirloin steak, other steak, other beef and veal." Underreporting was greater with general cues for certain items, particularly nonfood items. On the other hand, the level of reporting for many food items was not affected by the type of cues (Tucker and Bennett 1988).

Strategies: To get accurate recall about the past, it is necessary to get people to retrieve the mental records of what they actually did. Several strategies to get respondents to access their memories of experiences have proved useful in our investigations at BLS. One strategy has respondents recall a critical personal event that occurred in the reference period in order, to anchor the period. A second strategy has a respondent consult a calendar when attempting to recall. A third strategy has respondents decompose events recalled into smaller events to ensure that what is being recalled is a real experience and not a stereotypical schema. Research funded by BLS has found that

respondents vary in the extent to which they employ the strategy that they were instructed to use. only one-third of the laboratory subjects instructed to use a decomposition strategy when responding to questions on their hours worked used the strategy. Also, the vast majority of proxy respondents presented with this strategy ignored it because they did not have the knowledge necessary to use it.

Expertise: In a laboratory study of household respondent pairs using the CPS questionnaire, proxy responses disagreed with those of the self-respondent approximately one-third of the time

(Boehm 1989). In another laboratory study, when responents were instructed to use the decomposition procedure, the vast majority of proxy respondents ignored the procedure, since they did not have the knowledge necessary to use it (Edwards, et al. 1989). Self-respondents were found to overreport and proxy respondents to underreport the hours worked. Also, proxy respondents were more likely than self-respondents to make errors, and their errors

tended to be larger (see also Tanur 1990). As might be expected, proxies fail in areas they are less likely to know about. For example, proxies underreport more when the I person reported on worked weekends or worked extra hours. Also, proxy error was greater when the respondent was unrelated to or from a different generation than the person to whom the data related (Edwards, et al. 1989).

C. Judgment

People may recall correctly but not realize the recalled information is correct. They may recall correct information, know it is correct, but express it inappropriately because they misconceive how responses are to be expressed. It was noted above that field research on the CE Diary Survey indicated specific cues were often more effective and led to less underreporting than general cues (Tucker and Bennett 1988). Laboratory research has indicated that judgment is also a factor. When given specific cues, laboratory subjects were sometimes unsure of where to record

products on the form. Whether this hinders reporting is still an open question, but the accuracy of reports is affected (Tucker, et al. 1989). The specific cues also may make the task more onerous.

D. Communication

The importance of communication to cognition has largely been recognized in social psychology and anthropology. A considerable amount of survey research has shown that respondents' inclination to answer questions may be affected by the social desirability of the answers. In some cases, respondents may be disinclined to answer because they do not want to share certain kinds of information. In other cases, they may not want to present themselves in a bad light. In other cases yet, they may want to adapt their response to what they perceive to be the expectations of the interviewer.

While BLS has yet to complete an investigation of communication, it has recently begun several such investigations.

First, the laboratory is conducting research into the psycholinguistic factors that persuade a respondent to provide confidential information to a survey (Herrmann, et al. 1990). This research will indicate the degree of trust elicited by different protection terms (confidential, private, secret, concealed,

nondisclosed). Second, we are examining the influence of interviewer errors an the errors of respondents using techniques developed by Cannell (Cannell, et al. 1989). For example, tape recordings of CE Survey interviews are being analyzed to determine whether the quality of answers produced by respondents varies with the quality of the interviewers' presentation of a question. Third, like other agencies we are investigating the use of computer-assisted telephone interviewing (CATI) for some BLS surveys. Research is underway for the CPS, CPI-Housing, and Continuing Point-of-Purchase surveys to determine if people respond in the same manner in a computer-assisted telephone interview as they do in a personal interview. It has been suggested that the

personal interview ensures better attention from the respondent, but it has also been suggested that CATI elicits information that otherwise might not be disclosed because the respondent feels less personally involved when interacting with an interviewer on the telephone. In various ways our research is addressing these alternative expectations about CATI.

III. Future directions

Prior to the establishment of the laboratory, BLS sponsored a Questionnaire Design Advisory Conference to seek advice on the types of questionnaire research that should be undertaken for the CE and CPS (Bienias, et al. 1987). The conference participants all advocated the incorporation of cognitive concepts into the BLS research program and suggested that research focus on the issues of respondent rules, respondent and interviewer roles, questionnaire form and content, and statistical estimation.

In addition, our ongoing research program has taught us that many aspects of the data collection process require a broader

integrated-systems approach rather than a cognitive approach to research. The accuracy and efficiency of survey responses are affected not only by cognitive variables (e.g., abstractness of terms, retrieval cues) but also by other kinds of variables (e.g., physiological, perceptual, emotional, motivational, social, societal, cultural, and economic; see Royce 1973). In some cases, these variables affect responding because they interact with the quality of cognitive processes underlying responding. In other cases, these other variables leave cognitions unaffected but instead interact with a respondent's inclination to report accurately about these cognitions.

A. Looking beyond the cognitive approach

An integrated-systems conception of cognition has been advocated increasingly in recent years by scholars in anthropology (Cole and Scribner 1974), psychology, and neuroscience. Some noncognitive psychological and societal factors that may affect the

response process are: Physiological condition, perception, emotional state, motivation, familial roles, and societal norms.

Physiological condition: The accuracy and efficiency of cognitive responses are affected by the physical state of a person's body (Squire 1987). Physiological condition, as affected by physical health, influences a person's ability to understand, remember, reason, and analyze. A variety of routine health conditions (such as the common cold) may impair the accuracy and/or efficiency of cognitive processes (Cutler and Grams 1988). Cognitive processes are also impaired by commonly imbibed substances, such as coffee, tobacco, tranquilizers and antidepressants, and even certain antibiotics.

The CPRL has been sponsoring laboratory research on the effects of computer-assisted personal interviewing (CAPI) on the interviewer (Couper et al. 1990). Although the studies have been within the context of the Consumer Price Index survey, where

interviewers conduct interviews both on the doorstep of housing units and walking the aisles in retail establishments, the procedures developed, concerns raised, and results are generally applicable. For example, more than, 40 percent of the 46 interviewers who volunteered to be laboratory subjects stated that they had suffered neck, shoulder, and/or lower back problems in the 12 months prior to any contact with a portable computer. Moreover, approximately 75 percent of the subjects wore some form of corrective lenses, with bifocals presenting particular problems for interviewers trying to focus on the keyboard, screen, and respondent.

Perception: The quality of visual stimuli affects the ease of reading and comprehension. The role of perception is of special importance in many Federal surveys where data are collected via a self-administered form. For these surveys, the perceptual constructs may have significant effects on the quality of data. Wright (1980) suggests classifying form-design issues into three categories: the language of forms, overall structure, and the substructures within the forms such as the questions themselves.

In addition, there are perceptual issues related to the appearance of questionnaires, such as color and print font.

The presence of visual stimuli affects retrieval processes more than thinking about or imagining the stimulus. For example, psychological research indicates that the frequency at which academics use external aids, such as files and piles of papers on one's desk, has been found to be positively correlated with scholarly productivity (Hertel 1988). Survey research indicates that expenditure reporting increases with the use by respondents of an information booklet describing the types of items that belong to the categories being read aloud by the interviewer. More respondents appear to be willing to read the item lists than to listen to an interviewer read the list to them.

Respondents to the Occupational Safety and Health Survey face a very difficult task in deciding if an incident is an injury or an illness and if it is reportable or not. Currently, respondents receive a 22-page set of guidelines. Laboratory staff are now

investigating different methods for communicating the decision logic to respondents, i.e., flow charts or graphic representations of the decision paths. In addition, a simple user's guide (no more than 10 pages) is being prepared for respondents who are new to OSHA recordkeeping. Unlike the longer guidelines, this guide contains background on the 1970 OSHA act and provides examples on how to recognize, record, and report occupational injuries and illnesses.

Emotional state: Our cognitive ability to comprehend, retrieve, evaluate, and respond may be affected by our emotional state (Wolkowitz and Weingartner 1988), which in turn may be affected by recent events or prolonged stress. Stress, a major factor moderating emotional states, has been associated with cognitive failures in everyday life. Sometimes, emotional states may prevent people from producing correct responses, that they "know" at some level. For example, despite decades of controversy, it is now generally accepted that sometimes people repress memories.

Nontrivial levels of stress are currently experienced by interviewers. With the change over the next decade to increased CATI, the possibility of increased interviewer stress is real. In surveys like the CPS, the proportion of personal visit interviews will increase for most interviewers working in large metropolitan areas as many of their telephone interviews are transferred to a centralized CATI facility. Concerns about personal safety and administrative pressures to maintain high response rates are but two factors which may contribute to increased interviewer stress. In a centralized CATI facility, interviewers know their work is constantly being monitored. Recent news stories about the effects of constant observation and work quotas in the telephone industry indicate stress levels can be very high in these kinds of situations.

Motivation: We know little about respondents' motivations for responding to survey questionnaires. Census' recent experience of overestimating the mail-return rate in the decennial census is but one indicator of how little we know. At BLS, those of us working on the CE Interview Survey constantly wonder why anyone would agree

to an interview that is expected to last 2 hours. To investigate survey respondent motivation, a large-scale research project on household survey response has been initiated by Robert Groves at the University of Michigan, sponsored by the Bureau of Justice Statistics, the Bureau of Labor Statistics, and the National Center for Health Statistics. One part of the project is an examination of both interviewer (e.g., attitudes, behavior, and characteristics) and administrative (e.g., procedures, workload

levels, design parameters) influences on survey participation (Groves, R.M. and Cialdini, R. 1990). To examine the effects of alternative forma of persuasive communication on sample attrition rates and item response rates, BLS is conducting experiments using appeals that stress the use of Current Employment Statistics data by the trade associations representing the establishments (McKay 1990).

Familial roles: The roles people assume within the family have been found in recent years to affect cognitive processes. While it may be assumed in some surveys that people within a home are equally able to answer questions pertaining to the household, research shows that different family roles carry responsibility for knowing -about certain kinds of information. For example, wives tend to know more about the health and activities of children whereas husbands tend to know more about how community activities affect the household. Single parents tend to know the information possessed by both spouses in dual-parent households.

With the prevalence of proxy reporting in most household surveys, the importance of :learning about what information is exchanged within households and how should not be understated. Recent research on proxy reporting in the CPS indicates adults may be worse proxy reporters for youths than, for other adults in a household (Tanur 1990). Moreover, the proxy reporting of job search may be dependent upon the type of job search strategies being used by youth. At Tanur notes, there is no literature about family communication patterns and the issue of who in the family talks to whom about what.

Societal norms: Cognitive performance is affected by groups in several ways. For example, people are disinclined to perform memory tasks when the social stereotypes that apply to them indicate that they cannot perform well, such as the stereotypes associated with age or with gender. Also, people will sometimes knowingly give the wrong answer to a question because they recognize that their answer is contradicted by the other members of a group.

Moreover, social pressures sometimes dispose people to communicate falsely what they do or do not know in order to achieve social goals. For example, people may say they cannot recall some event or information to avoid, or speed up the questioning or to make a certain impression on the questioner. We do know that social desirability plays a role, but there has been little research into understanding the role (DeMaio 1984). We also know that the mode of data collection appears to have an effect on data, but we do not know why (Shoemaker, et al. 1989). Recent research by Suchman and Jordan (1990) shows clearly the influence of social

and cultural variables.

Evidence indicates that members of all cultures can equally perform all manner of cognitive tasks if the environment has provided the cultures equivalent education and experience. However, because cultures typically involve different educational systems, belief systems, and occupational opportunities, members of different cultures acquire different cognitive skills (Cole and Scribner, 1974). hus, members of different subcultures of a multicultural society will interpret certain concepts differently and answer differently.

B. Looking beyond the interviewing process

The research laboratory and laboratory techniques can be used

in a variety of survey design applications. Just as the responding process is affected by noncognitive variables, the survey process consists of more than just question answering. The entire survey design process, from defining the concepts to be measured through analyzing the data, involves the communication of concepts between people with different knowledge bases or an interaction between people and things. The process can benefit from a broad range of interdisciplinary research including both cognitive and other areas of psychology, other behavioral sciences, and human neuroscience.

The importance of the role of the interviewer has long been recognized. Data collection and training methods designed to control interviewer error, such as structured questionnaires and verbatim training, have been developed in an attempt to control interviewer error. Interviewer training typically stresses the need for neutrality, the use of specified questionnaire wording and administration procedures, and appropriate probing techniques. Recognizing the importance of this source of error, many BLS-sponsored laboratory studies conducted in the last two years have focused on the interviewer. These studies indicate the role of the

interviewer can be studied effectively with laboratory techniques. Thus, it seems natural to expand our research in this area.

IV. Summary

As survey researchers, we really know very little about the psychological processes underlying interviewer and respondent behavior. The few laboratory studies to date indicate the cognitive approach is very useful. With this approach we are learning about the roles of comprehension, recall, judgment and communication in the survey response process. Eventually, as we learn more, we can develop detailed models which questionnaire designers can use to assess new questions and forms for survey data collection.

Just as the research to date has shown that the cognitive approach is effective, it has shown that a more broad-based approach is necessary. Survey responses clearly emanate from all

behavioral systems within and outside the respondent. An understanding of how responding is affected by the cognitive system is not enough. A respondent's behavior is influenced by physiological, emotional, social, societal, and economic variables. A complete explanation of responding requires an understanding of all systems and how their influences are integrated overall to produce a response.

The adoption of an integrated-systems approach would be a natural step in the evolution of survey science. Consider the disciplinary history of economic statistics. First, there were economists producing simple descriptive statistics. The discipline of mathematical statistics was not really incorporated until probability sampling became the basis for sample designs. Then came the advent of computers. Just as we have expanded our use of statistical theory as applied to survey research beyond just sampling (e.g., to incorporating operations research techniques in sample design optimization and iterative methods such as raking in survey estimation) survey research may progress further by making use of not only cognitive psychology but also of knowledge of other

psychological and sociopsychological systems.

References

Bienias, J., Dippo, C., and Palmisano, M. (1987), Questionnaire Design: Report on the 1987 BLS Advisory Conference, Washington, DC: U.S. Department of Labor, Bureau of Labor Statistics.

Boehm, L. (1988), "CES Nonwage Cash Payment Prepilot Interviews," Internal memorandum to Alan Tupek dated December 16, Washington, DC: U.S. Department of Labor, Bureau of Labor Statistics.

Boehm, L. (1989), "The Relationship Between Confidence, Knowledge, and Performance in the Current Population Survey," in Proceedings of the Section on Survey Research Methods, American Statistical Association, in press.

Cannell, C., Fowler, F., Kalton, G., Oksenberg, L., and Bischoping, K. (1989), "New Quantitative Techniques for Pretesting Survey

Questions," in Bulletin of the International Statistical Institute, pp. 481-495.

Cole, M. and Scribner, S. (1974), Culture and Thought: A Psychological Introduction, New York: John Wiley and Sons.

Couper, M., Groves, R., and Jacobs, C. (1990, in press), "Building Predictive Models of CAPI Acceptance in a Field Interviewing Staff," in Proceedings of the 1990 Annual Research, Conference, Washington, DC: U.S. Department of Commerce, Bureau of the Census.

Cutler, S.J. and Grams, A.E. (1988), "Correlates of Self-Reported Everyday Memory Problems," Journal of Gerontology, 43, 582-590.

DeMaio, T. (1984), "Social Desirability and Survey Measurement: A Review," in Surveying subjective Phenomena, eds. C. Turner and E.

Martin, New York: Russell Sage.

Dippo, C.S. (1989), "The Use of Cognitive Laboratory Techniques for Investiating Memory Retrieval Errors in Retrospective Surveys, in Bulletin of the International Statistical Institute, Vol. LIII, Book 2, pp. 363-382.

Edwards, S., Levine R., and Allen, B. (1989), "Cognitive Strategies for Reporting Hours Worked, " in Proceedings of the Section on Survey Research Methods, American Statistical Association, in press.

Groves, R.M. and Cialdini, R. (1990), "Toward a Useful Theory of Survey Participation," unpublished manuscript.

Herrmann, D., van Melis-Wright, M., and Stone, D. (1990), "The Semantic Basis of Confidentiality," in Proceedings of the Section on Survey Methods Research, American Statistical Association, to appear.

Hertel, P. (1988), "External Memory," in M. Gruneberg, P. Morris,

and R. Sykes (eds.), Practical Aspects of Memory, New York: John Wiley and Sons.

Jabine, T., Straf, M., Tanur, J., and Tourangeau, R. (1984), Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines, Washington DC: National Academy Press.

Lessler, J., Salter, W., and Tourangeau, R. (1989). "Questionnaire Design in the Cognitive Research Laboratory: Results of an Experimental Prototype," Vital and Health Statistics, Series 6, No. 1 (DHHS Publication No. PHS 89-1076), Washington, DC: U.S. Government Printing Office.

Martin, E. (1987), "Some Conceptual Problems in the Current Population Survey," in Proceedings of the Section on Survey Methods Research, American Statistical Association, pp. 420-424.

McKay, R. (1990), "Application of Persuasive Communication Strategies to a Business Establishment Survey," in Proceedings of the section on Survey Methods Research, American Statistical

Association, to appear.

Miller, L. A. and Downes-LeGuin, T. (1989), "Reducing Response Error in Consumers' Reports of General Expenses: Application of Cognitive Theory to the Consumer Expenditure Interview Survey," Advances in Consumer Research, in press.

Mullin, P., (1990), "Proposal for Laboratory Research on the Feasibility of an Extended Interview Period for the CPS," unpublished memorandum to A. Tupek, in preparation.

Norwood, J. and Dippo, C. (in press), "Goverrment Applications," in Questions about Questions: Memory, Meaning and Social Interaction in Surveys, New York: Russell Sage.

Phipps, P. (1990), "Applying Cognitive Techniques to an Establishment Mail Survey," paper to be presented at the annual meeting of the American Statistical Association, Anaheim,

California, August.

Royce, J.R. (1973), "The Present Situation in Theoretical Psychology," in B.B Wolman (ed.), Handbook of General Psychology, Englewood Cliffs, NJ: Prentice Hall.

Shoemaker, H., Bushery, J., and Cahoon, L. (1989, in press), "Evaluation of the Use of CATI in the Current Population Survey," in Proceedings of the Section on Survey Research Methods, American Statistical Association.

Squire, L. (1987), Memory and Brain, New York: Oxford University Press.

Suchman, L. and Jordan, B. (1990), "Interactional Troubles in Face-to-Face Survey Interviews," Journal of the American Statistical Association, 85, 232-240.

Tanur, J. (1990, in press), "Reporting Job Search Among Youths: Preliminary Evidence from Reinterviews," in Proceedings of the 1990

Annual Research Conference, Washington, DC: U.S. Department of Commerce, Bureau of the Census.

Tourangeau, R. (1984), "Cognitive Sciences and Survey Methods," in Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines, T. Jabine, M. Straf, J. Tanur, and R. Tourangeau (eds.), Washington, DC: National Academy Press.

Tucker, C. and Bennett, C. (1988), "Procedureal Effects in the Collection of Consumer Expenditure Information: The Diary Operations Test," in Proceedings of the section on Survey Methods Research, American Statistical Association, pp. 256-261.

Tucker, C., Miller, L., Vitrano, F., and Doddy, J. (1989), "Cognitive Issues and Research on the Consumer Expenditure Diary Survey," paper presented at the annual American Association for Public opinion Research Conference.

Wolkowitz, O.M. and Weingattner, H. (1988), "Defining Cognitive Changes in Depression and Anxiety: A Psychobiological Analysis," Psychiatry Psychobiology, 3, 1-8.

Wright, P. (1980), "Strategy and Tactics in the Design of Forms,"

Visible Language, XIV 2, pp. 151-193.

THE ROLE OF A COGNITIVE LABORATORY IN A STATISTICAL AGENCY

Monroe G. Sirken

National Center for Health Statistics

Introduction

The statistical survey is an invention of the twentieth century. It produces a commodity, namely information, which many believe is the most important property in the modern world. Our Federal establishment, for example, would be unable to function nearly as effectively without the information being produced by surveys that are conducted by the Federal agencies represented at this Seminar. The Congressional and Executive branches use Federal surveys to monitor the nation's well-being, to evaluate the government's social, health and economic programs, and to plan legislation involving the collection of billions of tax dollars and the disbursement of billions of benefit dollars. Federal surveys could not have attained this level of acceptance and importance without the technological advances in survey methods that have occurred during the past half century. However, we can hardly afford to be complacent. As data producers, we are even more mindful than data consumers of the limitations of current survey technology. We

realize that further technological advances are essential to assure that Federal surveys will meet the growing needs for, more and better survey data.

There have been two major technological advances in survey methodology during the past 50 years and I believe a third may be in the offing. Each advance has introduced innovative technologies for improving the precision of the survey measurement process and was made possible by technology and theory transfers from the applied sciences. The "sampling" revolution in survey methodology that began in earnest during the 1930's came about as a result of technology transfers from the statistical sciences, and produced substantial advances in survey sampling and estimation methods. The "automation" revolution had its onset in the late 1960's. it came about as a result of technology transfers from the computer sciences, and has produced substantial advances in the methods of compiling and processing survey data. The "cognitive" revolution, which, as some of us believe got underway during the 1980's [Jabine, 1989], was made Possible by technology and concept transfers from the cognitive sciences. Whether called a revolution

or a movement, it has been introducing improved methods of designing data collection instruments and conducting questionnaire design research.

Federal Statistical agencies were major players in the "sampling" and "automation" revolutions in survey technology.: Now they are playing a major role in the "cognitive" movement by

developing and applying cognitive laboratory techniques to find better solutions to survey response problems. It is noteworthy that the cognitive movement is not confined to the U. S. government nor to the United States [Jobe and Mingay, 1991]. This paper, moreover, deals with only one part of the U.S. movement, namely, the work of the cognitive laboratory at the National Center for Health Statistics. The paper briefly describes the history and programs of the NCHS Laboratory and outlines the Laboratory's benefits to survey research, cognitive psychology, and Federal statistics.

History of the NCHS Laboratory

Until 1984, the role of cognition in the survey measurement process was largely ignored in the survey research programs of the National Center for Health Statistics. None of the earlier NCHS projects had been conducted in a cognitive laboratory, though one study [Laurent, Cannel and Marquis, 1972] used psychological theories to guide the development of interviewer and questionnaire techniques. Prior to 1984, survey response had been modeled as a two stage stimulus/response process with little attention paid to the effects that the respondents' mental processes had on the accuracy of their responses. In accordance with this psychological paradigm, survey research investigated the error effects of survey instruments and procedures almost exclusively in field tests. Since these field tests sought to replicate the actual conditions of the survey, they provided little opportunity to investigate cognitive issues, such as the following:

- What kinds of cognitive processing modes and strategies do respondents use in answering survey questions?

- How do the cognitive processing modes and strategies of survey respondents affect the accuracy of their responses to survey questions?

In 1984, with the support of an NSF grant, the NCHS embarked on a demonstration project that was motivated largely by the work of the Advanced Research Seminar on the Cognitive Aspects of Survey Methodology [Jabine, Straf and Tanur, 1984]. This project sought to demonstrate the utility of investigating the cognitive aspects of answering survey questions in a laboratory setting as a means of improving the design of Federal survey instruments [Sirken and Fuchsberg, 1984]. The project compared alternate versions of the dental supplement to the questionnaire of the 1986 National Health Interview survey.- One supplement was designed by the traditional field test method and the other by the proposed cognitive laboratory method [Lessler and Sirken, 1985].

The rationale for the demonstration project as expressed in the NSF grant proposal [Sirken, 1984] was:

"... because (1) questionnaire design is one of the weakest links in the survey measurement process, (2) past efforts to improve the quality of questionnaire have posed serious and difficult methodological problems, (3) the traditional field methods currently being used to improve questionnaire design are inadequate by themselves to handle many of these problems, and (4) complimentary methodologies that are not subject to the weakness of traditional field methods need to be developed, it is [therefore) essential to investigate the potential of using the (combined] techniques of the statistical and cognitive sciences in a laboratory setting as a complementary methodology for improving questionnaire design..."

The demonstration project was conducted in an interdisciplinary mode and in close collaboration with university

scientists so that, as the NSF grant proposal noted, another potential benefit was:

> "... it could go a long way in bridging the gap that exists between cognitive scientists academia and survey statisticians in Federal Statistical Agencies..."

This was critical to the ultimate success of the project because it was felt that gap between the disciplines had been largely responsible for the delay in applying cognitive methods in survey research.

At the successful conclusion of the demonstration project in 1986, NCHS established, with the support of a second NSF grant, the National Laboratory for Collaborative Research in Cognition and Survey Measurement. The National Laboratory's broad mission is to promote and advance interdisciplinary research on the cognitive aspects of survey methodology among Federal Statistical Agencies and the nation's universities and research centers. Interdisciplinary research with university scientists is promoted by a Collaborative Research Program which awards competitive

research contracts and appoints visiting scientists. Collaborative research with other Federal Agencies is promoted by the Questionnaire Design Research Laboratory which serves as the workplace for NCHS and other Federal Agencies to conduct intramural research [Royston, et al 1986]. The Collaborative Research Program has been largely funded by NSF grants and the Questionnaire Design Research Laboratory has been partially funded by reimbursable work agreements with other PHS Agencies [Sirken, et al 1990].

Activities of the NCHS Laboratory

Much of the work of the National Laboratory is based on a cognitive theory of survey response errors that can be stated as

follows: "survey respondents carry-out a series of mental tasks in the interval between being asked a survey question and providing a response. When these mental tasks pose serious mental burdens for

respondents they are likely to cause response errors." This view of the survey response process stimulated the development of cognitive methods for designing and pretesting questionnaire and for conducting questionnaire design research. Developing and testing survey instruments has short term objectives, namely, to detect and revise the design flaws before the survey instruments are field tested. in contrast, questionnaire design research objectives are long term, namely, to improve the designs of the next generation of survey instruments. These differences in objectives led to the development of distinctly different cognitive methods for developing and testing survey instruments and for conducting questionnaire design research.

Developing and Pretesting Questionnaires

The cognitive laboratory approach to developing and pretesting survey questionnaires is based on the premise that difficult, unreasonable or impossible the mental tasks implicit in some survey questions increase the likelihood of response errors. For example,

survey questions containing terms respondents do not understand, that are vague or ambiguous, that impose unrealistic demands on recall, that require complicated mental calculations, that contain too many elements for the respondent to think about simultaneously, that involve issues the respondent knows or cares little about, or that ask for embarrassing or threatening information-all impose cognitive burdens that are likely to result in invalid responses.

The realization that questionnaires obtain poor quality data when they ask respondents to perform difficult, if not impossible, mental tasks led to the development of a battery of laboratory techniques for investigating the cognitive burdens posed by survey questions [Bercini, In press, Royston, 1989] including think-aloud interviews, in-depth probing and focus group discussions, etc. These techniques are not new to questionnaire designers [DeMaio, 1983] but never before had they explicitly and systematically served as means of observing the manner in which respondents mentally process survey questionnaires and procedures.

Intensive interviewing techniques detect questionnaire design

flaws by observing the cognitive problems that result from these flaws. Poor questionnaire designs may impose difficult mental tasks at any cognitive stage of the response process including comprehending the questions recalling or estimating the information needed to answer the questions, and deciding whether or how to answer, the questions. Identifying the underlying cognitive difficulties experienced by respondents facilitates the process of revising the questionnaires appropriately.

Many questionnaire design problems detected and repaired by laboratory techniques are far less likely to be detected by traditional field testing methods. Consider the following question which was proposed for the National Health Interview Survey (NHIS), "During the past 12 months, have you been bothered by pain in your abdomen?" When laboratory respondents were asked this question, most answered it readily with a "Yes" or a "No". It was not until the laboratory interviewer probed into how respondents interpreted the term "abdomen" that it became apparent that respondents were

unsure of what section of the body to include. The interviews also determined that respondents had variable interpretations of the phrase, "bothered by," which in turn, affected whether they answered the question affirmatively or negatively. Intensive interviewing methods not only revealed that the question was apt to result in response errors, but also the underlying cause of the problem. When the cause of a question problem is understood, the solution is more likely to be found. In this case, part of the solution was a respondent flash card that showed an outline of the torso with the abdominal area shaded in.

Intensive interviews are conducted by laboratory trained questionnaire designers with many years of survey research experience. Paid subjects are recruited for the interviews. The topic and target populations of the survey determine the criteria for subject recruitment. Subjects are often selectively recruited to include those that would be most burdened by the survey questions or least successful in adopting effective mental strategies in answering the questions. Laboratory testing is usually carried out in interviewing waves of 5 to 10 subjects at a

time; the questionnaire is revised in consultation with the sponsor after each wave; and the testing continues until an acceptable version is obtained. Typically, flawed questions undergo 2-4 revisions before an acceptable version is ready for field testing. Field testing is essential in order to determine how the questionnaire will work under actual survey conditions. Additional laboratory testing may be needed to evaluate the questionnaire revisions that are suggested by the field test.

Depending on the complexity and scope of the questionnaire and on the number of conceptual problems associated with it, laboratory testing can be completed within several weeks or could span a longer period. For example, projects that involve special subject recruitment and testing may require a lead time of about six months or even longer. Also, laboratory projects are conducted collaboratively with survey sponsors and therefore involve frequent meetings to assure that the designed questionnaires satisfy the sponsors' research objectives.

Questionnaire Design Research

Cognitive methods of conducting questionnaire design research investigate why some survey questions and procedures pose cognitive tasks that are difficult, unreasonable or impossible for respondents to perform. In the same way that much has been learned in medicine by studying the cognitive aspects of amnesia and other memory disorders, so it is hoped that much can be learned in survey research by studying the cognitive aspects of questionnaires that' pose severe response burdens.

Questionnaire research seeks to improve the design of the next generation of survey questionnaires, especially those questionnaires dealing with topics for which better quality survey

data are needed. Causal relationships between the mental tasks performed by respondents and the accuracy of their responses are investigated in experiments. These experiment may be conducted in the cognitive laboratory or embedded in on-going surveys. The laboratory approach makes it possible to undertake many types of complex experiments that would be administratively impossible or prohibitively expensive to conduct as field experiments. Embedding cognitive experiments in on-going surveys makes it feasible to test laboratory findings under actual survey conditions.

Several features of cognitive laboratory experiments are noteworthy. They are interdisciplinary, involving the joint participation of cognitive psychologists and survey researchers. They generally involve testing questions that ask for the kinds of information that typically is poorly reported in surveys. They investigate those mental tasks implied by the survey questions that pose the greatest risks to accurate reporting. For example, if the question implied retrospective reporting, the focus would be on the cognitive aspects of the memory tasks and if the question asked for sensitive information the focus would be on the cognitive aspects of risk taking under conditions of uncertainty.

Generally, the subjects of laboratory experiments are recruited from population frames that contain information needed to, validate the experiment's findings. For example, the laboratory subjects for experiments on retrospective reporting of medical visits were selected from the files of a Health Maintenance Organization. Because the files provided access to the recruitment of subjects with known health conditions and doctor visit patterns [Means, et al, 1988]. Finally, the findings of the laboratory experiments are interpreted in terms of their potential contributions to cognitive theory as well as their implications for improving the design of survey instruments.

A recent project on dietary recall in nutrition surveys illustrates some of the benefits of conducting experiments in a cognitive laboratory. This complex multi-experiment project, involving randomization of subjects, diary keeping, and multiple

data collection sessions, could probably not have been undertaken as a traditional field experiment. The project investigated the cognitive burdens posed by the kinds of questions that are asked in household nutrition surveys [Smith, In press]. Generally these surveys collect dietary histories, food frequency inventories, and data on food portion sizes. Collecting these kinds of data imposes mental tasks involving free recall, frequency estimation, and magnitude estimation, respectively. Separate laboratory experiments were designed and conducted to assess the ability of respondents to provide accurate information on each of these tasks. The laboratory subjects participating in these experiments kept food diaries so their subsequent responses to dietary questionnaires could be validated.

For example, one of the nutrition survey experiments tested the effect of varying the portion size definitions on respondents' reports of the amount of food consumed. For each listed food item, respondents indicated whether their typical portion was small, medium or large in comparison with a defined medium portion size. Surprisingly, the food consumption reports in the experiment were

invariant to changes in the definition of medium portion size. These findings raise serious questions about the design of nutrition survey questionnaires and the quality of survey data on food consumption that are based oh portion size reports.

Over the past several years, laboratory experiments have investigated the cognitive factors involved in responding to difficult-to-answer questions on a variety of health related topics including utilization. of health services, cigarette smoking histories, illegal drug use, chronic pain episodes, and chronic disease prevalence.

A recent project on recall of doctor visit illustrates the benefits of embedding experiments in surveys. This split-ballot experiment was embedded in the pilot study of the National Medical Expenditure Survey. The experiment investigated the relative accuracy of retrospectively reporting doctor visits in a forward or in a backward temporal order [Jobe, et al, 1990]. It was suggested by the findings of previous laboratory experiments indicating that subjects varied in their preference between forward and backward

recall order but that backward recall seemed to produce more accurate reporting [Loftus, 1985].

The survey experiment assessed the accuracy of forward,, backward and free recall reporting strategies by comparing the medical visits reported by each strategy with the visits listed in medical records. The survey experiment did not confirm the findings of the laboratory experiments and showed little difference in accuracy between the alternative recall strategies. It was concluded that there was no evidence to suggest that survey instruments should be designed to favor either the forward, backward or free recall strategies.

Cognitive experiments involving survey material, whether conducted in laboratories or embedded in surveys, are valuable for several reasons. First, they provide in-depth knowledge about the cognitive processes respondents use in answering hard-to-answer survey questions. In particular, they often identify the kinds of question approaches that pose response burdens. And they suggest

methods of designing the questionnaires to reduce the response burdens and response errors. Secondly, because validation information is almost always collected (e.g., diaries, medical record matches, and biochemical markers) the response error effects of different questionnaire designs and cognitive strategies can be assessed. Third, the cognitive bounds on the abilities of respondents to perform specified kinds of mental tasks (comprehension, recall, etc.) posed by survey questions can be assessed.

Benefits of the NCHS Laboratory

The activities and programs of the NCHS cognitive laboratory during the past five years have benefitted survey research, cognitive science and Federal statistics in variety of ways. Some of the benefits are briefly outlined in these summary remarks.

Survey research has benefitted from the development of methods for investigating the cognitive aspects of the survey response

process. Intensive interviewing methods were perfected for designing and pretesting survey instruments in a laboratory setting, and experimental methods were perfected for conducting laboratory experiments and for embedding experiments in on-going surveys.

Cognitive science benefitted from the opportunities afforded its scientists by the NCHS laboratory to participate in the interdisciplinary research projects in cognition and survey measurement. Cognitive psychologists participating in these projects had opportunities to test cognitive theories with real world survey phenomena either in laboratory experiments or in experiments embedded in on-going surveys. And it is believed that the gains in cognitive psychology will ultimately benefit survey research and the quality of Federal surveys.

The activities of the NCHS laboratory fostered an appreciation and respect for the importance of conducting cognition and survey measurement research within and outside the Federal establishment. For example, the NCHS laboratory played a vital role in designing

and testing NCHS survey instruments during the past several years, and it is being viewed increasingly as a PHS laboratory with a mission to service the needs of agencies throughout the Public Health Service. As the first cognitive laboratory of its kind devoted to survey research, the NCHS laboratory served as a point of reference, if not the prototype, for the cognitive laboratories

that have since been established at other statistical agencies including the Bureau of the Census, Bureau of Labor Statistics and Statistics Sweden. Information dissemination has always been a high priority activity and during the past five yearsi the NCHS laboratory staff and collaborators published nearly 50 reports, and presented more than 100 papers at meetings and conferences.

Whether the existing movement in cognition and survey research, of which the NCHS laboratory is a part, will evolve into a full-fledged cognitive revolution with an impact equal to the sampling and automation revolutions remains to be determined. We

will know that. the cognitive revolution has, occurred when it becomes apparent that the cognitive sciences are providing scientific support to survey response research comparable to the support the statistical and computing sciences have been providing to research in survey sampling and in the automation of survey data.

References

Bercini, D.H. Presented at the EPA/AOWNA Symposium on Total Exposure Assessment Methodology. Pretesting Questionaire in the Laboratory: An Alternative Approach. In print. Toxicology and Industrial Health.

DeMaio, Theresa J. (Ed.) (1983). Approaches to Developing Questionnaires. Statistical Policy Working Paper 10. Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget. Washington, D.C.

Jabine, Thomas B. (1990). Cognitive Aspects of Questionnaire

Development. Presented at the EPA/AOWNA Symposium on Total Exposure Assessment Methodology. In print. Toxicology and Industrial Health.

Jabine, T.B., Straf, M.L., (1984). Tanur, J.M. and Tourangeau R. (Eds.). (1984). Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines. Washington, D.C. National Academy Press.

Jobe, J.B., White, A.A., Keileyi C.L., Mingay, D.J., Sanchez, M.J., and Loftus, E.F. (1990). Recall Strategies and Memory for Health Care Visits. Milbank Memorial Fund Quarterly/Health and Society, 68, 171-199.

Laurent, A.C., Cannell, C. and Marquis, K. (1972). Reporting Health Events in Household Interviews: Effects of an Extensive Questionnaire and Diary Procedure. Vital and Health Statistics, Series 2, No. 49 (DHHS Publication No. PHS 91-1079). Washington, D.C., U.S. Government Printing Office.

Lessler, J.T. and Sirken, M.G. (1985). Laboratory-Based Research on the Cognitive Aspects of Survey Methodology: The Goal of the National Center for Health Statistics Study. Milbank Memorial Fund Quarterly/Health and Society, 63, 565-581.

Loftus, E.F. and Fathi, D.C. (1985). Retrieving Multiple Autobiographical Memories. Social cognition. Vol. 3, pp. 280-95.

Royston, P.N. (1989). Using Intensive Interviews to Evaluate Questions. In F.J. Fowler, Jr. (Ed.), Heath Survey Research Methods (pp. 3-7) (DHHS Publication No. PHS 89-3447). Washington, D.C., U.S Goverrment Printing Office.

Royston, P.N., Bercini, D.H., Sirken, M.G. and Mingay, D. (1986). Questionnaire Design Research Laboratory. American Statistical Association, 1986 Proceedings of the Section on Survey Methods Research, pp. 703-707.

Sirken, Monroe G. (1986). National Laboratory for Collaborative Research on Cognition and Survey Measurement. Grant Proposal to the National Science Foundation. Washington D.C.

Sirken, Monroe G. (1984). Laboratory Based Research on the Cognitive Aspects of Survey Methodology. Grant Proposal to the National Science Foundation. Washington, D.C.

Sirken, M.G. and Fuchsberg R. (1984). Laboratory Based Research on the Cognitive Aspects of Survey Methodology. In Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciples. Washington, D.C. National Academy Press.

Smith, A.F. (in press). Cognitive Processes in Long-term Dietary Recall. Vital and Health Statistics, Series 6, No. 4 (DHHS Publication No. PHS 91- 1079). Washington, D.C., U.S. Government Printing office.

DISCUSSION

Elizabeth Martin

U.S. Bureau of the Census

In their two papers, Monroe Sirken of the National Center for Health Statistics, and Cathryn Dippo and Douglas Herrmann of the Bureau of Labor Statistics, document the activities of the cognitive laboratories which were established in 1984 and 1988, respectively, at their two agencies. The cognitive laboratories represent a commitment to survey data quality which is accredit to the two agencies. And Monroe Sirken and Cathryn Dippo, as two of the main instigators and initiators responsible for establishing

the laboratories, deserve credit and appreciation for their effort and accomplishment. The record of achievement by the two laboratories is a good one. Dippo and Herrmann organize their paper around a clear and comprehensive discussion of the sources of cognitive problems which can introduce errors in the response process; it id impressive how many of these problems have already been tackled in the BLS Collection Procedures Research Laboratory in its short history. Excellent research on a range of topics is also being conducted at the NCHS National Laboratory for Collaborative Research in Cognition and Survey Measurement, though in his paper Sirken does not actually describe the research. The NCHS lab lives up to the "collaborative" in its name; the number and caliber of academic researchers who have been involved in their projects are very high.

The growth of laboratory-based research on cognitive aspects of survey methodology is described by Dippo and Herrmann as a "movement" and by Sirken as a "revolution." These characterizations accurately reflect the enthusiasm and ferment of activity and new ideas in this area. However, "revolution" may not

be the most useful metaphor to describe how cognitive psychology is affecting (or, more importantly, should affect) survey research. In fact, the metaphor of "revolution" reflects and reinforces a weakness of the work currently going on in the new cognitive laboratories.

By emphasizing discontinuity with the past, researchers are, led to ignore relevant work which preceded many of the methods and ideas of the current "movement." Sirken characterizes survey research as (until recently) "based almost exclusively on the behaviorist paradigm" with "respondent's mental states... virtually ignored." This isn't accurate. Survey researchers, at least those practicing in academic or commercial settings, have hypothesized about and investigated psychological states intervening between survey questions and respondents' answers at least since World War II. (Jean Converse's Survey Research in the United States: Roots and Emergence, 1890-1960 provides a fascinating and useful history which traces the intellectual

origins of survey research.) Much of this work is still very relevant, and should be built on rather than ignored. For example, Dippo and Herrmann state that, "except for social desirability, the survey field is just beginning to investigate factors that affected communication of responses." They would benefit from reviewing the survey literature on the topic of communication, beginning with Herbert Hyman et al's comprehensive, Interviewing in Social Research, published in 1954. The methods used in the cognitive laboratories also have roots in the past. For example, Naomi D. Rothwell used very similar methods to conduct research on questionnaire design at the Census Bureau during the 1960s and 1970s. It is a bit of an overstatement for Sirken to claim in his paper to have invented the cognitive laboratory, without acknowledging similar, earlier activities.

In the field of survey research, there is a tradition of applying ideas from psychology to survey measurement issues. For the new work in the, cognitive laboratories to advance the state of the art of survey measurement, it should build on this tradition.

This would also increase its credibility to many survey researchers.

A danger of the "revolution" metaphor is it suggests a philosopy of "out with the old, in with the new." In some cases, this leads researchers to forget what they know about good survey practice. Compared to a survey, the cognitive laboratories generally rely on more intensive, less structured interviews with smaller numbers of respondents. This approach can be very informative about the nature and sources of cognitive errors in surveys. However, the "samples" usually are very small and not selected according to probability methods. one must be cautious drawing inferences from the results of most of the cognitive lab studies to date. For instance, I think Dippo and Herrmann are overstating the case when they conclude that, "research done at BLS shows clearly that proxy recall is different than self recall, both in terms of amount and kinds of information recalled." Laboratory findings such as this are more usefully thought of as hypotheses which should be subjected to more rigorous testing in a sample survey, and/or experimentally.

It is important to keep in mind that standards of evidence and proof still apply to research conducted in the cognitive laboratories. In some writings, the word "cognitive" is repeated so often as to suggest that the writer believes the word itself is sufficient to establish the merits of the research. But the researcher is still obliged to make his or her case on the evidence. For example, Sirken presents an example of a question on marijuana use which he says was improved by cognitive testing. How do we know it is better? He presents no evidence or logic to support his claim. In the long run, if the cognitive "movement" is to be taken seriously, it must demonstrate, not simply assert, the

value of its products, and be wary of the temptation to oversell itself.

I believe there are two common goals behind the activities in the cognitive laboratories. One goal is to improve particular

survey measurements. The second is to develop a theoretical foundation (beyond sampling theory) for improved survey design. The latter, broader aim requires that we develop better measures of nonsampling errors, and a better understanding of the effect of alternative survey designs on nonsampling errors. Methods and ideas from cognitive psychology are tools for achieving both specific and general improvements, but are not an end in themselves. Other social sciences (for example, social psychology) also have relevant knowledge to contribute.

With these goals (and the previously-stated cautions) in mind, what then is new and revolutionary about the work being done in the cognitive laboratories? First, this research has yielded new appreciation of the vulnerability of factual survey questions to biases and errors. I think it is fair to say that most government statisticians and academic survey methodologists probably have taken for granted the validity of simple factual questions. The research on problems of comprehension, recall and other cognitive difficulties is contributing to a more sophisticated understanding of how much we have yet to learn about the error properties of

survey measurements. Second, and more important, the research in the cognitive labs represents a new and more extensive set of methods for pretesting survey questionnaires and procedures. This in itself is a great leap forward. Traditionally, pretests of survey questionnaires have been ad hoc and informal, based on interviews with a few respondents and with no real guidelines beyond common sense to decide when one has succeeded or failed. The cognitive laboratories are changing that. Close and in-depth examination of Problems of respondent comprehension, recall, and' judgment, is shedding new light on the causes of these problems and (better yet) new ideas about how to correct them. The new methods which are being used and developed in the cognitive laboratories form a logical series of pretests prior to fielding a survey, proceeding from intensive, informal interviews, to small-scale experiments testing alternative questions or designs, to large-scale field experiments. In addition, as Cathryn Dippo points out in her remarks, testing can be integrated into the main survey itself, to provide ongoing information about nonsampling errors. The new methods thus make Possible a more scientific and, systematic approach to pretesting, and they promise to yield improvements in

the quality of data collected by the federal government.

DISCUSSION

Murray Aborn

National Science Foundation (retired)

I am grateful to my co-discussant, Elizabeth Martin of the Census Bureau, for providing the perfect lead-in to my own commentary on the papers presented at this session. Dr. Martin reminded us of the importance of viewing any disciplinary development from the perspective of its historical predecessors, and in this connection she succeeded in moving the advent of CASM

(Cognitive Aspects of Survey Methodology) -- writ large -- back several decades from the year most commonly cited as the date of its birth -- namely, 1980.

More consequential than revising our perception of the chronology of CASM (again writ large) is the difference Dr. Martin's remarks point up between the characterization of CASM in the paper presented by Cathryn Dippo and Douglass Herrmann of the Bureau of Labor Statistics, and the one presented by Monroe Sirken of the National Center for Health Statistics. Dr. Martin's remarks implicitly characterize CASM as a reawakening of old concerns, and thus place her in strong agreement with Dippols and Herrmann's labeling of CASM as a "movement," in contrast with Sirken's labeling of CASM as a methodological "revolution." Indeed, there is much to support the view of CASM as a movement; for instance, the enthusiasm of its adherents and the growing frequency with which its ideology is being endorsed by sectors of the statistical community and users of statistical data generally who have heretofore tended to ignore the psychosocial underpinnings of survey-taking (see, for example, Suchman and Jordan, 1990).

However, this does not mean that Sirken's description of CASM as representing a revolutionary development is totally incorrect. It may merely be premature, for the potential of CASM as a true breakthrough -- as a true revolution in survey research -- is clearly present in the programmatic and research agenda laid out for it in the seminal CASM document prepared by the National Academy of Sciences (see Jabine, et al, 1984). At the present time, only half the CASM prospectus is being actively pursued; namely, those objectives having to do with the adoption of certain recent advances in cognitive science into the survey design and instrumentation process. What we have seen little of to date is action on those objectives having to do with the use of surveys as naturalistic test beds for laboratory-based theories of the functioning of the neuronal mind and, ultimately, the emergence of a new paradigm for social/behavioral research in which survey-taking plays an important role in understanding such basic cognitive phenomena as how the brain stores memories and how mental imagery influences perception and recall, and in which developments in cognitive science relating to such branches of the field as

natural language semantics are used to produce greatly improved methods for achieving high-quality survey measurement. In other words, fulfillment of the "cognitive revolution" alluded to in Monroe Sirken's paper is clearly in prospect, but is yet to materialize.

I shall have a bit more to say on this subject at the close of my commentary; meanwhile, however, it is my opinion that much of the force behind Dr. Martin's view of CASM as a reawakening of old survey concerns -- as a "movement" more so than a "revolution" -- stems from the present truncated status of the programmatic agenda initially prescribed for the field. This gives CASM the appearance of a one-sided effort to adopt, in fairly superficial terms, some of the investigative techniques employed in recent laboratory-based cognitive psychology, and incorporate them in the conventional procedures for constructing and pretesting survey questionnaires. Under such a perspective, not much may appear to have been added to

what has long been known to be of influence in survey responding, and audiences such as the one attending the present session may rightfully feel that CASM amounts to little more than another real-life example of the familiar tale of "The Emperor's New Clothes" which, albeit a story-from the literature of childhood, embodies a profound adult theme concerning human gullibility and our tendency to accept uncritically what experts -- genuine and otherwise tell us is true, novel, or significant.

Now, let me examine the Emperor's New Clothes proposition against the CASM-engendered activities at the BLS and NCHS laboratories reported in the papers by Dippo and Herrmann and by Sirken. Reducing a sample of these activities to their most generic properties (in the sense of survey factors which, induce

response error), I would break them down into the following classification:

COLLECTION PROCEDURES RESEARCH LABORATOY (BLS)	QUESTIONNAIRE DESIGN LABORATORY (NCHS)
- Question Ambiguity (The extent to which a question may be interpreted in more than one way.)	- Question Wording and Order (The differential results induced by synonymous variation rearrangement of sequence.)
- Long-term Recall (The length of time over which the respondent is required to retrieve from memory.)	- Memorial Decay (The validity -- or veridicality -- of information supplied from short- and long-term memory.)

- Emotional Loading

(The degree of psychological stress which a question may place upon the respondent.)

- Affective Sensitivity

(The likelihood that a question may be embarrassing or impinge upon the respondent's privacy.)

- Subcultural Norms

(Question comprehensibility across ethnic subgroups.)

- Linguistic Complexity

(The effect of gramatical construction on the respondents ability to comprehend.)

- Social Desirability

(The extent to which a question is likely to elicit a normative rather than an idiopathic response.)

- Lexical Level

(The extent to which a question requires the respondent to have specialied -- in this case medical -- knowledge.)

Now, it is hard to believe that the many survey researchers trained in social psychology and cognate fields of social science

are oblivious to influences -- such as those charted above -- regardless of whether intellectual, technical, and/or cost factors make it impractical to subject such nonsampling sources of error to adequate control, or to estimate the proportion of total survey error due to their ubiquitous presence.

To take the phenomenon of Social Desirability, for example, it does not require a social scientist to comprehend the universal tendency of people represent a societally acceptable facade when questioned about attitudes and behavior. The popular press and many humorous books have for decades poked fun at surveys by ridiculing the informational value of asking such survey items as, "Do you bathe at least once a week?" or "Do you brush your teeth every day?"

To take some other examples, did it require CASM to alert survey researchers to the difference in results when a question is

phrased one way as opposed to another? Or to the difficulty of most respondents to deal with questions presented in grammatically complex form? Or to the impingement of certain areas of questioning on the sensitivity of respondents? Or to memorial decay overtime? Or to a respondent's understanding of questions embodying medical terminology?

I can't resist regaling the audience with a personal anecdote illustrating how ordinary, and even old-fashioned, if you will, is appreciation of the fact that few individuals not trained or highly educated in medicine cannot comprehend medical lexicography, and that one is apt to get ludicrous results from asking questions embodying medical terminology.

More than 25 years ago, when employed at the National Institute of General Medical Sciences, I shared an office with a public health epidemiologist who had just returned from a tour of duty in Puerto Rico. He told me of an effort to obtain data on the extent of interruption to normal life activities due to amoebic dysentery, which was then prevalent in most rural areas of Puerto Rico. Having never before conducted a survey, his group of public

health officials put together a series of questions utilizing such terms as diarrhea and defecation to get estimates of frequency. When the obtained results showed an average of only one to two bowel movements per day, the survey takers knew something was wrong and quickly realized that it was likely due to the language employed in identifying the disease.

The Public Health people reran a small subsample of respondents using the term "bowel movement" in the questionnaire, and obtained a slightly higher, but still medically incredible, estimate of frequency. Finally a native informant suggested that they phrase all questions pertaining to diarrhea in terms of La Mange or "The Curse" as it was known in the rural areas of the island and when they did this, the average reported frequency shot up to a more medically believable 11 or 12 occurrences per day.

If sheer knowledge of the fact that such variables as 1evel of lexical comprehension, differences in subcultural norms, and the tendency to respond in socially desirable ways are sources of error

in survey research, what, then, is it that is truly new about the CASM movement?

There are, to my mind, three major issues that have been brought to the fore by the CASM movement, coupled to the addition of new technical procedures which have proved powerful in cognitive research in psychology and artificial intelligence. And, as I have mentioned before and will emphasize at the close of my remarks, there is the potential for bringing about a truly interdisciplinary

effort to understand just what goes on in the interactional dynamics for survey and respondent.

The three major issues which have surfaced as a result of CASM are:

1. A reawakening of the essential conflict between survey questionnairing and ordinary conversation owing to the

need for artificially imposed standardized conditions of administration from the standpoint of survey statistics on the one hand, and the natural world existence of individual differences in mentality on the other.

2. The extent to which laboratory-based treatments and results can be transferred to the field in the case of survey-taking. This issue is of general importance to social science, as well as being particularly relevant to survey research insofar as the laboratory setting, which provides greater conditions of control and flexibility, creates possibilities for a more systematic approach to instrumentation, and hence to survey measurement.

3. The degree to which the contemporary shift in the underlying paradigm of survey research's cognate substantive discipline -- i.e., psychology -- requires a realignment away from behaviorism and toward cognition. CASM represents a bold attempt to test this issue and assay its yield, but there has thus far been far too

little involvement of cognitive psychology per se apart from the importation of certain investigative techniques.

I by no means wish to detract from the accomplishments reported in the papers by Dippo and Herrmann and by Sirken based upon the importation of the techniques employed in contemporary cognitive psychology, into the innovative laboratory facilities now ensconced in such two prestigious governmental agencies as BLS and NCHS. Much thought and expertise have been applied to the transfer of technology represented, by the successful adoption of such cognitive probes and methods as: (1) Focus Groups; (2) Part-set Cueing; (3) Protocol Analysis; and (4) Think-aloud procedure.

But in my opinion, this could be just the beginning of a truly revolutionary development in survey research and, through its influence, on social science more broadly. The laboratory-based techniques and procedures you have heard presented at this session are derived from research begun in the early 1960's by Nobel Laureate Herbert Simon and Alan Newell that resulted in the General Problem Solver and led, to the foundations of the field of artificial intelligence (Barr and Feigenbaum, 1982). The more

recent work of Simon (Simon, 1987), shows the even greater, potential of cognitive technology to uncover human information processing systems.

However, there is reason to be both pessimistic and optimistic about the future of CASM. On the one hand, the statistical framework of survey research -- the dominant framework for the field -- is concerned with drawing inferences about populations -- about whether the sample of a population is large and representative enough to permit accurate and valid conclusions to be reached about the distribution of characteristics in the population from which the survey sample was drawn. On the other hand, the cognitive framework is concerned with drawing accurate and valid inferences about individuals about respondent "truthfulness," if you will.

Therefore, one framework calls for instrumentation designed to enhance person-to-person comparability, while the other calls for

instrumentation designed to enhance the assessment of person-to-person variations on each survey variable.

It is the work of the two survey/cognitive research laboratories reporting here today that represents one of the two reasons I find to be optimistic about the future of CASM. Such facilities offer the best opportunities for reconciling the conflicting survey conceptual frameworks described above.

The other reason I find to be optimistic lies in the pronouncement appearing in a neuropsychological book which has become a national bestseller in addition to its importance to the scientific literature on brain-behavior relationships. I refer to -- and endorse to you as top-quality literature as well as a work of cognitive science importance -- Oliver Sacks' The Man Who Mistook His Wife for a Hat. I close my remarks by quoting from a passage in this work that, I believe, should stimulate cognitive scientists to become fuller participants in CASM, recognizing that survey centers and facilities are ideally suited to cognitive explorations and offer the prospect of a vital new interdisdipline.

After presenting and analyzing the case of The Man Who Mistook His Wife for a Hat, Sacks concludes, as I do here, that:

> cognitive sciencesiare themselves suffering from an agnosi similar to the one afflicting the man who mistook his wife for a hat. That man may thus serve as a warning and parable of what happens to a science which eschews the judgmental, the particular, the personal, and becomes entirely abstract and computational (Sacks, 1987, p. 20).

I hope that cognitive psychologists will take heed of Dr. Sacks' warning and see the opportunity that survey research offers to offset the present trend toward abstract computationalism.

References

1. Suchman, L. and Jordan, B., "Interactional Troubles in Face-to-Face Survey Interviews," JASA, Vol. 85, No. 409, pp. 232-253, 1990.

2. Jabine, T., Straf, M., Tanur, J., and Tourangeau, R. (eds). Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines, Washington, D.C.: National Academy Press, 1984.

3. Barr, A. and Feigenbaum, E.A., (eds.) The Handbook of Artificial Intelligence. Stanford, CA: Heuristech Press, 2:184-192, 1982.

4. Simon, H., "The Steam Engine and The Computer: , What Makes Technology Revolutionary," EDUCOM Bulletin, 22(l):2-5, 1987.

5. Sacks, O., The Man Who Mistook His Wife For A Hat, New York: Harper and Row, p. 20, 1987.

Reports Available in the Statistical Policy Working Paper Series

1. Report on Statistics for Allocation of Funds (Available through NTIS Document Sales, PB86-211521/AS)
2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques (NTIS Document Sales, PB86-211539/AS)
3. An Error Profile: Employment as Measured by the Current Population Survey (NTIS Document Sales PB86-214269/AS)

4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (NTIS Document Sales, PB86-211547/AS)

5. Report on Exact and Statistical Matching Techniques (NTIS Document Sales, PB86-215829/AS)

6. Report on Statistical Uses of Administrative Records (NTIS Document Sales, PB86-214285/AS)

7. An Interagency Review of tizie-Series Revision Policies (NTIS Document Sales, PB86-232451/AS)

8. Statistical Interagency Agreements (NTIS Documents Sales, PB86-230570/AS)

9. Contracting for Surveys (NTIS Documents Sales, PB83-233148)

10. Approaches to Developing Questionnaires (NTIS Document Sales, PB84-105055/AS)

11. A Review of Industry Coding Systems (NTIS Document Sales, PB84-135276)

12. The Role of Telephone Data Collection in Federal Statistics (NTIS Document Sales, PB85-105971)

13. Federal Longitudinal Surveys (NTIS Documents Sales, PB86-139730)

14. Workshop on Statistical Uses of Microcomputers in Federal Agencies (NTIS Document Sales, PB87-166393)

15. Quality in Establishment Surveys (NTIS Document Sales, PB88-232921)

16. A Comparative Study of Reporting Units in Selected Employer Data Systems (NTIS Document Sales, PB-90-205238)

17. Survey Coverage (NTIS Document Sales, PB90-205246)

18. Data Editing in Federal Statistical Agencies (NTIS Document Sales, PB90-205253)

19. Computer Assisted Survey Information Collection (NTIS Document Sales, PB90-205261)

20. Seminar on the Quality of Federal Data (NTIS Document Sales, PB91-142414)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161 (703) 487-4650

"1"David A.Pierce is Senior Statistician, Micro Statistics Section, Division of Research and Statistics, Federal Reserve

Board, Washington, DC 20551, and a member of the Federal Committee on Statistical Methodology and its Subcomittee on Data Editing in Federal Statistical Agencies. Any views expressed do not necessarily reflect those of the Federal Reserve System.

2 The sampling design in the original CATI sample was stratified simple random sampling. The reinterview sample was a random sample of CATI respondents within strata. The bias was approximated by expanding the difference in reconciled and CATI response at the sample unit level.

Statistical Policy Working Paper 20 - Seminar on Quality of Federal Data - Part 3 of 3

Statistical Policy

Working Paper 20

Seminar on Quality of Federal Data

Part 3 of 3

Federal Committee on Statistical Methodology

Statistical Policy Office

Office of Information and Regulatory Affairs

Office of Management and Budget

March 1991

MEMBERS OF THE FEDERAL COMMITTEE ON STATISTICAL METHODOLOGY

(February 1991)

Maria E. Gonzalez, Chair

office of Management and Budget

Yvonne M. Bishop

Energy Information Administration

Warren L. Buckler

Social Security Administration

Charles E. Caudill

National Agricultural Statistics Service

Daniel Kasprzyk

Bureau of the Census

Daniel Melnick

National Science Foundation

Robert P. Parker

Bureau of Economic Analysis

David A. Pierce

Federal Reserve Board

Cynthia Z.F. Clark

National Agricultural Statistics Service

Thomas J. Plewes

Bureau of Labor Statistics

Zahava D. Doering

Smithsonian Institution

Wesley L. Schaible

Bureau of Labor Statistics

Robert M. Groves

Bureau of the Census

Fritz J. Scheuren

Internal Revenue Service

Roger A. Herriot

National Center for Education Statistics

Monroe G. Sirken

National Center for Health Statistics

C. Terry Ireland

National Computer Security Center

Robert D. Tortora

Bureau of the Census

Charles D. Jones

Bureau of the Census

PREFACE

In 1975, the Office of Management and Budget (OMB) organized the Federal Committee on Statistical Methodology. Comprised of individuals selected by OMB for their expertise and interest in statistical methods, the committee has during the past 15 years determined areas that merit investigation and discussion, and overseen the, work of subcommittees organized to study particular issues. Since 1978, 19 Statistical Policy Working Papers have been published under the auspices of the Committee.

On May 23-24, 1990, the Council of Professional Associations on Federal Statistics (COPAFS) hosted a "Seminar on the Quality of Federal Data." Developed to capitalize on work undertaken during the past dozen years by the Federal Committee on Statistical Methodology and its subcommittees, the seminar focused on a variety of topics that have been explored thus far in the Statistical Policy Working Paper series. The subjects covered at the seminar included:

Survey Quality Profiles

Paradigm Shifts Using Administrative Records

Survey Coverage Evaluation

Telephone Data Collection

Data Editing

Computer Assisted Statistical Surveys

Quality in Business Surveys

Cognitive Laboratories

Employer Reporting Unit Match Study

Approaches to Developing Questionnaires

Statistical Disclosure-Avoidance

Federal Longitudinal Surveys

Each of these topics was presented in a two-hour session that featured formal papers and discussion, followed by informal dialogue among all speakers and Attendees.

Statistical Policy Working Paper 20, published in three parts, presents the proceedings of the "Seminar on the Quality of Federal

Data." In addition to providing the papers and formal discussions from each of the twelve sessions, this working paper includes Robert M. Groves' keynote address, "Towards Quality in a Working Paper Series on Quality," and comments by Stephen E. Fienberg, Margaret E. Martin, and Hermann Habermann at the closing session, "Towards an Agenda for the Future."

We are indebted to all of our colleagues who assisted in organizing the seminar, and to the many individuals who not only presented papers and discussions but also prepared these materials for publication. A special thanks is due to Terry Ireland and his staff for their work in assembling this working paper.

Table of Contents

Wednesday, May 23, 1990

Part 1

KEYNOTE ADDRESS

TOWARDS QUALITY IN A WORKING PAPER SERIES ON QUALITY 3

Robert M. Groves, The University of Michigan and U. S.

Bureau of the Census

Session 1 - SURVEY QUALITY PROFILES

THE SIPP QUALITY PROFILE 19

Thomas B. Jabine, Statistical Consultant

INITIAL REPORT ON THE QUALITY OF AGRICULTURAL SURVEY PROGRAM. . . . 29

George A. Hanuschak, National Agricultural Statistics

service

DISCUSSION 40

Barbara A. Bailar, American Statistical Association

DISCUSSION 46

Nancy A. Mathiowetz, U. S. Bureau of the Census

Session 2 - PARADIGM SHIFTS USING ADMINISTRATIVE

RECORDS

PARADIGM SHIFTS: ADMINISTRATIVE RECORDS AND CENSUS-TAKING 53

Fritz Scheuren, Internal Revenue Service

AN ADMINISTRATIVE RECORD PARADIGM: A CANADIAN EXPERIENCE. 66

John Leyes, Statistics Canada

DISCUSSION. 77

Gerald Gates, U.S. Bureau of the Census

DISCUSSION 83

Edward J. Spar, Market Statistics

Session 3 - SURVEY COVERAGE EVALUATION

CONTROL MEASUREMENT, AND IMPROVEMENT OF SURVEY COVERAGE 87

Gary M. Shapiro, U. S. Bureau of the Census; Raymond R.

Bosecker, National Agricultural Statistics Service

QUALITY OF SURVEY FRAMES 100

Judith T. Lessler, Research Triangle Institute

DISCUSSION 108

Fritz Scheuren, Internal Revenue Service

DISCUSSION 114

Joseph Waksberg, Westat, Inc.

Session 4 - TELEPHONE DATA COLLECTION

QUALITY IMPROVEMENT IN TELEPHONE SURVEYS 123

Leyla Mohadjer, David Morganstein, Westat, Inc.

COMPUTER ASSISTED SURVEY TECHNOLOGIES IN GOVERNMENT:

AN OVERVIEW 137

Marc Tosiano, National Agricultural Statistics Service

DISCUSSION. 155

William L. Nicholls II, U. S. Bureau of the Census

DISCUSSION. 161

James T. Massey National Center Health Statistics

Part 2

Session 5 - DATA EDITING

OVERVIEW OF DATA EDITING IN FEDERAL STATISTICAL AGENCIES 167

David A. Pierce, Federal Reserve Board

EDITING SOFTWARE (An excerpt from Chapter IV of Working

Paper 18) 173

Mark Pierzchala, National Agricultural Statistics

Service

RESEARCH ON EDITING 180

Yahia Ahmed, Internal Revenue Service

DISCUSSION 184

Charles E. Caudill, National Agricultural Statistics

Service

DISCUSSION 186

Richard Bolstein, George Mason University

Session 6 - COMPUTER ASSISTED STATISTICAL

SURVEYS

OVERVIEW OF COMPUTER ASSISTED SURVEY INFORMATION COLLECTION 191

Richard L. Clayton, U. S. Bureau of Labor Statistics

A COMPARISON BETWEEN CATI AND CAPI 197

Martin Baum, National Center for Health Statistics

COMPUTER ASSISTED SELF INTERVIEWING 202

Ralph Gillmann, Energy Information Administration

COMPUTER ASSISTED SELF INTERVIEWING: RIGS AND PEDRO,

TWO EXAMPLES. 205

Ann M. Ducca, Energy Information Administration

DATA COLLECTION 209

Cathy Mazur, National Agricultural Statistics Service

DISCUSSION. 212

Robert N. Tinari, U. S. Bureau of the Census

DISCUSSION 216

David Morganstein, Westat, Inc.

Thursday, May 24, 1990

Session 7 - QUALITY IN BUSINESS SURVEYS

IMPROVING ESTABLISHMENT SURVEYS AT THE BUREAU OF LABOR

STATISTICS 221

Brian MacDonald, Alan R. Tupek, U. S. Bureau of Labor

Statistics

A REVIEW OF NONSAMPLING ERRORS IN FEDERAL ESTABLISHMENT

SURVEYS WITH SOME AGRIBUSINESS EXAMPLES 232

Ron Fecso, National Agricultural Statistics Service

DISCUSSION 243

David A. Binder, Statistics Canada

DISCUSSION. 247

Charles D. Cowan, Opinion Research Corporation

Session 8 - COGNITIVE LABORATORIES

THE BUREAU OF LABOR STATISTICS' COLLECTION PROCEDURES

RESEARCH LABORATORY: ACCOMPLISHMENTS AND FUTURE DIRECTIONS 253

Cathryn S. Dippo, Douglas Herrmann, U. S. Bureau of Labor

Statistics

THE ROLE OF A COGNITIVE LABORATORY IN A STATISTICAL AGENCY 268

Monroe G. Sirken, National Center for Health Statistics

DISCUSSION. 278

Elizabeth Martin U. S. Bureau of the Census

DISCUSSION 281

Murray Aborn, National Science Foundation (retired)

Part 3

Session 9 - EMPLOYER REPORTING UNIT MATCH

STUDY

INTERAGENCY AGREEMENTS FOR MICRODATA ACCESS:

THE ERUMS EXPERIENCE 291

Thomas B. Petska, Internal Revenue Service; Lois

Alexander, Social Security Administration

SAMPLE SELECTION AND MATCHING PROCEDURES USED IN ERUMS 301

John Pinkos, Kenneth LeVasseur, Marlene Einstein,

U. S. Bureau of Labor Statistics; Joel Packman, Social

Security Administration

RESULTS, FINDINGS AND RECOMMENDATIONS OF THE ERUMS PROJECT 309

Vern Renshaw, Bureau of Economic Analysis; Tom Jabine,

Statistical Consultant

DISCUSSION 318

W. Joel Richardson, Charles A. Waite, U. S. Bureau of the

Census

DISCUSSION 324

Thomas J. Plewes, U. S. Bureau of Labor Statistics

Session 10 - APPROACHES TO DEVELOPING

QUESTIONNAIRES

TOOLS FOR USE IN DEVELOPING QUESTIONS AND TESTING

QUESTIONNAIRES 331

Theresa J. DeMaio, U. S. Bureau of the Census

TECHNIQUES FOR EVALUATING THE QUESTIONNAIRE DRAFT 340

Deborah H. Bercini, National Center for Health Statistics

DESIGNING QUESTIONNAIRES FOR CATI IN A MIXED MODE

ENVIRONMENT. 349

Gemma Furno, U. S. Bureau of the Census

DISCUSSION 360

Carol C. House, National Agricultural Statistics Service

Session 1 1 - STATISTICAL DISCLOSURE - AVOIDANCE

DISCLOSURE AVOIDANCE PRACTICES AT THE CENSUS BUREAU 367

Brian Greenberg, U. S. Bureau of the Census

THE MICRODATA RELEASE PROGRAM OF THE NATIONAL CENTER

FOR HEALTH STATISTICS 377

Robert H. Mugge, National Center for Health Statistics

(retired)

DISCUSSION 385

George T. Duncan, Carnegie Mellon University

Session 12 - FEDERAL LONGITUDINAL SURVEYS

FEDERAL LONGITUDINAL SURVEYS 393

Daniel Kasprzyk, U. S. Bureau of the Census; Curtis Jacobs, U. S. Bureau of Labor Statistics

THE ADVANTAGES AND DISADVANTAGES OF LONGITUDINAL SURVEYS 407

Robert W. Pearson, Social Science Research Council

LONGITUDINAL ANALYSIS OF FEDERAL SURVEY DATA 425

Patricia Ruggles, Joint Economic Committee

DISCUSSION 438

Michael Brick, Westat, Inc.

DISCUSSION. 447

Marilyn E. Manser, U. S. Bureau of Labor Statistics

TOWARDS AN AGENDA FOR THE FUTURE

Stephen E. Fienberg, Carnegie Mellon University 455

Margaret E. Martin 462

Hermann Habermann, Office of Management and Budget 465

Part 3

Session 9

EMPLOYER REPORTING

UNIT MATCH STUDY

INTERAGENCY AGREEMENTS FOR MICRODATA ACCESS:

THE ERUMS EXPERIENCE

Thomas B. Petska

Internal Revenue Service

Lois Alexander

Social Security Administration

The Employer Reporting Unit Match Study (ERUMS) was a pilot record linkage study carried out under the auspices of the Federal Committee on Statistical Methodology of the Office of Management and Budget. The study linked records of employers and their reporting units from three agencies: the Bureau of Labor Statistics (BLS), the Social Security Administration (SSA) and the Internal Revenue Service (IRS). The primary linkages involved samples of the agencies' records for employers in the State of Texas covering their-activities in 1982.

For the ERUMS Workgroup to gain access to the data sets needed for the study, arrangements had to be developed that would comply with the confidentiality provisions and statutes of the Federal and State agencies that controlled these data sets. This paper gives an overview of these arrangements and agreements. In the first section, background information on the statistical content and confidentiality provisions of each of the data sets is provided. In the second section, the actual arrangements for the release of

confidential microdata are described. The last section provides a summary of what we have learned about such data sharing arrangements.

Background Information

The goal of ERUMS was to demonstrate the feasibility of matching employer and reporting unit data from different agency record systems as a means of obtaining more precise information about the coverage and content of the data in those systems. A purpose was to examine and I evaluate differences in wage and employment data at the state and county level as reported to those agencies. Despite the many difficulties encountered in establishing the data access agreements, ERUMS demonstrated that data such sharing Projects can be successful under current laws.

1. Data Sets

The ERUMS study was a three-way data linkage study in which individual microdata records from BLS, SSA, and IRS were matched by Employer Identification Number (EIN).

a. BLS provided a 1982 Unemployment Insurance (UI) Address File, which, for each state, consists of data for individual employers and their reporting units, which are often equivalent to "establishments". The data for this file are submitted to BLS by the State employment security agencies that operate the Federal-State UI Program. BLS uses the data submitted by the states as a basis for statistical reports on employment and wages and uses the UI Address File as a national sampling frame for its establishment surveys.

b. SSA provided an edited file of Form W-3 annual reports for 1982 and the Single Unit and Multi-Unit Code Files.

The Form W-3 file provided data on individual employers and, in some cases, for each of their reporting units, which are frequently equivalent to establishments. The Single Unit Code File contains a record for most entities that have filed an application for an Employer Identification Number. The Multi-Unit Code File contains a record for each reporting Unit of multi-unit employers who are participating in the Establishment Reporting Plan, a voluntary program under which employers report wage information on Form W-3 separately for each of their reporting units.

c. IRS data used for ERUMS were from a Census-edited file based on Forms 941 and 943 for Tax Years 1981-83. These forms are used by employers to report each quarter (annually for Form 943) to IRS on income taxes withheld from wages and other payments to employees and on taxes under the Federal Insurance Contributions Act (FICA) under the Social Security system. Extracts of data from these forms are provided annually by IRS to the Census Bureau for use in the latter's County Business Patterns

Program and other statistical programs. The Census Bureau edits the files, particularly the industry codes, and imputes certain missing data. This file was made available to the IRS Statistics of Income (SOI) Division for use in its business employment and payroll studies and was used for ERUMS. In addition, copies of Form 940, Federal Unemployment Tax Return, were obtained for a substantial proportion of the ERUMS sample cases.

2. Data Sharing Issues

For the ERUMS Workgroup to gain access to the data sets needed for the study, it was necessary to develop working arrangements that complied with the provisions of confidentiality statutes, regulations, and policies of the Federal and State agencies that controlled these data sets.

Although interagency exchange of identifiable microdata was the key to ERUMS, such data sharing is restricted by Federal confidentiality laws which generally permit agencies to disclose statistical information only in summary or other unidentifiable form. Since ERUMS was designed to link and compare information about individual employers collected separately by the different agencies, the Workgroup had to develop and implement lawful methods of transferring data on identifiable business units among the participants. A related task was to minimize the disclosure of identifiers in making those transfers and linkages.

The Workgroup was particularly interested in the different ways an employer may report establishment or multi-unit enterprise data to various State and Federal agencies. To examine these differences, the Workgroup needed to compare employers' reports to the BLS State UI programs, the SSA FICA reporting, and the IRS employment tax returns. Members Of the Workgroup included employees of these agencies, plus employees of the Bureau of

Economic Analysis, Office of Management and Budget, the Bureau of the Census, and the Committee on National Statistics of the National Academy of Sciences.

The Workgroup planned to analyze the information that corresponded to each EIN as it was reported to each agency. The analysis and findings would be entirely statistical in nature with no reference to the individual (identifiable) cases. Nevertheless, the planning, processing, and analysis phases each required access to identifiable data.

3. Confidentiality of Federal and State Tax Records

In the ERUMS study, the Employer, Identification Number (EIN) was the identifier that was common to all the reporting systems. It was used to define the sample drawn by BLS and was used as the basis for retrieving, linking and comparing records containing information from the SSA and IRS files. By law, the EIN is a tax identification number, and even when standing alone is protected by Internal Revenue Code confidentiality restrictions.

ERUMS required access to data from W-3 records which by law are Federal tax records that are processed and maintained at SSA in conjunction with the computation of Social Security retirement benefits. Since these are tax records, it was necessary to satisfy IRS that the selection by SSA of sample cases, SSA's disclosure of W-3 data to BLS, and the use of employer data by other members of the Workgroup met the requirements of the Internal Revenue Code dealing with disclosure of tax information. (See No. 4 below.)

BLS selected Texas as the State whose records it would sample, and it obtained written permission from the Texas State Employment Security Agency to use their UI records in the project. The Texas

Unemployment Compensation Act requires Texas employers to maintain records and file reports to the Texas Employment Commission with detailed information about the business operations and the number and compensation of employees. Texas law prohibits disclosure except for administering the Act, and it makes improper disclosure

punishable by fines or imprisonment.

4. Other Confidentiality Considerations

Since the Workgroup was composed of employees from several agencies and organizations, confidentiality laws did not apply to them uniformly. In varying degrees, certain laws, regulations, and policies affected each agency's access to identifiable records from particular sources and provided differential access to various individuals in the Workgroup. A recurring theme was the necessity at each phase of the process to identify the persons who needed to use identifiable data and to ensure that no others had access at that time.

Besides affidavits and other written procedures to protect the confidentiality of records, certain technical safeguards were adopted to minimize disclosure risk. The first of these methods was to avoid identifying sample cases by EIN to persons who performed processing in the participating agencies but were not

directly associated with the Workgroup. This method was adopted to conform to, the Internal Revenue Code requirements for tax information under the agreement BLS had with the State of Texas. At BLS this led to a decision not to process the data on the mainframe computer system at the Department of Labor that is operated by a private contractor. Instead, BLS used a mini-computer which was accessible only to BLS employees who were members of the Workgroup.

State agencies periodically submit to BLS UI address files that compile identification data for all reporting units at the most-detailed level that is available from employers' reports. BLS compiles these reports under a pledge of confidentiality that allows the data to be used only by authorized persons for statistical purposes.

Once BLS selected the Texas sample, it had to create a finder list so that SSA could extract corresponding records from its W-3 and related files for employers in the sample. The technical staff who performed these operations at SSA have routine access in their

usual jobs to the employer records maintained at SSA. However, they did not need to know which of the employers' records comprised the sample selected by BLS from the Texas UI file. To avoid identifying those cases that were actually in sample, furnished SSA with a listing of 7 of the 9 digits of sample EINS. SSA staff then extracted records from the W-3 and related files for all records in which these 7 digits appeared without knowing which

employers were actually in the BLS sample. This procedure effectively masked the identities of sample cases derived from State UI files, and thus significantly limited the number of SSA employees who were required to sign BLS non-disclosure affidavits.

Agreements for Interagency Data Sharing

Access by the Workgroup to the data sets needed for the study

was accomplished through three interagency agreements plus an additional access arrangement.

The Workgroup had originally planned a tripartite arrangement through interagency agreements of SSA and BLS with IRS. However, IRS counsel raised objections that such a multi-party agreement would be unduly cumbersome, and approval would probably not be forthcoming. As an alternative, IRS proposed to contract exclusively with BLS for the performance by BLS of services that required access to tax data. SSA staff would be designated as special agents of BLS to process the data. Bilateral BLS/IRS and BLS/SSA agreements would also have to be drafted under this arrangement.

The drafting of these arrangements proved to be a delicate task. By law, the purposes of IRS participation in the project and its service contract with BLS had to be related to IRS administration of the tax laws. Section 6103(n) of the Internal Revenue Code (IRC) allows IRS to disclose tax return information to persons outside of the agency as long as it is for purposes of tax administration [1]. Specifically, this purpose is to conduct

statistical studies based on return information, which Section 6108(a) of the IRC authorizes IRS to perform [2]. A case was made that the ERUMS study was one such purpose

1. BLS and Texas Agreement

BLS has cooperative agreements with 50 State Employment Security Agencies to use employment statistics collected by the states for its labor economics research. The 1982 data used in the ERUMS study was furnished to BLS in its ES-202 program by the Texas State Employment Commission under a cooperative agreement. It was necessary for BLS to obtain authorization from the State Commission to use the microdata for the ERUMS study and to provide access for the Workgroup members. Under this cooperative agreement, the access and use of the data were subject to the confidentiality requirements of the Texas Employment Compensation statute as well as those set out in the BLS Commissioner's Order No. 2-80.

Each UI program is operated under state law that must conform

to certain minimum federal standards, with reports that enable BLS to monitor state compliance. Under the Texas program, each

employing unit is required to file (and update periodically) a status report with the Texas Employment Commission, describing the type of ownership, location, and nature of business. On a quarterly basis, employers are required to file detailed reports on wages and contributions. Multi-Unit employers are asked to file a voluntary statistical supplement that provides detailed employment, wage, and contribution reports for each establishment. The ES-202 reports are compiled by BLS and form the basis for the UI Address file that BLS maintains. This is a micro-level employer file that contains first quarter information for each reporting unit, and the 1982 file provided the Texas sampling frame for the ERUMS sample.

The confidentiality of statistical data collected under the cooperative agreement is protected by interrelated state and federal procedures. At the state level, these UI reports are

collected under the Texas Unemployment Compensation Act which limits the availability of its UI reports to public employees in the performance of public duties, except, as the Employment Commission may find necessary in its administration of Texas law. At the federal level, BLS receives and maintains these confidential reports under the authority of the BLS Commissioner's Order that pledges confidentiality and prohibits disclosure except to authorized persons for statistical purposes. This Order precludes any use of identifiable information for non-statistical purposes, such as investigation or enforcement.

Under this cooperative agreement with the State of Texas, it was necessary for BLS to obtain permission from the Texas Commissioner to select employer sample cases and to make information about them available to BLS and SSA employees in the ERUMS Workgroup and later to others in the Microdata Access Group. In Addition, BLS procedures establish the confidentiality of the identities and all information pertaining to employers in the sample. Members of the Workgroup who were not BLS employees were appointed as BLS agents pursuant to another interagency agreement

with BLS. Like BLS employees, other Workgroup members were required to sign a Non-Disclosure Affidavit before they would be given access to the microdata.

2. IRS And BLS Agreement

The initial draft of the statement of purpose by IRS representatives was, acceptable to IRS counsel since its justification for sharing of confidential tax information was defined as for purposes of tax administration, which is permissible under section 6103(n) of the Internal Revenue Code [1]. However, the case that was made for IRS tax administration purposes was not acceptable to other Workgroup participants because they felt that this did not clearly describe the purposes of the ERUMS project in general or SSA's role in particular. In the, subsequent draft, care was taken to define contractual purposes in language that covered

the statistical purposes of the several participating agencies and

that provided for the exchange of records to create a common pool of data for a variety of analytical purposes, including those related to tax administration.

In this agreement, IRS contracted with BLS for the performance of those parts of the ERUMS project that required access to tax data, including the wage report information that was to be provided by SSA. Under this agreement, SSA staff could be designated as special agents of BLS to carry out their part of the linkage and analysis operations. By law, the purposes of IRS participation in the project and its service contract with BLS had to be related to IRS administration of the tax laws.

The terms of a contract between IRS and BLS which needed to be acceptable to SSA enabled BLS to receive tapes containing tax information from IRS and SSA and to combine them with records in the UI Address File maintained by BLS. It imposed strict safeguard procedures and required BLS to provide IRS with a list of all persons permitted to see confidential tax return data. This list included SSA employees who were required to sign affidavits as

agents of BLS.

3. BLS and SSA Agreement

The third agreement was a Conditions of Use agreement between BLS and SSA which enabled SSA to release data from its employer files to BLS and authorized BLS to link data from these files to data in the UI Address File and data to be furnished by IRS. Like the IRS/BLS agreement, it limited access at each stage of the project to those persons who needed to use identifiable data, kept the number of such persons to a minimum, and required adequate physical security procedures. This agreement, which needed to be acceptable to IRS, enabled BLS to use SSA files for the ERUMS project. Under this agreement, SSA would furnish BLS with SSA's Single Unit Code File, Multi Unit Code File, and Employer Report (W-3) Record. The agreement authorized BLS to link data from these statistical files with data in the BLS Unemployment Insurance Address File and with data to be furnished by IRS, and prohibited any other linkage.

4. Microdata Access Group

In the planning and matching stages of the project, the persons who needed to have access to microdata were those members of the Workgroup who were performing the record matching and verification. At Workgroup meetings, members generally reviewed data in the form of frequencies and other summaries to track the progress of the matching operations and to plan future steps. Occasionally, discrepancies appeared or questions arose concerning

classification of a particular employer or possible mis-match of data. Those matters were usually referred to particular members to resolve, with access to microdata as needed on an ad hoc basis.

When the matching steps were completed and time came to plan the analysis, new arrangements were needed to enable a different

group of persons to examine identifiable microdata. The Microdata Access Group (MAG) was formed for this purpose. At this point, IRS agreed that its contractor, BLS, would be permitted to make Workgroup members its agents as needed for the analysis stage. This ehabled the Workgroup members who were employees of BEA and the Committee on National Statistics to become sworn agents who, like the employees of BLS and SSA, would be permitted to examine and analyze microdata. Thus, of the three agencies sharing microdata (BLS, SSA, and IRS), IRS was the only one that did not have access to the matched microdata file. This group met periodically to plan and perform the analysis, prepare findings, and to report its findings back to the full Workgroup.

Once the terms of all contracts were agreed upon, the contracts and the conditions of use agreement were signed by officials of the participating agencies, and the way was cleared for the data transfers.

Summary and Conclusions

To say that the process of discussion and negotiation leading to the signing of the ERUMS access agreements was painstaking, sensitive, and costly in terms of staff time and delay in the study's completion is an understatement. The disclosure aspects of the study severely tested the will and resolve of the affected agencies.

In retrospect, the signing of interagency agreements between IRS and BLS and between SSA and BLS documented a process of negotiation by which the study plan was adapted to the requirements of the varios confidentiality laws that impinged on it. In addition, it summarized a process in which a combination of technical and procedural safeguards were fitted to meet the requirements of the Federal and State agencies that were involved in the data sharing.

While the participants in the ERUMS study all feel a certain, degree of Accomplishment due to their collective persistence, none are quite so upbeat about the long duration of the study. Clearly, the long incubation period for the interagency data sharing agreements was a major contributor. However, it is important to

recognize that the prolonged negotiation for interagency agreements did not result from lack of cooperation among the participants. On the contrary, it reflected the complex mosaic of legal restrictions on use and interagency dissemination of records.

Once it became evident that a single multi-party agreement would be unworkable for the overall project, the plan was broken down into component steps of disclosure, record linkage, and analysis. Each failure to reach an agreement required a step back to re-examine the study imperatives and to adapt the procedures to the practical and legal necessities at each stage.

In addition to adding to the overall time and resources consumed by the project, these delays further contributed to supplemental delays, including:

1. Personnel turnover among the project participants due to

the extended length of the project's schedule necessitated slower progress on the technical issues.

2. The acquisition of IRS Form 940 data was adversely impacted since these have a 5 year retention and were scheduled for destruction by the time the sample EIN's were determined.

On the positive side, however, ERUMS demonstrated that such data sharing projects can be successful under current laws if there is creativity, flexibility, and most of all, persistence.

Notes and References

[1] Section 6103(n) of the Internal Revenue Code (IRC) allows for the provision of confidential tax return information for purposes of tax administration. Specifically, it reads:

"Certain Other Persons. -- Pursuant to regulations prescribed by the Secretary, returns and return

information may be disclosed to any person, including any person described in Section 7513 (a), to the extent necessary in connection with the processing, storage, transmission, and reproduction of such returns and return information, and the programming, maintenance, repair, resting, and procurement of equipment, for purposes of tax administration."

[2] Section 6108 of the IRC has three parts which call for the publication of statistical compilation of tax return information at regular intervals, but, unlike Section 6103(n), such information cannot identify a particular taxpayer. This Section is the primary "mandate" for IRS' Statistics of Income (SOI) program.

a) Publication or other Disclosure of Statistics of Income.

-- The Secretary shall prepare and publish not less than annually statistics reasonably available with respect to the operations of the internal revenue laws, including classifications of taxpayers and of income, the amounts

claimed or allowed as deductions, exemptions, and credits, and any other facts deemed pertinent and valuable.

b) Special statistical Studies. -- The Secretary may, upon written request by any party or parties, make special statistical studies and compilations involving return information (as defined in section 6103 (b)(2)) and furnish to such party or parties transcripts of any such special statistical study or compilation. A reasonable fee may be prescribed for the cost of the work or services performed for such party or parties.

c) Anonymous Form. -- No publication or other disclosure of statistics or other information required or authorized by subsection (a) or special statistical study authorized by subsection (b) shall in any manner permit the statistics, study, or any information so published, furnished, or

otherwise disclosed to be associated with, or otherwise, identify, directly or indirectly, a particular taxpayer.

Section 6108(a) has been interpreted as a tax administration purpose for the Statistics of Income (SOI) Program (unlike 6108(b) and 61O8 (c)); hence, if a 6108 (a) study requires the use of "outsiders", then a 6103(n) contract can be initiated as was done for the ERUMS study.

SAMPLE SELECTION AND MATCHING PROCEDURES IUSED IN ERUMS

John Pinkos

Kenneth LeVasseur

Marlene Einstein

U. S. Bureau of Labor Statistics

Joel Packman

Social Security Administration

Introduction

The first paper in this session described the experience with developing interagency agreements, the third described the findings resulting from the study while this one describes the sample selection and matching procedures used.

In addition to describing the sample selection and matching procedures, the followinq will explain what the ERUMS Workgroup considered when developing the protect design. This paper also describes the sampling frames, data, and manual matching conducted by the ERUMS Workgroup.

The ERUMS project was a pilot study, designed to develop and test procedures for linking and comparing employer and reporting unit data from different administrative record systems. The study from its inception was exploratory in nature, and the ERUMS Workgroup members hoped to observe and document the similarities and differences discovered between the records in the systems being studied and, thus, between the systems, themselves.

The scope of the project included employer reporting unit data from the Bureau Labor Statistics and Social Security Administration employer data files which have similar coverage. Internal Revenue Service data, which were edited by the Bureau of the Census, were used to assist in the analysis of the sample. The ERUMS committee members included staff from Office of Management Budget (OMB) , Bureau of Labor Statistics (BLS), Social Security Administration (SSA), Bureau of Economic Analysis (BEA), Internal Revenue Service (IRS), Census and the Committee on National Statistics (CNS). Developing the sample design, selecting the sample, and performing the machine and manual match were conducted by SSA and BLS staff who were cleared to work with the confidential data. To conduct

the final analysis of the data this group was later expanded to include staff from BEA and the CNS.

There are two reasons for providing an account of the ERUMS sample selection and matching procedures. The obvious reason is that the results, like those of any research study, are dependent on the procedures used, and anyone interested in the results is entitled to a full description of how the study was carried out. The other reason, equally or perhaps more important, is that ERUMS was a venture into uncharted territory, and we believe that future projects of this kind will benefit from the availability of a detailed road map of the procedures that were developed to match and compare employer and reporting unit records from BLS, SSA, and IRS for statistical purposes.

Sample Design Considerations

A major design consideration affecting the size and scope of the project was the limited staff time and resources each of the participating agencies was able to contribute. The committee realized from the beginning, the meat of the project would be in the manual review of the reporting units from each of the administrative record systems. To keep the workload manageable, the Workgroup decided to limit the study to one State rather than several. It was also decided that this State should be large and be one which could share its data with federal statistical agencies for research purposes. The State selected was Texas.

Probability sampling was used at all stages of selection and provided two benefits. It ensured that sample results could be used to produce unbiased estimates for the study population, and it made possible estimation of sampling errors. Additionally, the Workgroup felt it would be useful for both analytical and methodological purposes to produce weighted estimates. Consideration was given to designing a baseline sample where a sample from one agency (e.g., BLS) would be drawn and then a search for the selected sample members would be conducted on the other

agency's files (e.g., SSA). This approach would provide matched units on both files as well as those on the BLS file but not the SSA file. This method, however, would not identify those units on the SSA file but not on the BLS file.

The baseline sample approach was abandoned and it was decided that samples would be selected in two stages. The stage one sample was an equal probability sample of the population which was then stratified by match status. The stage two sample was a systematic subsampling from these strata. This method of sampling provided a means for over- sampling selected types of records which were of more interest to the project and it also resulted in a manageable sample size. As a final design consideration, the committee wanted to ensure that records from both SSA and BLS had an equal chance of selection. Additionally, the Committee wanted to develop an approach that would minimize the number of computer searches

required to select the sample and relevant data elements from these

large administrative record files.

The sample design used was one that selected separate samples from the BLS and SSA files using the same get of random pairs of numbers. The purpose of this design was to measure overlap between the two frames and, more importantly, to measure the amount of nonoverlap between the two frames. The nonoverlap included those sample members on one frame but not the other. This design also minimized the computer costs and allowed the committee to select the sample in one pass through each agency's data file. Once the sample was selected, the relevant data elements for each sample member were downloaded to a micro computer.

Sampling Frames

Both the SSA and BLS data files are compilations of administrative tax records. The SSA data file includes data from employer W-2 and W-3 wage reports, whereas the BLS file includes data from employers' State Unemployment Insurance tax reports. The identifying data element common to both the SSA and BLS files and assigned from a single source is the Employer's Identification

Number, or EIN. The EIN is a unique 9-digit number assigned to companies by IRS and is used to track federal tax payments.

When companies pay State Unemployment Insurance Taxes the State assigns an Unemployment Insurance (UI) Tax number to track payment. Since companies are given a federal tax credit for State UI taxes, they provide their EIN to the State UI tax department. On an annual basis IRS provides each State UI tax department with a file of all the EINs registered in the State. The UI tax department then reconciles the amount of State UI taxes paid by each employer against the IRS file of EINs and tax credits claimed by each employer. By definition, all companies on the SSA files should have an EIN reported, because this is what is required for an employer to be included on the file. On the BLS State file a few units did not have an EIN reported since only a State Ul tax number is required for an employer to be included on that file. The first quarter 1982 Texas file had EINs reported for 98.7 percent of all reporting units.

The sampling frame for BLS was all the EINs reported in the

Texas first quarter 1982 U.I. Name and Address File. The sampling frame for SSA was all the EINs reported in the Single Unit or Multi Unit Code file with wage reports for calendar year 1982. The SSA files are continuous files linked over time, whereas the BLS file in 1982 was a snapshot of one calendar quarter. Effective with first quarter 1989 data, the BLS began linking data quarterly and now has a continuous data file.

The sampling rate was determined by the Workgroup's decision that 400 EINs would be a manageable sample size and that about one-. half of the sample should have EINs classified as multis, or companies with multiple locations. EINs classified as multis were of particular interest because there is more variation in reporting practices.

To derive the sampling rate, the committee looked at the first quarter 1982 Texas file, which had 267,487 EINs classified as

single units and 3,125 EINs classified as multi units. A sampling rate of 6 in 100 was selected since it provided approximately 188 EINs that were multi units.

As previously mentioned, it was decided to select a two- stage sample. The first was an equal probability sample of the population. This first-stage sample was selected from all EINs that had 1 of 6 random pairs of numbers in positions 7 and 8 of the EIN. The sampling rate of 6 in 100, when applied to both the BLS and SSA frames provided a combined stage one sample of 19,964 EINS. The stage one sample was then machine matched and each EIN was assigned a status classification. The initial status classifications are shown below:

MATCH STATUS IN:

Table A

Group	BLS	SSA

1	Single	Single
2	Single	Inactive
3	Inactive	Single
4	Multi	Single
5	Single	Multi
6	Multi	Inactive
7	Inactive	Multi
8	Multi	Multi

EINs that were inactive in both systems obviously had no chance of entering the ERUMS sample.

Another view of the status classifications is shown in attachment A, which is a 3x3 grid having classifications, single, multi, and No Wage Report (NWR) on each scale for both the BLS and SSA files. Records with no wage reports on the SSA file were considered inactive. The bottom right cell on the grid is not applicable since these would be records that did not exist on either file.

Based upon the interest of the Workgroup three of the basic classifications or cells were subdivided and are shown as the shaded sectors on the 3x3 grid (see attachment A). County and SIC became matching criteria for those EINS that were single on both files. The number of reporting units became a criterion for those EINS that were multis on the BLS file but were single on SSA file and those EINs that were multis on both.

These eleven match status classifications became the strata used for the second stage sample. The second stage sample selection had equal probability within each stratum. The sampling rates used varied by stratum, from selecting all to selecting 1 in 173.78. Given the exploratory nature of ERUMS, the intent of the Workgroup was to pull a larger sample of EINs classified as multis and nonmatched records. These cases were expected to present more

difficulties. Therfore, the Workgroup wanted to, have enough of these cases to learn what the situations were and to test methods of dealing with them. The final sample contained 401 EINS, including 201 classified as having multi units on, either the BLS or the SSA files. The remaining 200 EINs were those not classified as multis on either the BLS or SSA files.

Once the sub sample was selected, the Workgroup began the review and analysis phase, which included labor-intensive manual matching.

The working group reviewed reported employment and SIC and geographic codes for each of the 401 EINS. To assist in this process, the Workgroup made arrangements to have access to IRS data for tax years 1981 through 1983. Data for 385 of the 401 EINs were made available.

During the review process the Workgroup attempted to uncover the reasons why records did not match or why records were on one file but not the other. In this process of looking very closely at

the actual records from each agency, the Workgroup learned much about the two systems and found reasons to reclassify some of the records which affected the final match status. For example, in the area of multiunits, the BLS system defines multis as companies with multiple locations within the same State whereas the SSA system defines multis as companies that have multiple locations in the United States. During the review of the multi-unit records, employment levels were considered and attempts were made to reconcile differences in reporting units by aggregating employment of the individual multi units to the EIN level. As a result of this review, the Workgroup decided not to use employment as a match criterion. It was also decided that for purposes of this study, a multi unit EIN would be an EIN that had multiple locations within the State of Texas. This reduced the number of SSA multi unit EINs in the final sample from 120 to 10. The remaining 110 records were reclassified as single EINS.

As noted, the Workgroup also compared SIC and geographic codes from both files. SIC codes were first examined to see why there were non-matches at the four-digit SIC level. In some cases, the non matched EINs were assigned SIC code in related industries; in other cases, the industry code reflected la larger aggregation of the reporting unit. Another, and perhaps more important factor that accounted for differences at the 4-digit level was, both BLS and SSA have policies for SIC coding exceptions.

The BLS in 1982 had 11 exceptions to 4-digit SIC coding which meant a 3-digit SIC code was assigned in certain industries in lieu of the 4-digit SIC code. This represented 43 4-digit industries. These are industries which either have a significant amount of overlapping in their industrial activities or are industries that historically had been difficult to collect sufficient information from to assign a 4-digit SIC. The BLS currently has reduced the number of 4-digit coding exceptions to 6, which represents 17 4-digit SIC industries.

The SSA SIC coding exceptions exist in some agricultural industries and Public Administration, which are coded to the 1

digit level. This affected 64 4-digit industries. Approximately 63 other 4-digit industries were coded at the 3-digit level for one reason or another, typically insufficent information. In addition to reviewing SIC codes, the Workgroup also looked at geographic codes and tried to explain why some records did not match between files. Maps and coding manuals were consulted and the review showed there was some inherent misreporting of county codes by employers. Texas has more than 37 cities with the same name as a county but these cities ate not located in those counties. Houston, for example, is in Harris County not Houston County and Austin is in Travis County, not Austin County. Counties named Houston and Austin are located elsewhere in the State. In some cases the reason for non matching records was that the reporting unit was coded in an adjacent county. Texas has a very large number (254) of counties. For those employers who keep their records by city or are not familar with the county names, it is easy to see the potential for some misreporting.

The Workgroup also looked very closely at the cases having inactive EINs on either the BLS or SSA files. Inactive EINs for

the BLS were defined as those that appeared on the SSA file but did not Appear on the BLS File. Inactive EINs for SSA were defined as those on the SSA file with no wage reports for 1982.

When reviewing the BLS inactive EINs, the Workgroup used SSA SIC and employment data to determine if the employer was exempt from Unemployment Insurance coverage. They also looked at IRS data to determine if the employer became active after the first quarter of 1982 and at the first quarter 1983 Texas file to see if the employer reported in 1983.

When reviewing the SSA inactive EINS, the Workgroup was able to use a more nearly complete SSA wage report file that included wage reports that were either delinquent when the sample was selected or were in the process of reconciliation with IRS. As a result of these additional data, 44 of the 99 EINs originally classified as inactive on the SSA file were determined to be

active. The Workgroup also used the BLS 1982 and 1983 first quarter Texas files to conduct name searches to see if the same employer reported under a different EIN. The Texas files were also used to see whether zero employment was reported, which might have indicated no wages were paid. Additionally, IRS data were then used to see what level of employment was reported to IRS.

The last step in the review and analysis phase was to determine the final match status of the 401 EINS. As a result of the review, it was decided to collapse the 11 categories shown in Attachment A down to the basic 8 cells shown in Table A.

As part of the final analysis, committee members worked on completing the documentation for the project and discovered that an additional 2,608 EINs that were on the SSA file but not the BLS file were inadvertently omitted from the first stage sample and, consequently, from the second stage. Adding cases to the stale 1 and 2 samples at that point in time would have further delayed completion of the study, so the Workgroup decided the best way to deal with this problem was to reweight the sample cases in the two

affected strata and rerun the results tables.

MATCH STATUS CLASSIFICATIONS

KEY: NWR = No Wage Report

SIC = Standard Industrial Code

RU = Reporting Units

RESULTS, FINDINGS, AND RECOMMENDATIONS OF THE ERUMS PROJECT

Vern Renshaw

Bureau of Economic Analysis,

Tom Jabine

Statistical Consultant

The other papers in this session have examined the administrative arrangements and the sample selection and matching procedures for the Employer Reporting Unit Match Study (ERUMS) This paper reviews the study's results, findings, and

recommendations.

The main purpose of the ERUMS project was to provide information on the technical and administrative feasibility of interagency record linkages. However, the ERUMS Workgroup hoped that the study would also shed some light on at least three areas of substative concern.

1) We hoped that geographic and industry information for reporting units contained in the Bureau of Labor Statistics (BLS) Unemployment Insurance (UI) Address File could help evaluate the potential statistical usefulness of a) reporting unit data supplied by multi unit employers participating in the Social Security Administration (SSA) Establishment Reporting Plan (ERP) for forms W-2 and W-3; and b) State data supplied to the Internal Revenue Service (IRS) on Form 940. SSA has been concerned about the quality of its reporting unit data because resources for maintaining the ERP had been inadequate for some time and the State data supplied on IRS Form 940 had never been used for statistical

purposes.

2) We hoped that information from LRS and SSA files could help evaluate the completeness of employer coverage in the UI Address File. The UI Address File leaves out or estimates employer information that is not received by its statistical deadline, whereas information for late reports was generally available in the IRS and SSA files used for ERUMS.

3) We hoped that the analysis of matched records could help evaluate the consistency of industry and geographic coding in the BLS, IRS, and SSA systems.

The extent to which the ERUMS project could actually shed light on these areas was limited by several factors. First, ERUMS was a pilot study based on a small sample drawn from a single State (Texas) for a single year (1982). The results, therefore, could

not be expected to reflect precisely the status of the data systems for the entire country or for subsequent years. (BLS has taken steps to improve the UI Address File since 1982.) Second, both the information content and processing procedures differed somewhat among the data systems. The W-2/W-3 data were for calendar 1982, for example, while the UI Address File that was used contained data only for the first quarter of 1982.

Finally, a number of unanticipated problems were encountered in carrying out the study. The most limiting of these problems resulted from the slow implementation of ERUMS. For example, by the time the final sample of employers was selected, many IRS Form 940s for 1982 had been destroyed. Therefore, it was not possible to evaluate the State data contained on the Form 940s.

Another unanticipated difficulty arose because the initial SSA files used in the matching process omitted some wage reports and were generally inadequate to determine if employers were actually reporting multiple units in Texas. These initial files were later

supplemented with more complete information, but the supplementation occurred after the final sample had been I drawn; consequently the size of the sample was smaller than intended for some categories of employers, especially for multi unit employers.

Finally, it proved to be more difficult than had been anticipated to account for differences in employer coverage among the data files. In part, this was because estimated data were not identified in the UI Address File (a deficiency being corrected) and because there was no documentation of such phenomena as dates when employment started for employers (or ended, or was changed by reorganization, etc.) or dates when forms filed by employers were received by the processing agencies.

The clearest conclusion to emerge from the ERUMS project related to the poor quality of SSA's ERP data for multi unit employers. It was evident that SSA would need to take steps to improve quality control it the SSA system were ever to be useful for developing data by geographic and industry classification. The other findings of the ERUMS project were not so stark as those

relating to the poor quality of SSA's establishment data, but the study could well reinforce the concerns of those who worry about the inconsistencies in industry coding that occur when employers are coded independently by different agencies.

In the following sections of the paper the results, limitations, findings, and recommendations, of the ERUMS project are discussed in somewhat greater detail. Tables A-1 to A-8, which are referred to in the next two sections, appear in Chapter III of the ERUMS final report (Statistical Policy Working Paper 16). In order to meet space limitations, we have included Only Table A-4 with this paper.

Results

As explained in detail by Pinkos et al in the second paper of this session, the ERUMS sample was a two-phase sample of employers, as defined by unique Employer Identification Numbers (EINs). Most

of the results presented in this paper are estimates based on the Phase II sample of 401 EINS, weighted to account for the disproportionate sampling used in the second phase of the sample selection.

Of the Texas EINs that were active in 1982 in the BLS or SSA systems, 67.1 percent were active in both systems, 27.6 percent were active only in the SSA system and 5.3 percent were active only in the BLS system (Table A-1). Only about 1.0 percent of all active EINs were classified as multi unit in one or both systems, and most of these were classified as multi unit only in the BLS system (Table A-4).

For the matched single unit EINS, i.e., those that were active in both systems, an estimated 81.6 percent had the same State and county codes in both systems. The remaining cases were about equally distributed in three categories: same State, different county; same State with no county code in the SSA file; and different State (Table A-5). An estimated 70.2 percent of the matched single unit cases had the same two-digit industry codes.

About half of the remaining cases were not classified by industry in the SSA system (Table A-5). When matched against the IRS/Census-edited Form 941/943 file, about three-fourths of the matched single units from both the BLS and SSA files had two-digit industry codes that agreed with those in the IRS/Census file. However, when the SSA unclassified cases were excluded from this comparison, the proportion of SSA cases that agreed with the IRS/Census two-digit code was somewhat greater than the corresponding proportion for the BLS matched single unit cases (Table A-8).

Only a few EINs (nine sample cases) were classified as multi unit in both the BLS and SSA systems. Matching individual reporting units for these cases proved to be difficult. Overall, the nine sample employers had 105 Texas reporting units in the BLS system and 60 in the SSA system for 1982.

Of the active SSA EINs not found in BLS's first quarter 1982 UI Address File, it was estimated that 69.2 percent had reported no first quarter employment to IRS on Form 941 and therefore would not

normally be expected to appear in the BLS system (Table A-6). For another 10 percent of these employers, the analysis suggested that they may not have met requirements for UI coverage in Texas either because they had no operations in Texas, because of nonprofit status or because their payrolls were too small. For the remaining 20 percent, the reasons for their absence are not always clear, but

it may have resulted in part from lags in incorporating new employers in the UI State agency and BLS files.

Most of the employers who were included in the 1982 UI Address File but did not file 1982 W-2/W-3 wage reports (22 sample cases) appeared to have ceased hiring employees, gone out of business, or gone through other changes that altered their reporting to IRS and SSA. Half of the employers in this group reported no employment in the 1982 UI Address File. Many of the remainder had filed their final Form 941 with IRS (at least for the period 1981-1983) for a quarter in 1981.

An analysis of the sample EINs that appeared in SSA's Multi Unit Code File provided some indication of the extent to which multi unit employers were participating in SSA's Establishment Reporting Plan (ERP) in 1982 (Table A-7). An estimated 35.9 percent of these EINs had been incorrectly added to the Multi Unit Code File as the result of a processing error that has since been corrected. Most of the remaining employers had initially agreed to participate in the ERP, but more than half of this group did not provide separate data for each reporting unit in their W-3 wage reports for 1982.

Limitations

Several factors limit the broad applicability of the ERUMS findings. The results reflect the reporting requirements and operating procedures associated with the agency record systems in 1982. There have been significant changes since then. In particular, BLS has taken several steps to improve the timeliness and the completeness and accuracy of data in its UI Address File.

The study was based on data for a single State, Texas, and on a small sample of employers and reporting units. The UI system gives the States some latitude in their record-keeping practices, so indications of the coverage of employers in the record systems of the Texas State Employment Agency in 1982 should hot be assumed to apply fully to the UI systems of other States at that time. The small sample size means that estimates based on the Phase II sample are subject to relatively large sampling errors. Because of limited resources and the complexity of the Phase II sample design, we were able to compute sampling errors only for a few key estimates (see Table A-4).

The analysis of the results was complicated by differences in concepts and coverage in the record systems used in the study. These differences occurred in the basic filing requirements for the UI and SSA/IRS systems, the time reference of the basic BLS and SSA files used for matching, the definition of reporting units in the BLS and the SSA/ERP systems, and the structures of the BLS and SSA industry classification systems. In addition, certain file

deficiencies and operational problems made the analyses more difficult. About 1.3 percent of the records in the 1982 UI Address File for Texas did not have EINs and therefore were not included in the Phase I sample of EINs from that file. I In the SSA files, a significant proportion of employers lacked county and industry codes. The most serious problem was that a high proportion of multi unit employers were not reporting separately in 1982 for each reporting unit, so that we were unable to do a thorough comparison of reporting units for multi unit employers active in both the BLS and SSA systems.

Although these differences and file deficiencies made the analyses more difficult, the fact that we succeeded in identifying and documenting them is an indication that the ERUMS project succeeded in its main goal, which was to demonstrate the feasibility of doing matching studies as a means of evaluating the suitability of administrative record systems for statistical uses.

The data on amounts of employment and payroll available from SSA, BLS and IRS files were used in reviewing the unmatched sample cases and trying to understand why they were not present in both SSA and BLS files. However, the employment and payroll data were not added to the data file for the 401 sample EINs that were used to develop the estimates presented in this report. Therefore, all of the results shown are estimates of numbers of employers or reporting units classified by attributes such as match status, and geographic and industry codes in the different systems included in the study. We did not attempt to estimate what proportions of aggregate employment or payroll were accounted for by employers who were unmatched or had different geographic or industry codes.

Findings

The detailed analyses of the ERUMS data did not suggest that large numbers of employers who report wages in one of the payroll

tax systems were failing to report in the other system when they should have been. They do, however, suggest that late reports and different procedures for processing the reports in the two systems created potential problems for using both of the systems data files for statistical purposes.

Perhaps the clearest finding was that it is not possible to maintain a usable establishment reporting unit plan for multi unit employers in the absence of systematic procedures for monitoring employer reporting and updatig files for changes in the number, location and industry of each employer's reporting units. SSA's Establishment Reporting Plan clearly lacked the necessary resources to do this in 1982 and there is no reason to think that the situation has improved since then.

There, was a moderately high but by no means perfect correspondence between county and two-digit industry codes for single unit employers included in both the BLS and SSA systems. A substantial proportion of the differences arose from the absence of

county or industry codes in the SSA system. Comparisons of industry codes at the three and four-digit level were not attempted because of the differences in the industry classification systems used by the two agencies.

With some qualifications, we were successful in matching the records of employers, as defined by their EINS, in different systems. However, we were not successful in matching BLS and SSA records for reporting units, the main reason being the incompleteness of SSAs data for reporting units provided under the voluntary ERP. Other reasons were the lack of a common identifier, analogous to the EIN at the employer level, for reporting units and the slight differences in the reporting unit definitions used by BLS and SSA.

We learned what we believe are some important lessons for others who may wish to match business records from different agency sources, whether for research or operational purposes. First, the plans and the necessary interagency agreements should be developed well ahead of the earliest date at which the files to be linked are

expected to be available. In particular, the development of interagency agreements for the exchange of identifiable records is a painstaking process and considerable time may be needed for their completion and approval.

Second, successful matching requires in-depth knowledge of all of the record systems involved and of the specific files that exist within those systems. An interagency team approach, with full exchange of information, is essential because there is unlikely to be a single individual who has all of the necessary information, even for the files of a single agency.

Finally, whenever possible, it is essential to pretest matching procedures before embarking on large-scale operational applications.

Recommendations

ERUMS was designed primarily as a demonstration project and

was therefore limited in its coverage and scope. Nevertheless, the Workgroup believes that the study results, along with other information acquired in the course of the study, justified the inclusion in its report of five formal recommendations addressed specifically to the BLS and SSA record systems for employers and reporting units. These recommendations were:

1. SSA should undertake a full review of the current status and uses of the Establishment Reporting Plan and decide either to continue it with adequate resources for maintenance and improvement of quality or to discontinue it entirely.

   (note- such a review was begun by SSA prior to the completion of the ERUMS project. As a result of that review, SSA is taking steps to prepare for the

termination of the ERP.)

2. BLS should review the State Employment Security Agencies' procedures for identifying employer births (including those resulting from mergers and changes of organization) and seek ways of reducing the apparent lag between filing of applications for EINs and inclusion of new employers on State Agency and BLS lists used as frames for statistical surveys and reports.

3. Data in the UI Address File on employment and wages paid should be labelled to distinguish imputed data from data reported by employers.

4. The EIN should be identified as a key item in the UI Address File and efforts should be made to achieve 100 percent reporting initially and current reporting of changes in EINS.

5. BLS and SSA (if it continues the Establishment Reporting Plan) should strive to obtain data from employers for,

their establishments as defined in the 1987 Standard Industrial Classification (SIC) Manual Both agencies should code industry for all establishments, without exception, at the 4-digit SIC level of detail. Whether or not the Establishment Reporting Plan is continued, SSA should code all employers identified on Forms SS-4 at the 4-digit level of detail.

(see parenthetical note following recommendation 1 concerning the current status of the ERP)

In a broader context, the ERUMS Workgroup concluded that current efforts to collect economic data at the establishment level are dispersed among Federal and State agencies, are poorly coordinated, and place unnecessary burden on employers. The Workgroup believes that further, more intensive and extensive interagency matching studies have an important role to play in resolving these problems and in determining the possible effects on statistical programs of prospective major changes in administrative reporting systems for employers. We therefore recommend that:

6. Further matching studies should be directed at acquiring information that will support the eventual development of a mandatory reporting system to meet the needs of all Federal and State statistical programs for establishment lists, including SIC codes. An interim goal should be that all agencies requiring or requesting employers to provide data at the establishment or reporting unit level adopt common definitions of units and data items to be submitted for these units.

Three agencies the BLS, the Census Bureau and the National Agricultural Statistics Service -- play a dominant role in the direct collection of establishment-level economic data. Recent initiatives of these agencies, under the general guidance of OMB's Statistical Policy Office, have been directed at greater coordination of their respective list-building and maintenance activities. Further integration of business lists will require

fuller understanding of the similarities and differences of the three systems, based on matching of individual establishments and reporting units in the different systems.

1/Numbers in parentheses are standard errors of the percents.

* Indicates a standard error of less than 0.05 percent.

DISCUSSION

W. Joel Richardson

Charles A. Waite

U. S. Bureau of the Census

Introductory Comments on ERUMS

First of all, I would like to thank the many people who have been involved with ERUMS. Their commitment and resourcefulness have helped to make the ERUMS project a success. As Vernon has detailed, several recommendations were presented that undoubtedly will improve the business files of the Bureau of Labor Statistics (BLS) and the Social Security Administration (SSA). But more importantly, the ERUMS study provided valuable experience in the technical aspects of matching interagency data sets. I am hopeful that this experience will help to further the efforts of data-

exchange initiatives among federal statistical agencies in the coming years.

When the preliminary planning for ERUMS began in 1983, the Census Bureau expected to be one of the participating agencies. Our business employer files were to be matched along with those of the BLS, SSA, and the Internal Revenue Service (IRS). However, there were significant problems concerning the release of our confidential data. Though we realized the importance of ERUMS, we could not resolve these data-access problems soon enough to allow us to be an active participant. As an alternative, the Census Bureau obtained observer status, which enabled us to closely follow the progress of ERUMS.

Before critiquing the three papers, I'd like to expound on the value of the ERUMS study to the federal statistical community. Warren stated that a major goal of ERUMS was to test the feasibility of matching employer records from the business lists of different government agencies. This goal was, accomplished in ERUMS, and the results showed that the matching of the two distinct

data files is possible. Additionally, the ERUMS evaluation revealed problems associated with matching the interagency data files. I expect that these findings will be valuable in future matching studies.

A matching study should be the first step in any data-sharing proposal -- before a data sharing proposal is accepted by the participating agencies, it is essential to confirm the comparability of the data sets and to resolve any conceptual an definitional differences. In my view, the ERUMS project showed that the BLS and SSA data sets are comparable, and that an effective matching operation is possible.

Although there are obvious discrepancies between the data sets

-- only 67.1 percent of the EIN records were active in both systems

-- significant benefits could be realized through data sharing.

First, greater consistency in the industrial classification codes, geographic location indicators, and related data values could be

achieved by sharing the data for matched records. Second, unmatched records could be researched in an effort to ensure the completeness of each of the employer universes. Though numerous issues would need to be explored and settled, such a data-sharing plan could result in greater comparability among the data series.

Currently, the administration has a legislative proposal in Congress that would permit limited data sharing between the Census Bureau and the Bureau of Economic Analysis (BEA). The primary purpose of the proposal is to provide BEA with confidential access to the Census Bureau's establishment information. This information will augment and improve the data on foreign direct investment that BEA collects and publishes.

There are other versions of the legislative proposal in Congress to share Census and BEA data -- not only with each other, but, in at least one version, with the Government Accounting Office (GAO) and the Committee on Foreign Investment in the U.S. (CFIUS). We are concerned that response rates may decline if our microdata are made available to such policy-making organizations as GAO and

CFIUS. For this reason, the Census Bureau does not support this legislative proposal.

The BEA collects foreign-investment data at the enterprise level. The Census Bureau conducted a feasibility study that showed BEA enterprise-level data could be linked successfully with Census Bureau establishment data. By integrating our establishment-level data with BEA enterprise data, BEA will be able to present foreign direct investment statistics at a much finer industry and geographic level. This is one of many possible data-sharing plans that could provide significant cost and qualitative benefits to Federal statistical programs.

I would like to believe that the administration's legislative initiative, together with successful match studies such as ERUMS, will provide the impetus for increased data sharing among Federal statistical agencies in the future.

Interagency Agreements for Microdata Access: the ERUMS Experience

Tom Petska's presentation focused on the interagency agreements required to comply with the confidentiality provisions that govern the three sets of data. Clearly, the matching of individual records in the ERUMS project could not take place until these confidentiality issues were resolved.

Tom has presented thoroughly the problems associated with sharing the individual records from different agencies. It is apparent that these legal agreements represented a major barrier in the ERUMS project. To their credit, the ERUMS workgroup was able to overcome, the confidentiality problems and to formulate a workable plan -- IRS contracted with BLS to perform the match, and SSA staff were designated as special agents of BLS to process the data. The IRS is permitted to disclose tax information to outside contractors as long as it is for purposes of tax administration, and the ERUMS study was considered to be a statistical study

related to the administration of IRS tax laws. Unfortunately, considerable time was spent in determining this solution and in drafting the required legal agreements. This added considerably to the length of the ERUMS study. Future matching studies may face similar obstacles in gaining access to confidential data.

As an example, the Census Bureau obtains the EIN and related data values for many small employer businesses from the IRS. Any future studies undoubtedly will rely on the EIN to match the records, because the EIN is the one key identifier common to U.S. data systems. But as Tom has pointed out, the EIN itself is protected by Internal Revenue Code confidentiality provisions. For this reason, the EIN and related data that the Census Bureau obtains from the IRS cannot be released to other statistical agencies such as the BLS. Only those business records whose EIN and related data have been confirmed through direct respondent contact would be eligible for release. This would impact on the completeness of any matching studies between the BLS and Census Bureau data sets, because a portion of our business universe has not been directly canvassed.

The BLS was permitted access to IRS records in the ERUMS project because of tax-administration purposes. Although additional studies possibly could be conducted using similar arrangements, it would require the support of the IRS and other agencies that furnish the administrative data. Otherwise, future studies may require changes to relevant statutes and regulations before microdata access is authorized. Such changes are difficult to obtain.

I do have one minor point on the paper concerning the confidentiality provisions of the BLS data. The ERUMS study used matched BLS records from only one state -- the state of Texas. Although Tom outlined the disclosure provisions associated with the data records from Texas, it was unclear whether these provisions were typical of the other 49 states. We understand that BLS affords each state with certain latitude as to the collection of the unemployment data. If the states also have different confidentiality provisions -- specifically, provisions that strictly prohibit the release of data to Federal agencies other than BLS -- the ERUMS project may not have been possible using

records from these states.

One of the goals of ERUMS was to gain experience in the procedure of obtaining access to the confidential data of the various data sets. To this end, the ERUMS study was a success. The study revealed the problems associated with obtaining the access to the microdata for matching purposes, and also determined a workable solution that overcame these problems. However, I expect that disclosure problems will continue to be a major obstacle in future matching initiatives.

Sample Selection and Matching Procedures for ERUMS

John Pinkos's presentation focused on the sample selection and matching procedures in ERUMS. As John has pointed out, a major constraint affecting the sample size was the limited staff time and resources. Because considerable analysis was inevitable for the

sampled records, the-ERUMS members agreed to select a relatively small sample. As it turned out, 401 cases were selected.

By limiting the sample to one state, and oversampling from certain categories of records that were of particular interest, ERUMS was able to create a manageable set of sample records that were sufficient to meet the study's objectives. I expect that future matching studies will benefit from the details, of the procedures used in ERUMS.

Three sources of data were used in the study -- BLS data, SSA data, and IRS data. Cases were selected first from the BLS data files and then independently from the SSA data file. Using this technique -- specifically, by selecting independently based on certain digits of the EIN -- the ERUMS sample included records that were present in only one of the two data systems, as well as records that were present in both systems. Records present in only one of the data systems were a critical part of the study, as these represented potential differences in employer coverage between the two data files.

The ERUMS study, however, did not sample from the IRS data set. The IRS data were used only to help analyze the BLS/SSA cases selected in the sample. The IRS file was not included in the sample selection because of the difficulties in gaining access for such a purpose. Although this decision was unavoidable, it may have compromised the results of ERUMS somewhat.

The IRS data file represents a complete universe of business employers in 1982 -- all employers who filed payroll tax returns in with no exclusions as to the size of the business or the nonprofit status of an organization, were included on the IRS file. Without this complete file of businesses, ERUMS was left to compare records from the BLS and SSA data sets. Although differences were identified and quantified, the study could not make valid estimates

on the completeness of the two data sets as compared to the universe of businesses on the IRS file.

A similar point exists for the matching of multiunit records from the BLS and SSA data sets. The ERUMS study showed that about l percent of all active EINS were classified as multi unit in-one or both systems. Most of these were classified as multi unit only in the BLS system. One of the findings of the study was that the SSA multiunit file is deficient, and steps should be taken to either improve the quality or to discontinue it entirely. Because of the obvious deficiency in SSA's multiunit file, no legitimate conclusions could be reached on the accuracy of the BLS multiunit file.

One last point on John's paper, he discussed briefly the comparison of industry classification and geographic location from the BLS and SSA files. I would liked to have seen some general table that presented these results. Even if the results were presented at broad industry and geographic levels, it would have provided some general information on the comparability of these critical data elements.

Results, Findings and Recommendations of the ERUMS Project

The agencies involved in the ERUMS project have gained valuable experience in the technical aspects of linking data files and in the administrative requirements for gaining access to the data. For this reason alone, the ERUMS project should be considered a success. In addition to the experience gained, the ERUMS project presented several recommendations that will help to improve the business files of the BLS and SSA. I understand that the BLS has already taken several measures to improve the timeliness, completeness, and accuracy of the data in its Unemployment Insurance Address File.

Vernon's presentation detailed the recommendations that were identified in the ERUMS study. In one of the recommendations, he stated that BLS should review the procedures for identifying births in an effort to improve the timeliness of including new employers in the BLS lists. I suggest that the BLS review procedures for identifying deaths as well. Up-to-date operational status is a critical element of business employer records.

The final recommendation in vernon's presentation covered the need for additional matching studies to acquire information that will support the eventual development of a reporting system to meet the needs of all Federal and State statistical programs. Because of certain legislative barriers -- for example, Title 26 strictly prohibits the release of IRS data to other statistical agencies -- and significant operational problems, such a far-reaching goal may not be plausible in the foreseeable future.

The Census Bureau supports a more achievable goal of data-sharing among Federal statistical agencies, and would welcome the opportunity to conduct additional matching studies in an effort to further data-sharing initiatives. Before proposing the Census/BEA data-sharing initiative, we conducted a matching study that confirmed the feasibility and value of linking our establishment-level data with BEA's enterprise-level data. This preliminary study was a necessary step in the Census/BEA data-sharing

initiative. Additional matching studies may promote other data-sharing initiatives in the Federal Government.

The ERUMS project, which effectively matched interagency data files, may help provide the impetus for increased data sharing in the coming years. With the necessary legislative changes, pertinent data from each of the employer files could be shared among statistical agencies. Such a data-sharing plan would provide major advantages, including greater comparability among economic data series, less respondent burden on the business community, and a reduction in overall Government costs.

Summary

Comparisons between data sources are beneficial because they highlight conceptual differences and identify the limitations and strengths of the data sets. The ERUMS project successfully met both of these objectives. In addition, ERUMS provided valuable experience it the technical aspects of matching interagency data sets.

Our current mission should be to use this experience to further the efforts of data sharing in the Federal Government. Data sharing offers major advantages to Federal statistical agencies. By supplementing business data sets with applicable information from the data sets of other agencies, the Federal statistical system will attain greater comparability in related economic data series. The ERUMS project showed that interagency. data sharing is a viable option. I would like to congratulate the many people who have been involved with ERUMS for a job well done.

DISCUSSION

Thomas J. Plewes

U.S. Bureau of Labor Statistics

I appreciate the opportunity to appear at this public

unveiling of the Employer Reporting Unit Match Study (ERUMS) report. This is an event that has been long-awaited by all of those who have been involved in this multi-agency, multi-year, and multi-faceted project. I expect that no participant has awaited this day more anxiously than Warren Buckler, who, along with the folks here at the speaker's table and many in today's audience, has spent a great deal of time over the past few years in conceiving, giving birth, and nurturing this little study. Indeed, to carry the metaphor further, is hard to figure out where we stand now on the continuum from project conception to death. Is this session a commencement ceremony, or is it a eulogy? As my commentary will soon indicate, I hope that we are gathered for a commencement ceremony for the statistical community has learned important lessons about sharing and about the basic quality of two major business lists in this project at some significant cost. It would be a shame if the lessons learned were not put to use in implementing critically needed program improvements.

I would like to accomplish two objectives in the short time I have allotted as a discussant. First, I want to step back to

examine the environmental framework in which this study took place and contemplate the arena into which the report now has been thrust. My second goal is to draw specific conclusions from the exercise and suggest specific steps that should be taken as a result of the work that has been done.

What is the environment in which we must consider this study? It is a complex environment, characterized by:

1. Little sharing of business directory information between Federal government agencies, but a growing pressure to develop, procedures for sharing so as to reduce the burden on respondents. These pressures are building to the extent that I believe sharing will surely be mandated. That mandate may come in the form of legislative action, a fiat from the Office of Management and Budget using its authority under the Paperwork Reduction Act, or of most profound consequence, through a centralization of the statistical agencies.

2. A reliance on lists characterized by their primary usage as administrative data sources which focus the support Of the administration of the law or function. We have built our elaborate business directory programs and constructed our business survey frames on databases that have been

developed with only a distant secondary concern for the statistical uses of the data.

3. Difficulty in separating statistical from enforcement purposes. If we, as statistical agencies, make the data better and create an environment for comparing lists, we enhance their use for enforcement and administrative purposes also. This aspect will be particularly troublesome when we involve, as we eventually must, the Internal Revenue Service in sharing schemes. The participation of the IRS in the ERUMS process gave us an indication of the lengths to which IRS will go to protect the tax data, and of the difficulties this injected in

the ERUMS process.

4. A growing concern over confidentiality of establishment records.

5. A lack of consistency of definitions and coding that extends throughout the statistical system, but has a most profound impact on sharing of administratively-derived lists. Administrative differences in the programs lead to inconsistent definitions of even the most simple of terms, such as "employment", "address", "wages" and the like.

6. An expanding recognition that errors and omissions in the business lists are a significant source of error in the survey process. The Federal Committee on Statistical Methodology's Working Paper 15, "Quality in Establishment Surveys" documented this, and the Tupek-MacDonald paper this morning discussed the effect that the Bureau of Labor Statistics' Business Establishment List improvement

project will have on BLS survey quality.

These environmental elements pose formidable challenges to statistical agencies that want to improve the efficiency of their operations and reduce burden on their reporters. For example, in terms of frames for surveys of nonagricultural businesses, there are at present two major government lists -- the Census Bureau's Standard Statistical Establishment List (SSEL) and the BLS Business Establishment List (BEL) -- and one major private sector list -- the Dun & Bradstreet file -- with a myriad of lesser known and more specialized lists for more limited purposes. We can look at the SSEL as a representation of the of the SSA/IRS administrative data files with considerable value added by the Census Bureau. Likewise, the BEL may be seen as a representation of the State unemployment insurance files with considerable BLS value added. If these Federal government files do not match, and we suspect they do not through analysis of the macrodata, the problem can be with the basis administrative data files, with the value added, or both. Over the years, Fritz Sheuren's various administrative database

comparison projects have documented the systemic differences in the files very well. They must be borne in mind. Fixing the files once we have identified the root difficulties is quite another matter. The statistical agencies do not own them, and they are exceedingly expensive to change (in terms of budget and response burden). Indeed, quite often only a revision in law or nationwide program practice will do the trick.

Fixing the "value added" portion is somewhat more possible, but it too is expensive in terms of budget and people. often there are good reasons for not fixing the way we add our value, such as the need to assure the continuity of historical data series.

Definitions are another challenge. If we want to share lists, we must think in terms of three types of problem. In some cases, repair is relatively simple. We heard today, for example, that our definitions of multi-unit employers are already in close proximity. The EIN and SIC systems are also bedrock. Our challenge in those

instances where there is close concordance between the files is to maintain the definitional base in a standardized, current and relevant manner.

In other areas, we must change the way we do business but, if we are willing, our task will be reasonably easy. One match problem that ERUMS identified was that the project was comparing annual SSA reporters with lst Quarter UI reporters. This is one Of the problems that we can fix with time and resources, because the data are there.

In a few important other cases, however, we are quite limited in our ability to bridge definitional gaps. For example, when coverages are based on Federal laws, State laws, and judicial precedent regulating the administrative database, we would be forced to justify a change in the insurance or tax program on statistical grounds.

Certainly, confidentiality concerns have a presence in the equation. We, glimpse in the Petska-Alexander paper the importance that necessary confidentiality protection schemes had in this

project, and the price those schemes exacted in terms of time and precision. That's one of the reasons I like the Petska-Alexander paper so much. It outlines the practical implications of maintaining a pledge of confidentiality when cooperating on a project of importance to the statistical agencies. Everything, as they so well point out, had to be invented. There are no text book examples of interagency agreements on confidentiality. The solutions which the project team developed were carefully crafted to stay within the very restrictive IRS law and were implemented with an eye toward the reality of the environment. Thus, there are really two stories in the Petska-Alexander paper. One story is about the difficulties that the team encountered in sharing confidential data. The other, written between the lines, is about

the sense of cooperation and dedication that allowed the cumbersome solutions to move forward.

The Petska-Alexander paper starkly reminds us that the role of

confidentiality policy is important but little understood. We may be hopeful that the current situation will be short-lived. The National Academy of Science's Committee on National Statistics had taken on these issue with the formation of an expert panel. Until we are able to benefit from that report, however, we are left with the fact that understanding of confidentiality of business records has not progressed very far as either science or practice. Only recently has a literature on the subject of confidentiality begun to emerge, but most of it addresses the more emotional topic of confidentiality of information about individuals. The literature pays little attention to issues surrounding confidentiality of business records. Without such a foundation, the statistical agencies have mostly assumed that the issues of confidentiality of business records are the same as those for individuals. This assumption has played an important role in justifying past limits on sharing between the Federal agencies.

The second paper, by Einstein, Levasseur, Packman, and Pinkos, also attempts to stand back with benefit of hindsight and make some sense out of what was a convoluted process. Since 3 of the 4

authors work with me, these comments may not be as critical as others may have rendered, for all along the way I "bought-in" to the approaches taken and the effort expended. Nonetheless, I view the documentation that this paper offers in a somewhat different light than the authors, and draw slightly different conclusions.

The matching process, as described, makes a good deal of statistical sense. The team selected a two-stage sample selection process, stratified into 9 groups. The second phase, a subset of about 400 cases of the first selected on a probability basis, provides for detailed analysis. Some of the specific steps in the process were to meet the confidentiality restrictions, but not all.

The process that the team established should serve as a first step toward developing an on-going statistical process control system, if and when sharing does take place. Many of these same activities should be continued in a recurring program to meet the objectives of total quality management. Thus, the work of the team has long-term, permanent implication. The authors seemed to recognize this when they stated that "we believe future projects of

this kind will benefit from-the availability of this detailed road map". Probably so, but I speculate that future researchers will look at the road map and decide against making the journey. That is why I would take pains to separate the enduring aspects that should be the foundation of a quality management system from those that were necessary to meet more bureaucratic objectives.

The contribution of the Renshaw-Jabine paper is to Yield some hope, in that it reminds us how close we are to an ability to share, while providing some sober reflection about some major tasks still lying ahead if we are to share. Their bottom line is that the systems are reasonably close in coverage -- eventually most employers emerged in the systems. There were troublesome differences in multi-unit identification, in county coding, and in industrial classification at the 2-digit level, but I would label these of moderate concern. Indeed, under the BEL initiative, BLS has taken steps to correct many of the inadequacies in its data, investing with the States in improving SIC coding, interpretation

of SICS, and, more recently, in fixing the multi-establishment identification problems. Unfortunately, with lack of resources, the Social Security Administration has not been able to make the same investment, so many of the difficulties in the SSA file may have multiplied.

In summary, we ought not let this expensive experience lie on the shelf. We have learned a great deal about two files -- lessons that should be extended to files maintained by the Bureau of the Census. And we need to get on with fixing some of the obvious flaws in the administrative data. Most importantly, we have learned that maintaining confidentiality is possible, that matching is feasible, and that the will is present at the staff level in the agencies to make it all come together. Now it is time for leadership. As Senator Bennett Johnston said in an argument before Congress, "There's a time to stop talking the talk and start walking the walk." We have the map. Let's start walking.

Session 10

APPROACHES TO DEVELOPING QUESTIONNAIRES

TOOLS FOR USE IN DEVELOPING QUESTIONS AND TESTING QUESTIONNAIRES

Theresa J. DeMaio

U. S. Bureau of the Census

As the collection of information through surveys becomes more prevalent in our society, increasing numbers of people find themselves in a position to develop questionnaires. Writing a questionnaire seems like such a simple task -- many people think that anyone without training or experience can do it. But developing a good questionnaire -- one that can obtain good quality, information that meets the objectives of the survey -- is not as easy as it looks. Many different kinds of abilities, including subject matter expertise, writing capabilities, and knowledge of social psychological principles are necessary to develop a simple, cohesive questionnaire in which the questions are clearly worded. Developing a good questionnaire is not a solitary task -- simply a matter of sitting down at your desk for a few minutes or even a few hours. There are a number of procedures that can be used to involve potential respondents in content or question development,

and to test and evaluate questionnaire drafts before they are finalized.

The purpose of Statistical Policy Working Paper #10, Approaches to Developing Questionnaires. is to provide practical information about these methods. The report contains descriptions of 11 different techniques, which can be used at various stages of questionnaire development. The report is structured in three parts: tools to develop questions, procedures for testing the questionnaire draft, and techniques used to evaluate the questionnaire draft. This structure was somewhat artificially imposed for ease of presentation in the report. In fact, there is no one ideal way to go about the process of developing a questionnaire. Depending on a number of factors, such as whether you're working from scratch or from an existing questionnaire, how much time and funds are available for survey development, these techniques can be used in many different combinations. In terms of improving the content of a survey questionnaire before it goes out into the field, the important thing is that testing and developmental work be conducted, not necessarily that it be done according to the structure presented in the report.

Having made this disclaimer, I am nevertheless going to discuss the techniques that are presented in the first two sections of the report -- that is, tools for developing questions and techniques for testing the questionnaire draft. I'm going to generally describe the methods contained in the report, and mention some additional techniques as well.

Developing Questionnaires

Part I of the report describes three tools for developing questions. The report presents these methods as useful in developing new questionnaires. I'd like to expand on this a little and suggest that these techniques can be used in the early stages of questionnaire development of any survey. Most surveys are conducted more than once; subsequent rounds of data collection begin with an existing questionnaire draft that is subject to

revision. These later rounds each have early stages of questionnaire development, complete with an existing questionnaire draft. In these cases too, the methods described in Part I of the report may be appropriate.

Unstructured individual interviews

Unstructured individual interviews are one-on-one conversations between a researcher and a member of the population for the survey or proposed survey. I use the term "conversations" because the discussion is unstructured; rather than having a set of specific questions, the researcher uses a topic outline that collects information on various aspects of these topics in whatever order, and using whatever terminology the respondent suggests. Respondents may also bring up additional issues related to the general topic, which might be incorporated into the topic outline for later interviews. The goal is an unstructured setting in which the researcher finds out how the respondent perceives the topic of interest, what terminology the respondent uses to talk about the topic, whether the respondent is knowledgeable and able to provide

information on the topic. By working from a blank slate, the researcher is not constrained by the content and terminology of an existing questionnaire, and the true frame of mind of the respondent is more likely to surface.

Qualitative Group Interviews

Many of you may be familiar with qualitative group interviews under a different name, such as focus group interviews, group depth interviews, or focussed discussion groups. Essentially these are unstructured interviews with a group of respondents rather than a single respondent, led by a group moderator. About 8 to 12 people participate in a group, and the moderator uses a topic outline to guide the discussion. Qualitative group interviews are used for many research purposes other than questionnaire development. When used to assist in questionnaire construction, the goal is the same as the goal of unstructured individual interviews -- to elicit the terminology used by respondents in thinking about the topic in question, to determine aspects of the topic that respondents consider important, and to get a reading on how respondents react

to aspects of the topic that survey planners consider important.

The difference between qualitative group interviews and unstructured individual interviews is, obviously, the group setting the diversity of opinions held by group members may stimulate interaction among them that elicits more information than could be obtained through interviews with each member separately. In order for these groups to be successful, however, the ability of the moderator is an important consideration. The idea is to stimulate discussion among all the participants and to avoid domination of the discussion by some people who may be more vocal than others.

Participant Observation

Participant observation is a technique that is used as an independent method of data collection, as well as a tool for questionnaire development. It has been extensively used around the

world. The basic elements of the technique are suitable for questionnaire design purposes, especially in developing questionnaires for use by members of other cultures or subcultures living within our own country. For example, the homeless population is a subculture that is currently the object of much interest, and for which the use of participant observation techniques is relevant. Indeed, these techniques have been successfully used in research on homelessness being conducted at the Census Bureau.

There are several distinguishing characteristics of participant observation research. First, the researcher must speak the respondents' language. This is not limited to English as opposed to a foreign language, but also refers to dialects, slang, or professional jargon. Second, the researcher associates with the members of the community he or she studies and engages in their activities. Ideally the researcher lives among the respondents; at a minimum, he or she develops contacts in the community over a long period of, time. The participant observer may also use the ethnographic interview technique during the course of his or her research. This involves using unstructured interviews (the

methodology I previously described) with "key informants." These are members of the community who are willing to talk at length with the researcher or introduce the researcher to other community members.

From this brief description, it should be obvious that participant observation is not a methodology that a person can "pick up" by reading an introductory textbook. The expertise required in the use of this technique dictates the involvement of trained ethnographers. While that may limit its use somewhat among U.S. statistical agencies, there are several ways it can be incorporated in a project. First, participation observation can be conducted as part of a project by trained anthropologists hired to serve on the project staff. In the homeless project I referred to a moment ago, we hired an anthropologist to work with a survey

methodologist, and this combination has worked out very well. A second way to make use of this technique is to consult with

ethnographers who have prior experience among the culture of interest, and take advantage of this previous experience rather than conducting original fieldwork. This could be done either by hiring the person on staff or doing it on a consultant basis.

Think Aloud Interviews

Another technique suitable for the early stages of questionnaire development has gained in popularity since the Working Paper was completed in 1983. This is the think aloud interview. Also referred to as protocol analysis, this method is an extremely valuable source of information about how respondents understand the survey questions put to them, and how they go about answering the questions. The purpose of the technique is to get respondents to talk out loud and verbalize their thoughts as they respond to questionnaire items. The data of interest here are respondents' reactions to the items, their thoughts as they formulate answers to the items, and what decisions they make in answering the questions.

Use of the technique requires a questionnaire draft. Since the results of these interviews are crucial to the questionnaire development process, the person doing the interviewing is generally a researcher or questionnaire designer. For interviewer-administered surveys, the questioner first explains to the respondent that rather than just answering the questions, he or she should actually think out loud -- that is, say what he/she is thinking as he/she answers each question. Respondents differ in their ability to verbalize their thoughts, and some may require a, bit of probing to uncover how they arrive at the answer to a question. At times it may take skillful questioning to probe completely what is on a respondent's mind. The interviews are generally tape-recorded (with the respondent's permission), since it is difficult to take notes and concentrate on probing the respondent's answers at the same time.

This technique can also be adapted for self-administered interviews. In this case, the questioner is basically an observer. The respondent is instructed to complete the questionnaire, reading

the questions and instructions out loud as well as verbalizing the responses. I've done quite a few of these interviews, and they really are quite helpful in detecting layout problems (not noticing skip instructions, etc.) in addition to uncovering problems with the questions.

This technique is used with relatively small numbers of respondents. Ten or fewer think aloud interviews provide large amounts of information and can uncover systematic misinterpretations or other problems. Use of the technique is an

iterative process -- once the questionnaire designer conducts five to ten think aloud interviews, problem areas will generally surface. Then, after revisions to the questionnaire are made, additional interviews can be conducted to detect problems with the revisions. Or alternatively, some other method can be used for the next round of questionnaire development.

Testing Questionnaires

Whatever methods are used to develop a questionnaire draft, it must be subjected to testing before it can be used in the field. There are a number of ways that this can be done, involving various levels of time and effort. Part II of the report mentions three techniques: informal testing, pilot studies, and split sample testing. I'11 describe each of these briefly and also add another selection to the menu.

Testing Multiple Questionnaires

In the questionnaire testing phase, the content of the questionnaire may be pretty much set except for fine tuning, or substantive questions may remain about how best to ask about a topic. When the latter is the case options that involve the testing of alternative questionnaires should be considered.

Experimental Group Session

The experimental group session is a small-scale method of testing alternative questionnaire versions, applicable only to the development of self-administered questionnaires. It may be conducted with respondents who are selected for their demographic characteristics and are not representative of any larger population.

In an experimental group session, respondents come to a central location (usually a large room containing tables or desks) for the purpose of completing a questionnaire. A group session is held, and 20-30 respondents participate at one time. The session is experimental, since more than one questionnaire version is randomly administered. A moderator conducts the session, and questionnaires are randomly distributed to the participants. After the questionnaire has been completed, a debriefing form may be administered to collect additional information about how the respondent interpreted specific questions. Multiple sessions are conducted until the total number of respondents is large enough

(about 500 or so) to facilitate statistical comparisons of the responses to the alternative questionnaires.

This methodology does not duplicate the response situation of a self-administered questionnaire where the respondent receives a form in the mail and returns it that way as well. For one thing, the respondent in the group session does not have access to other household members or to personal records, which may be necessary to answer some items. For another thing, once a respondent has joined the group, he or she generally completes the questionnaire and turns it in, while at home the questionnaire might remain unanswered. Despite these limitations, however, the methodology has definite advantages in the early stages of questionnaire development. It takes a relatively short time to arrange and conduct the sessions. If the statistical analysis is conducted quickly, it can provide rapid feedback about large differences in response to the alternative questionnaires, for use in later revisions of the survey instrument.

Split Sample Testing

In some situations, the questionnaire designer needs a large sample of respondents and a more formal test of different question wordings, question concepts, or methods of categorizing responses. This is particularly important in developing a major new survey instrument (such as SIPP when it was introduced several years ago), or in revising an existing questionnaire such as the decennial census form. When the nature of the survey requires large-scale testing of different versions of a questionnaire, the vehicle of choice is the split sample test.

Split sample testing, also referred to as split ballot or split panel testing, involves the use of multiple questionnaire variants, each administered to a portion of the sample. The entire questionnaire need not be different, but the alternative questionnaires should contain different versions of the items that are the focus of the test. In fact, the questionnaires should not contain too many differences, since all the variations in a questionnaire can affect response. One way to deal with this issue

is to limit the number of questions that are tested. Another might be to use automated data collection methods such as CATI or CAPI, which provide the means to randomize several experimental series of questions with respect to each other.

Once the content of the questionnaires is established, the alternative questionnaires are randomly distributed among the sample population, to decrease bias due to factors other than those being tested. The procedures for data collection are basically the same as for a survey containing a single questionnaire with one exception: control procedures must be established to I ensure that each sample case is assigned to the proper treatment group.

In a split sample test, the responses to the question variants that are the focus of the test are of keen interest. Thus,

statistical analysis of the data is an important aspect of the evaluation. In addition, observation of interviewers and

interviewer debriefing can be used for a personal interview survey, and information gained from these methods can help inform some of the statistical results of the analysis.

Testing Single Questionnaires

The final two techniques I would like to mention involve testing of questionnaires that are ready for "fine tuning." That is, major uncertainties about the content of a questionnaire do not exist, although changes in the instrument may be recommends base on the results of a field test

Informal Testing

Informal testing as its name implies, is a relatively casual method of evaluating a questionnaire. It is relatively small in scope, involving between 50 and 300 interviews. The cases for interview are selected purposively, rather than through any kind of systematic sampling methods. This may be accomplished by selecting participants from a broad range of subpopulation groups, in the

case of a test for a national survey, or limiting the participants to narrow population segments, if for example, a survey of food stamps or social security recipiency is being tested.

The informal nature of the test also carries over into the evaluation system. While some basic, quantitative information is calculated from the questionnaire responses, such as item nonresponse rates, and the number of "don't know" responses most of the evaluative information is based on observational feedback. There are several ways of obtaining this feedback. Observers, including the questionnaire designers, can accompany interviewers in the field, or in the case of telephone interviews, they can be tape-recorded. Specially-designed evaluation forms can be completed by interviewers and/or observers. Also, interviewers and observers can be debriefed after the interviewing is completed. Most of the information collected through these methods is subjective, based on the impressions of the staff present at the interviews. The informal testing procedure also allows unstructured discussion with the respondent at the end of the interview. In response to probing by the interviewer or observer

the respondent can provide information about his or her problems with the questionnaire, the meaning of specific items on the questionnaire, or other items of information. Observers who are involved in the survey as questionnaire designers, subject matter specialists, or sponsors may use their background knowledge to guide their probing and obtain useful information for evaluating the questionnaire.

Pilot Studies

In contrast to an informal test, pilot studies are much more formal and are conducted on a larger scale. Pilot studies are generally conducted further along in the questionnaire development cycle, and the goal is to duplicate the final survey design from beginning to end. This includes data collection from a larger sample, scientifically selected to represent the survey universe, and execution of data processing and perhaps tabulation procedures

as well. Needless to say, this is a lot more time-consuming than an informal test, and is not attempted until the questionnaire is in a final or nearly final state. Where the informal test seeks to uncover problems with terminology and question interpretation, at this point the questionnaire design issues of concern relate to how well the survey instrument performs in conjunction with the other aspects of the survey -- for example, errors in key codes or problems with response range categories for numerical data.

Evaluation of the results of a pilot study is much more quantitative than the analysis of an informal test. In a pilot study, the data capture, editing, and imputation programs are performed and, to the maximum extent possible, the data analysis plan is executed. This tests all the software developed for the survey and checks to see that the various stages of data processing are properly coordinated. Frequently, time constraints limit the amount of analysis conducted on pilot study data; however, the more effort is expended at this stage, the less likely you will be to find surprises when the survey is actually fielded. In addition to this formal evaluation of a pilot study, some less rigorous

evaluative tools are also used observation of interviewer training sessions is generally conducted and modifications to the training are suggested, as necessary. Also, observers may accompany interviewers for a personal visit survey, and both interviewers and observers are debriefed.

Another use of a pilot study might be to phase in a new questionnaire in a continuing survey. Rather than adopting the new questionnaire wholesale, overlapping samples can be designed, in which a portion of the respondents receive the new questionnaire and the rest receive the old one. The purpose here is not to test a questionnaire, but to collect information about the alternative questions and measurement strategies. The goal is to calibrate the old and the new questionnaires, to provide quantitative information about differences in response, which might affect the time series for the survey.

Discussion

The descriptions I've presented demonstrate the wide range of

options available in the questionnaire development process. The intent of the report was not to suggest that each of these should

be used in the development of a single questionnaire. Rather, we wanted to familiarize questionnaire designers with the techniques, to encourage their use, and to promote the value of testing questionnaires in general.

As I said at the beginning of this paper, there is no one ideal procedure to follow in preparing to field a survey instrument. It is generally best to start by talking to respondents, with or without a draft questionnaire, to find out if the vocabulary and the intent of questions are understood. Think aloud interviews, unstructured individual interviews, or other techniques that involve in-depth one-on-one discussions with respondents are extremely helpful here.

These techniques are not limited to the earliest stages of

questionnaire development, however. A draft questionnaire can be revised based on think aloud interviews, used in a field test, and the revised version also used in additional think aloud interviews. It is an iterative process that can continue as long as you find problems that need fixing.

Similarly, there is no magic formula for field testing an instrument. An informal test followed by a pilot study might be warranted based on the characteristics of the survey. Perhaps a series of informal tests might be considered for some complicated surveys. Or informal tests might be followed by a split sample test.

Depending on the circumstances of a particular survey and the time and budget allowed for survey development many possibilities are available. The important point is that testing facilitates problem detection, and fixing problems in a questionnaire will improve the quality of the data that is obtained.

TECHNIQUES FOR EVALUATING THE QUESTIONNAIRE DRAFT

Deborah H. Bercini

National Center for Health Statistics

This paper reviews the section of the report, Approaches to Developing Ouestionnaires, called "Techniques for Evaluating the Questionnaire Draft." How do these techniques differ from "Tools for Developing Questionnaires" and "Procedures for Testing the Questionnaire Draft" described in earlier sections of the report? In many cases, they do not. When there is a lot of time before a survey goes into the field, a particular method might be described as a tool for developing the questionnaire. When time has run out or the survey is already in the field, the same method would be referred to as a technique for evaluating the questionnaire draft. It doesn't really matter. In fact, the beauty of some of the techniques covered in this paper is that they can be adapted for use anywhere in the questionnaire design process.

What are these techniques? At first glance, the chapter headings shown in Figure 1 represent an apparently unrelated assortment of methods. However, there is a common thread. What links these techniques is that each uses an external source of information to evaluate the performance of the questionnaire. In this case, "external" refers to data that originates outside of the answers to the questionnaire items themselves. The first three techniques rely on the insights of the survey participants, that is respondents and interviewers. Next, observer evaluations are provided by an outsider who is not part of question-response process. The last technique, record checks, steps even further away from the interview or data collection situation by comparing questionnaire responses to an independent criterion, usually administrative records.

Approaches to Developing Ouestionnaires predated the emergence of laboratory or cognitive evaluation methods. The cognitive approach has provided a theoretical framework for understanding and reducing many kinds of response errors. Although this framework

was not in place in the early 1980's, some of the techniques that follow are similar, if not identical, to those used in today's labs.

Figure 1

Technique	Source
Frame-of-Reference Probing	Respondents
Response Analysis Surveys	Respondents
Interviewer Debriefings and Interviewer Questionnaires	Interviewers
Observation and Monitoring	Survey designers, etc.

Record Check Studies Administrative records

Frame-of-Reference Probing

Frame-of-reference probing is such a technique. This method evaluates the questionnaire by probing for how respondents understand key concepts, terms, definitions and instructions. The probes may be in the form of structured questions developed before the interview, or ad hoc, spontaneous questioning by an interviewer. Although this technique could be applied at any stage of questionnaire development, the report deals with it primarily in the context of field testing and the survey itself. The probes can be inserted after selected survey questions or they can be grouped at the end of the interview.

When probing is unstructured, it is usually done by the survey researchers or questionnaire designers because of their greater insight into question objectives. Standardized, structured probes can be administered by a field interviewer as part of the data collection process.

The use of frame-of-reference probing requires some planning. The first decision concerns when in the development process to probe. While probing will yield useful information at any time, clearly it will have the most impact when it is done with early drafts. If there are problems with fundamental concepts and key terms, it makes sense to detect them soon enough to work on a solution. If it turns out that respondents have difficulty with an entire topic or questionnaire approach, then modifications might be needed in the data objectives, not just question wording.

Structured probes, of course, need to be developed in advance. These may range from general, all purpose questions such as "What does so and so mean to you?" to more specific individualized probes. Even when probing is going to be of the unstructured or ad hoc variety, a plan for which questions and which terms to probe is advisable. In a field setting, respondents' time is limited and so are the number of probes that tan be asked. Based on previous

testing, researchers are likely to have some notion of potential trouble spots to cover. This information can be used to develop a protocol which specifies the criteria for probing.

No single evaluation technique is comprehensive or perfect. Each has its strengths and limitations. The particular strength of probing is that it can identify problems in the questionnaire that are often missed by methods that rely on respondents giving overt indications of difficulty. Consider the question, "During the past year, have you had pain in the abdomen?" This question was tested in the Questionnaire Design Research Laboratory at the National Center for Health Statistics. Most laboratory respondents answered it readily with a "yes" or a "no." It was not until interviewers probed for how respondents interpreted the term "abdomen" that it was discovered that very few respondents knew exactly where their abdomens were (1].

Probing also has the potential for identifying the underlying

causes of response problems, not just the fact that a problem exists. Returning to the example, the problem was variable interpretation of a key term. The underlying cause was lack of knowledge. When the cause of a question problem is understood, the solution is likely to suggest itself. In this case, the solution was a respondent flash card that showed an outline of the torso with the abdominal area shaded in.

Skilled probing is cost-effective. It can unearth quantities of information on how the questionnaire is working in a relatively short time. However, exclusive reliance on standardized probes tends not to produce very useful insights. Specialized probes require more time to develop but yield more valuable results. Also, if no ad hoc probes are used, unanticipated problems will be missed. The results of unstructured probing are, of course, subjective and anecdotal and require some skill to interpret.

Today, probing is one of the primary tools of the cognitive or laboratory approach to questionnaire evaluation. But its use is not limited to comprehension issues. Question laws related to

vague concepts or unfamiliar terms art certainly common, but they are by no means the only sources of response error. Probing techniques can be used effectively to detect question problems that affect many components of the response process. These include recall, estimation, judgement, decision making and motivational factors (2].

Probing and other intensive interviewing methods have now evolved into a major and separate phase of the questionnaire design process that usually precedes the field testing phase. This approach gives questionnaire designers freedom to explore the response process in depth. Constraints on probing are significant when it is done during a field interview that was designed for another purpose.

Response Analysis Surveys

The response analysis survey (RAS), like frame-of-reference

probing, evaluates the questionnaire from the respondent's

perspective. The report describes it as a technique used to evaluate mail surveys, especially mail surveys of establishments. In effect, the response analysis survey is a survey about a survey in which personal interviews are conducted with a sample of mail survey respondents. Interviewers administer a structured questionnaire which asks questions about how respondents would go about answering the mail survey questions.

The typical RAS would collect data on how establishment records are maintained, what kinds of information they contain, how difficult it is to retrieve this information and so on. It also attempts to find out if respondents can understand what is being asked of them, their willingness to provide the information and other aspects of respondent burden.

If the RAS is being conducted to evaluate an on-going survey or to prepare for the next cycle of a periodic survey, researchers use available data on response errors as a guide when developing

questions for the RAS. Data collection and analysis proceed as in a regular survey. The results are interpreted and then used to redesign the questionnaire for the next mail survey.

The strengths and limitations of the RAS parallel those of frame-of-reference probing. The RAS asks the respondent to analyze the response task, and, in doing so, both techniques are capable of detecting covert response problems and their underlying causes. And the RAS, like probing, loses some of its potential when only structured probes are used. On the other hand, a formal response analysis survey will produce valuable, objective data that will reliably indicate where questionnaire revisions are needed.

Variations or adaptations of the RAS concept can be used to, evaluate any self-administered questionnaire, not just mail questionnaires for establishment surveys. Laboratories at several government agencies test self-administered questionnaires using a combination of observation, think-aloud, and structured and unstructured probing methods. The Government Accounting Office, for example, uses an interesting combination of observation and

laboratory-style probing. They watch respondents complete the questionnaire noting all kinds of non-verbal behavior such as sighs, grunts, head shaking and other signs of impatience, skipped questions, and so on. Afterwards, the interviewer goes back to each of the questions that provoked the reaction and asks the respondent to elaborate.

Learning from Interviewers: Interviewer Debriefings and Questionnaires

No assessment of interview survey questionnaires is complete without the interviewer's input. The report chapter, "Learning from Interviewers," presents two techniques for gathering information from interviewers - interviewer debriefings and structured post-interview evaluations. The latter are questionnaires completed by interviewers about various features of the survey questionnaire. Either method can be employed with

pretest questionnaires or during an on-going survey.

The interviewer debriefing session is a forum in which interviewers can relate their experiences in administering a questionnaire or data collection procedure. Their scope and formality can vary with the scope and formality of the testing operation that they accompany. In large field tests involving many interviewers, a single comprehensive debriefing is usually held when interviewing is completed. With more informal testing, it is possible to conduct multiple debriefings throughout the pretest period. Questionnaires can be revised on the spot, tested, and then revised again.

Interviewer questionnaires can take several forms also. They can be directed to a specific issue or, problem, such as nonresponse. Or they can consist of questions designed to get at suspected difficulties with particular survey items. Another, format is a questionnaire made up of standardized ratings that interviewers apply to each survey item.

Any survey planner who ignores what interviewers have to say

about the questionnaire is taking a great risk. Interviewers are in the best position to comment on how the questionnaire and other survey procedures affect respondent cooperation. For the most part interviewer performance is judged on response rates and completion rates, so interviewers will naturally be sensitive to factors that affect performance in these areas. Interviewers also have excellent insights into the logistics of questionnaire administration and are quick to spot things that impede the efficient flow of the interview. If survey planners, for whatever reason, do not heed interviewers' major objections, they will pay a price. A questionnaire that interviewers find unnecessarily difficult to administer will lead to poor interviewer performance, and therefore, lower data quality.

There are also limitations to what can be learned from interviewers. Interviewers are often more adept at making bad questions work than they are at finding flaws, unless the question is so bad that the interviewer can't figure out how to ask it and the respondent can't or won't answer it. The interviewer's job is to get a response and they are good at it. They are less likely to

notice subtle problems with question wording, interpretation, and so on, as long as the respondent gives a codable response.

Reliance on post field test interviewer debriefings to detect question problems is a poor evaluation strategy. Only those problems that visibly disrupt the interview will be mentioned. And as in any group situation, the most vocal will dominate. It can be difficult to achieve concensus as to what the problems are because interviewers can have varied experiences depending upon the sample cases they have interviewed. Using interviewer evaluation questionnaires to supplement the debriefing can compensate for some of these drawbacks.

Getting interviewer input need not be confined to the field situation. It is possible to ask interviewers to evaluate draft questionnaires before they ate field tested. This can be done in a laboratory setting with "real" respondents or with researcher

respondents. Although subject to some of the same sorts of limitations mentioned above, it may be possible for interviewers to identify some flaws in this way. It could be especially useful to ask interviewers to try out questionnaires that have been adapted to new data collection modes, such as CAPI, for example.

Observation and Monitoring

Observation of face to face interviews or monitoring telephone interviews-evaluates the questionnaire from the perspective of a third party. Observers are usually people involved in the survey planning process, from sponsors and subject matter experts to questionnaire designers and data analysts. At NCHS, observation usually takes place during field pretests, but the same methods could be used to evaluate on-going surveys. Most often, an observation program provides a qualitative, subjective assessment of questionnaire performance and related communications. An infrequently used, but more objective approach, is known as behavior or interaction coding in which standardized codes are used to evaluate question performance.

In all cases, observers need some preparation for their task. Attending the interviewer training helps as well as some coaching on specific situations or problems to watch out for. Observer forms can be useful, provided observers are not so busy recording minor detail that they miss more significant interactions. At the end of the testing period, observers may submit a written report summarizing their experiences, participate in a debriefing session, or both.

Sometimes observers can become active participants if they are also there as frame-of-reference probers. At the other extreme are the behavior coders who usually work from taped interviews.

The Survey Research Center at the University of Michigan has recently completed a study on pretesting techniques, among them, behavior coding. The goal of the study was to test techniques that would enhance the usefulness of the traditional field pretest. The

Michigan study used the coding scheme shown in Figure 2 to identify interviewer and respondent behaviors that are symptomatic of problem questions. Trained coders listen to taped interviews and apply the appropriate codes to each question. The numbers and types of codes for each question are, then tallied. The benefit of this technique, according to the study, is that is can provide objective indicators of flawed questions [3].

Figure 2

Interviewer Behavior	Respondent Behavior
Reads question with slight changes with answer	Interrupts question
Reads questions with major changes or does not complete it	Requests clarification or repeat of question

Gives qualified but adequate answer

Gives inadequate answer

Gives "Don't Know" answer

Refuses to answer

What are the strengths of the observation technique? It can readily be incorporated into existing pretest plant. Third party observers can interpret the interviewer-respondent dynamic in a way that the participants cannot. And for many survey sponsors and planners, actually seeing how the questionnaire performs in the field is the most convincing evidence that changes need to be made.

The limitation of the observation technique is evident from the name - it can only detect observable questionnaire problems. And when a problem is observed, the underlying cause may not be

obvious. Individual observers will have limited experience with the questionnaire unless they have observed a great many interviews. Therefore, agreement on what the problems are may be difficult to achieve.

Record Check Studies

Record check studies are used not so much to evaluate question wording, but to evaluate the validity of the data that is produced by the questionnaire as a whole. The accuracy of responses on a particular topic are checked against an independent criterion, usually administrative records. For example, data from a health survey that asks questions on doctor visits could be compared to respondents' records maintained by their health care providers. In these studies, it is assumed that the administrative record represents the "truth." Respondent reports that do not correspond

to record data are counted as errors and a high error rate would indicate that something is wrong with questionnaire or approach to data collection.

Record check studies pose numerous logistical challenges. One needs to obtain the cooperation of a records source. Preservation of confidentiality is often a problem. The structure and quality of the record system needs to be studied. Is it adequate? Matching criteria must be developed. Does the record system support the level of matching that is desired? Matching questionnaire data to record data is invariably more difficult than anticipated, and many discrepancies have to be resolved. Finally, the results require thoughtful interpretation. What are the implications for the questionnaire? What is it about the questionnaire or other aspects of the survey design that are contributing to response errors?

Record check studies can provide objective evidence that a questionnaire is collecting the information it is designed to collect. However, they can only be used to evaluate questionnaires on topics for which independent records are available. Clearly,

there are many types of human behavior that interest researchers for which no records exist. Compared to the other evaluation methods described above, record checks are relatively time consuming and costly. But the costs have to be weighed against the benefits. For large, expensive surveys where data precision is critical, evaluation by record check would make sense.

Several variations of the record check study are possible. Seeding the pretest sample with cases known to possess a target characteristic is a scaled down version of a record check study. It is not too difficult to implement if the characteristic is a simple one and if it is not highly sensitive, having arthritis versus having AIDS, for example. Methodological studies can use other validation sources besides administrative records. Some possibilities are respondent diaries, data collected from other family members on the same topic, biochemical markers, medical exams and so forth.

Conclusions

When there are so many ways to find out if a questionnaire is performing as intended, there is no good reason not to do it. Several of the techniques in the report are conducted in conjunction with field testing, so that time and cost factors are marginal. Probing techniques can be applied in so many different ways and at different levels of intensity that the technique can be adapted to almost any evaluation objective or questionnaire type. Laboratory facilities are advantageous, but not essential.

It should be evident that no single technique will tell you all you need to know about the adequacy of your questionnaire. An evaluation program that includes, several different sources of information on question performance will be the most successful.

References

[1] Bercini, Deborah H. "Pretesting Questionnaires in the

Laboratory: An Alternative Approach." (accepted for publication)

Toxicology and public Health.

[2] Royston, Patricia N. "Using Intensive Interviews to Evaluate

Questions." Proceedings of the Fifth Conference on Health Survey

Methods Research, Keystone, Colorado, 1989.

[3] Cannell, Charles et. al. "New Techniques for Pretesting

Survey Questions. NCHSR #HS 05616, Survey Research Center,

University of Michigan, 1989.

DESIGNING QUESTIONNAIRES FOR CATI IN A MIXED MODE ENVIRONMENT

Gemma Furno

U. S. Bureau of the Census

1. Introduction

The use of computer assisted data collection by Federal statistical agencies has increased dramatically since Approaches to Developing Questionnaires (Statistical Policy Working Paper No. 10) was written in the early 1980's. Utilization of computer assisted telephone interviewing, or CATI, is now commonplace in many agencies.

CATI is an interactive system whereby the questions appear on a terminal screen and the interviewer keys the answers directly into the computer. Branching paths are programmed into the system and the next appropriate question is automatically presented. Range and consistency edits can be programmed to allow for on-line editing of data. These telephone interviews are conducted from one or more centralized locations.

At the Census Bureau, CATI is often used for demographic surveys in conjunction with field personal visit and decentralized

telephone interviews which use a paper and pencil questionnaire. Typically, some portion of households interviewed previously are assigned to CATI in this mixed mode environment. How many are assigned depends on several factors, such as the sample design itself and optimum workloads for both the field and the centralized telephone facility. Personal interviews are reserved for first time contacts where a visit to establish rapport has been found beneficial, and to follow up cases, such as unable to contact and refusals, that could not be completed on CATI. Telephone interviewing from the interviewer's home is often used for returning cases not assigned to CATI.

This description fits the current usage of CATI in the national sample of the American Housing Survey, known as AHS. The American Housing Survey is conducted every two years. CATI was first introduced in 1987 and its use was expanded in 1989, when approximately 25 percent of the sample was initially assigned to CATI. The AHS questionnaire is lengthy and complex, containing over 125 main items, in addition to the household roster. The average interview time is approximately 30 minutes.

This paper describes our experiences collecting data on the American Housing Survey using both CATI and a paper and pencil questionnaire from the perspective of CATI questionnaire design. A summary of several data quality issues will be presented, followed by a discussion of issues encountered in designing an AHS

CATI questionnaire that was comparable to the paper version and which minimized any problems when the two data sets were merged.

2. Issues of Data Quality

Computer assisted data collection holds the promise of improving data quality in several areas. Those cited in the literature [2,5,6,8,9,12] that are most directly related to questionnaire design include the ability to:

1. control branching paths, thus helping to ensure that the correct questions are asked;

2. tailor question wording to the specific situation, thus relieving the interviewer from the burden of choosing alternative wordings, and helping to ensure that the questions are always worded correctly; and

3. evaluate the answers given for appropriateness and take corrective action through the use of on-line range and consistency edits, scripted probing, and dependent or reconciliation interviewing (using answers obtained in previous interviews to improve present answers).

Intuitively one would think that these capabilities should improve data quality. But much research still needs to be done to prove that it actually does, and to quantify the improvement [6,9]. It clearly has been shown that controlling branching paths does ensure that the appropriate questions are presented on the screen.

But this alone does not guarantee that the interviewer actually reads the question as worded, receives an acceptable answer or enters it correctly [1,5,6,9]. In reality, entries of "don't know" or "refused" are allowed for most items, range and consistency edits cannot catch all respondent or interviewer errors, and scripted probing and dependent or reconciliation interviewing have practical limitations.

Data from the AHS preedit reject operation illustrate these points. In AHS, the field preedit operation is designed to identify and correct certain clerical, keying and consistency errors in order to improve control of the sample and the quality of survey data before it goes through the regular range, consistency and blanking edits. Approximately half of these reject reasons involve consistency checks within the household roster. The CATI data was put through the preedit program to help evaluate and ensure its quality. The results of the preedit operation for 1989 show that 11.6 percent of the 8,794 completed CATI cases rejected with an average of 1.20 rejects per case, while 42.6 percent of the 49,279 cases completed on the paper and pencil form rejected with an average of 1.76 rejects per case.

In addition to vastly improving the control of the sample (no rejects for duplicate records, mistakes in the control number, sample status etc.), results for CATI also show that most reject reasons related to the questionnaire were reduced or eliminated. For example, missing data on several key items such as type of living quarters, roster line number, relationship code, reference person, and heating equipment were eliminated due to the automatic branching feature. Where consistency checks were programmed into the CATI questionnaire to identify roster errors such as inconsistencies between birthdate and age, two spouses recorded for the same person, unmarried person has a spouse, etc., the corresponding preedit reject was greatly reduced. However, where roster consistency checks were not in place, some errors remained. Time constraints and the size of the CATI questionnaire have prevented programming all the appropriate checks. The practicality of adding more of these checks will be investigated.

For other reject reasons not showing improvement, we found a number of explanations. For example, preliminary review indicates that for one reject the CATI interviewers accepted "don't know" entries at a much higher rate than the field interviewers. The item asked for the number of units in a multiunit building. This may well be related to telephone interviewing in general rather than CATI per se. However, now that the problem has been discovered, better interviewer training and/or a scripted probe for a "don't know" answer could be added to CATI for this item. In another instance, a high rate for a reject reason disclosed a flaw in the CATI questionnaire, which will be corrected.

Adding more roster checks and a careful review of the other reject reasons should lead to improvements that will further lower the number of CATI cases rejecting in the preedit operation, although some interviewer and respondent errors will inevitably remain.

Another indicator of quality in the AHS CATI data involves a

reconciliation study conducted for selected items. In the 1987 interview, these items were tenure, type of basement, number of bedrooms and bathrooms, heating fuel, heating equipment, rent and home value. If the 1987 CATI response failed certain tolerance limits, compared to the 1985 response, then the answers were probed at the end of the interview to discover the reason for the discrepancy.

Of the 6,432 cases completed on CATI in 1987, 54.8 percent failed the comparison on at least one of the items, triggering the reconciliation questions [11,13]. For all the items reconciled, an average of 49% of the respondents reported some plausible explanation for the discrepancies between the two survey periods. For example, a half bath was converted to a full bath, a different type of heating equipment has been installed, local real estate conditions affected rent or house values, etc.

However, that left an average of 51 percent of the respondents

reporting that the response was incorrect in either 1985 or 1987, thus the change between survey years was spurious. (One caveat is that some of these cases may actually represent real status changes that were incorrectly classified due to response error in the reconciliation question itself.) An interesting result it that respondents were almost as likely to point the finger at the answer they had just given in the 1987 interview as the 1985 one. Forty-nine percent said the 87 answer was wrong compared to fifty-one percent for the 85 answer. But the reconciliation questions did not attempt to ascertain why the 87 response was wrong - did the interviewer read the question incorrectly, did the respondent or interviewer misunderstand or did the interviewer enter it incorrectly? This result offers another reminder that a CATI questionnaire may offer the potential to improve data quality in some areas, but it is not a panacea. Future AHS reconciliation studies will try to better ascertain the cause of these errors.

3. Issues Encountered When Designing the AHS CATI Questionnaire

When designing a CATI questionnaire to be used in a mixed mode environment with a paper and pencil questionnaire, the paper form

and its associated procesting system has usually been in use for a number of years. In such applications, the CATI questionnaire generally is expected to conform to yield comparable data and expeditious processings of data from both modes.

In this situation, the CATI questionnaire has to serve "two masters". First, it should satisfy the basic objectives of CATI questionnaire design. For example, House and Nicholls stress that a CATI questionnaire must conform to the general accepted standards of questionnaire design while functioning as a complex computer program [7,8,10]. The program must ensure that the questions work correctly under all circumstances and that minimum demands are made on hardware resources while maintaining rapid response times. But secondly, in a mixed mode environment, a CATI questionnaire must meet these requirements while providing comparability with an existing paper and pencil version and minimizing any problems encountered when the data is processed. Usually the CATI data is reformatted, then merged at some point with the data collected on the paper form and processed through the existing system. A single processing system saves time and money and ensures that any complex

edit and data imputation/allocation procedures are consistently applied, regardless of collection method.

A. Numbering of CATI Questions and System Commands/Instructions

A basic issue of CATI questionnaire design is the numbering scheme used for the questions and system commands/instructions. This can have important implications when a complex questionnaire

is used in a mixed mode environment. Two possibilities are to utilize the actual question numbers, or if different, any processing code numbers.

The AHS questionnaire contains several sections of duplicate or parallel items because there are different sections of the questionnaire for renters and owners. Within both of these two major subsections there are several further subsets of parallel items based on type of housing unit. All questions on the paper

form have a unique item number, but a duplicate or parallel item shares the same processing code as only one version of the item could be asked in an interview.

The question arose as to how to handle these duplicate or parallel sets of questions in CATI - should the basic question be programmed only once and the system programmed to alter the question wording and its universe as appropriate, or should the design of the CATI questionnaire follow the paper form as closely as possible? We chose the latter course for AHS, that is, to follow the paper form as closely as possible, and thus utilize the item numbers rather than the processing codes. There were two reasons for this. First, the universes for the duplicate or parallel items are extremely complicated in AHS. Entering the basic question only once and programming the system to alter the wording and universes as needed would not have saved the CATI author any time as the system instructions and documentation would have just become more complicated, prone to error and difficult to test out. Secondly, with both CATI and paper questionnaires in use simultaneously, and separate training materials to be written, our

goal was to move easily from one questionnaire and set of materials to the other without confusion.

This compatibility between questionnaires proved especially helpful when it came to writing the specifications and programs to reformat the CATI output to be merged with the paper and pencil data.

B. Question Wording, Fills and Answer Categories

We encountered little difficulty in transferring the actual question wording to CATI. The paper and pencil form had already been adapted for telephone use in the field several years before. However, a few problems had to be dealt with.

When collecting data under both modes, it becomes difficult to change the wording close to the start of the survey if you want to keep the questions as comparable as possible. For example, the CATI interviewers found the wording of one question particularly awkward but it could not be easily changed because the paper form

was already printed. The sponsor did not feel comfortable changing the wording on only the CATI questionnaire. A revision had to wait

until the next time the survey was conducted when it was made to both.

Other situations involved question "fills." That is, using information previously obtained to tailor the exact wording to the situation. This is one of the jobs that CATI does best, but to display the question correctly, the system obviously must be programmed to distinguish among the wording choices. This sometimes required that answer categories be expanded. For example in Figure 1 below, the paper and pencil version of item 120g (on means of transportation) groups "cars, truck and van" into a single answer category. The interviewer substitutes the specific response in the following question, 120h. In CATI, three separate categories are required if subsequent questions are to use the answer given.

Figure I

Q120g - 120h On The Paper & Pencil Questionnaire

Figure 1 also illustrates the situation where a subquestion on the paper form is imbedded in the middle of the main question with

the answer categories numbered as if it is one continuous question. Figure 2 below shows what this series looks like in CATI. In CATI, the subquestion appears on a separate screen but the answer categories are numbered 1 and 2 instead of 2 and 3. CATI interviewers are used to seeing the categories in numerical order,

starting with one. It would have been confusing to present the question on a separate screen with the categories numbered any other way.

Figure 2

Q120g - 120h In CATI (Answered For Mary Smith)

>Q120g< How did MARY SMITH usually get to work Last week?

(MARK ITEM THAT ACCOUNTED FOR GREATEST DISTANCE TO LOCATION OF JOB AT WHICH PERSCN WORKED MOST HOURS LAST WEEK.)

<1> Car

<2> Truck

<3> Van

<4> Bus or streetcar

<5> Subway or elevated

<6> Railroad

<7> Taxicab

<8> Motorcycle

<9> Bicycle

<10> Other vehicle

<11> Walked only

<12> Works at home

===> 3   (System is programmed to
display Q120g 1 if 1-3
is answered here.)

>Q120g1< Did MARY SMITH drive alone or go with other?

<1> Alone

<2> Go with others

===> 2   (System is programmed to
display Q120h if 2 is
answered here.)

>Q120h< How many people including MARY SMITH usually ride in the van?

<01-97> 1-97

<98> 98 or more

===>7

C. The Reformattng Stage

Our processing goal for the CATI questionnaire was to produce

output that could easily be reformatted to look exactly like the data keyed from the paper and pencil form. This allowed for merging the two data sets at the earliest possible opportunity and running the combined file through the current processing system. We encountered several situations that affected the reformatting operation.

1. Reformatting the CATI Data

When the CATI question and answer categories closely corresponded to those on the paper form, reformatting was straight forward. However, if it did not, then more complicated reformat specifications had to be followed. Some simple examples of more complicated situations include questions where, as seen in figures 1 and 2, expanded CATI answer categories had to be collapsed back or separate CATI questions had to be combined to match the compact style and embedded questions of the paper form. Although some of this reformatting was performed in the CATI questionnaire itself, most was completed in batch mode after data collection. Performing most of these tasks later reduced the number of variables in the

CATI questionnaire, an important consideration since the questionnaire was quite large. While the Census CATI system does allow for very large and complex questionnaires, there are system limits which the AHS questionnaire reached.

Table and grid formats also frequently required reformatting. The CATI author had to ensure that these questions were programmed in such a way that the data could be successfully reformatted later.

2. Adding Special Data Items Not Needed for CATI Questionnaire

Not all items on the paper and pencil questionnaire are needed or are relevant to the CATI questionnaire. Some are standard housekeeping input items, such as sample designations or geography codes that can't be changed during the interview. Others are output items that summarize respondent or housing unit characteristics from previously answered items. The question arises as to whether such items should be carried on the CATI input or output files or merely added in batch mode when the CATI data

are reformatted.

In deciding which course to follow, we gave consideration to balancing the size and efficiency of the CATI questionnaire with the difficulty of adding these items later. The system size constraint and the ease of taking care of these items dictated that

we add them during reformatting while taking special care not to overlook relevant items.

As can be seen from these examples, complex questionnaires and processing needs often require special effort to ensure comparability of the CATI data to that collected on the paper form.

4. Summary

While great improvement in data quality was seen as a result of automatic branching and on-line consistency checks, there is

still room for improvement in the AHS CATI questionnaire. It must also be remembered that the questionnaire can not shoulder all the burden for improving data quality. Respondent and interviewer actions over the telephone, whether in a centralized or decentralized environment, must also be considered.

Our experiences on AHS showed a large, complex paper and pencil questionnaire successfully transferred to CATI. Constant effort by a number of people, including the specifications writer, the CATI author, and other programmers was needed to accomplish the task.

In AHS, it was the CATI questionnaire that was expected to conform to the paper and pencil form and the current processing system. This does not always allow CATI to be used to its fullest potential but it is a common situation that must be faced when CATI is added to an already existing paper and pencil survey. As new surveys are developed with a mixed mode of data collection planned from the beginning, different experiences will result.

References

1. Catlin, G. and Ingram, S. (1988), "The Effect of CATI on Data Quality: A Comparison of CATI and Paper Methods", Proceedings of the Fourth Annual Research Conference, U. S. Bureau of the Census, pp. 291-299.

2. Dillman, D. and Tarnai, J. (1988), "Administrative Issues in Mixed Mode Surveys," in Telephone Survey Methodology, Robert G. Graves et al, (editors), John Wiley & Sons, New York, NY.

3. Ferrari, P. (1984), "Preliminary Results from the Evaluation of the CATI Test for the 1982 National Survey of Natural Scientists and Engineers", unpublished research report, U.S. Bureau of the Census.

4. Ferrari, P. (1986), "An Evaluation of Computer-Assisted Telephone Interviewing Used During the 1982 Census of Agriculture,

unpublished report, Agriculture Division, U. S. Bureau Census.

5. Groves, R. and Mathiowetz, N. (1984), Computer-Assisted Interviewing: Effects on Interviewers and Respondents", Public opinion Quarterly, 48(1B), pp. 356-3691.

6. Groves, R., and Nicholls, W. (1986), "The Status of Computer-Assisted Telephone Interviewing: Part II - Data Quality Issues, "Journal of Official Statistics, 2(2), pp. 117-134.

7. House, C. (1985), "Questionnaire Design With Computer-Assisted Telephone Interviewing", Journal of Official Statistics, 1(2), pp. 209-219.

8. House, C. and Nicholls, W. (1988), "Questionnaire Design for CATI: Objectives and Methods", in Telephone Survey Methodology, Robert G. Groves et al, (editors), John Wiley & Sons, New York, NY.

9. Nicholls, W. and Groves, R. (1986), "The Status of

Computer-Assisted Telephone Interviewing: Part I - Introduction and Impact on Cost and Timeliness of Survey Data", Journal of Official Statistics, 2(2), pp. 93-115.

10. Nicholls, W. and House, C. (1987), "Designing Questionnaires for Computer-Assisted Interviewing: A Focus on Program Correctness", Proceedings of the Third Annual Research Conference, U.S. Bureau of the Census, pp. 95-1ll.

11. Nicholls, W. (1989), "The Impact of High Technology on Data Collection", CATI Research Report No. Gen-1, Computer-Assisted Interviewing Central Planning Committee, U. S. Bureau of the Census.

12. Van Bastelaer, A., Kerssemaklers, F. and Sikkel, D. (1988), "Data Collection With Hand-Held Computers: Contributions to Questionnaire Design", Journal of Official Statistics, 4(2), pp. 141-154.

13. Schwanz, D., Montfort, E. and Cannon, J. (1988), "Analysis of Operational Issues: 1987 AHS-CATI", CATI Research Report No.

AHS-1, Computer-Assisted Interviewing Central Planning Committee, CATI Research and Analysis Sub-Committee, U. S. Bureau of the Census.

DISCUSSION

Carol C. House

National Agricultural Statistics Service

Dillman and Tarnai (1988) define a mixed mode survey as one that uses two or more methods to collect data for a single data set which will be analyzed as a unit. Familiar examples include face-

to-face first wave interviews followed by telephone or mail on subsequent waves; and telephone follow-up to a mailed questionnaire. Most large Federal statistical agencies routinely use mixed mode surveys to collect data. The Furno paper focuses on the 1987 American Housing Survey which uses a combination of face-to-face, decentralized telephone, and CATI interviews in its mixed mode design. The National Agricultural Statistics Service (NASS) sometimes incorporates mail, centralized (non-CATI) telephone, CATI, face-to-face, and decentralized telephone interviews in a single survey.

Why are survey organizations choosing mixed mode designs over simpler single mode surveys? Their objectives appear to be to reduce survey costs, improve timeliness, and to take advantage of the relative strengths of different modes of collection. At the same time they want to preserve data comparability and integrate the mixed design into the data collection, data handling and data manipulation processes that are already in-place in the organization. This over simplifies the decision making process, but it fits most large Federal agencies.

I will use this set of objectives as a basis to evaluate the design of the CATI questionnaire discussed in the Furno paper. This paper describes adding CATI to an existing mixed mode survey featuring face-to-face and decentralized telephone interviewing. The CATI questionnaire was designed to work effectively in that specific environment. The author discusses issues related to question wording , editing and data processing because a CATI questionnaire design impacts all of these areas.

Cost Reductions and Improvements in Timeliness

The Census Bureau probably achieved most of the gains in these areas when they originally mixed face-to-face interviewing with decentralized telephoning. They may see additional cost savings by adding CATI to this mix, but any such gains are likely to be minimal. However, a specific discussion of cost were beyond the scope of this paper.

The literature usually asserts that timeliness can be improved

by CATI because the data is immediately entered into a computer without a separate data entry step. Secondly, and more

importantly, data is edited during the interview and is "clean" by the time the data collection is over. But what actually happens to CATI data from the American Housing Survey? Furno reports that, "Our processing goal for the CATI questionnaire was to produce output that could be easily reformatted to look exactly like data keyed from the paper and pencil form... and [to] run the combined file through the current processing system." Thus the "clean" data from CATI is dumped in with "dirty" data from other data collection modes and the whole file is run through the standard batch editing programs. This results in no improvements in timeliness.

Her approach is not uncommon. The goal of easy assimilation frequently takes precedence over improvements in timeliness as well as several other objectives of the mixed mode design. This decision may be necessary during early experimental uses of CATI,

but as CATI (and soon CAPI) become ongoing parts of a survey organization, we need to find ways to integrate CATI data into data processing programs with less duplication of effort.

Tapping the Strengths of Different Data Collection Modes

The American Housing Survey's instrument incorporates a number of CATI features to improve data quality. These include controlling the branching paths, tailoring question wording to the individual respondent, and online editing. Furno measures the gains in some of these areas through two different comparisons: by counting the number of rejects in a subsequent, "pre-edit" program; and by conducting a reconciliation study.

The pre-edit program is designed to make simple checks or clerical and keypunch errors prior to the data entering more sophisticated and complex editing Programs. Furno measures 43% rejects on the paper versions and 12% rejects on CATI. This demonstrates substantial improvements using CATI. However, one wonders why there should be any of these very simple errors in the

CATI data. The author indicates that not all of the checks were added to the CATI instrument, so this is an area for possible improvement in the questionnaire.

The reconciliation study was conducted at the end of the interview for selected items, comparing responses with corresponding answers obtained during the 1985 survey. These items (such as number of bedrooms in the house) were expected to be fairly constant over the two years. This study uncovered reporting errors on both the 1985 and 1987 surveys, with approximately the same number of errors occurring each year. This fact indicates that the CATI questionnaire (new in 1987) did not significantly improve the data quality in some areas. More detailed studies of this type will possibly uncover the causes of these errors and lead to improvements in both the CATI And paper questionnaires.

The reconciliation study was kept completely independent of

the main part of the questionnaire. The original responses on the survey were not changed based on reconciliation, although CATI technology would have made it easy to do so. This brings up a broader issue to consider for panel survey: is it appropriate to use previously collected data to edit or influence current responses? Is this practice any less appropriate on mixed mode surveys where certain modes (CATI) would use this earlier information and other modes would not? It is unclear whether using previously collected information would improve overall data quality or merely heighten inconsistency and variability in the error structure of a mixed mode survey. NASS is struggling with these issues and we would appreciate reaction and experiences from other groups.

Keeping Data Comparable Across Modes

This was one of the primary objectives of the designers of the American Housing Survey's CATI instrument and the Furno paper concentrates on these issues. Discussions include the ways to program questions on CATI that appear in tabular form on a paper

form; handling fills; using consistent answer codes; and handling last minute questionnaire changes.

Integrating A Mixed Mode Design Into Existing Survey Processes

When new technologies or new modes of data collection are added to an on-going survey it is important to cause as little disruption to the routine as possible. This was the situation with the CATI test on the American Housing survey, and the Furno paper describes the efforts to which the designers went to make CATI fit into the existing design. The CATI version of the questionnaire was always made to conform to the paper version, and the CATI processing and editing to conform to, and go through the existing batch programs. The disadvantage of this approach is that some (much?) of the advantages from CATI were lost in the mixed mode design.

CATI is here to stay in telephone surveys and CAPI is just arriving. These technologies will be used routinely in mixed mode designs. How do we handle the integration of CATI and paper once

the testing phase is over? Although it may be reasonable to make a CATI questionnaire conform to a paper version when early testing is going on, it is not reasonable to retain that unbalanced relationship later after 75% to 80% of the Contacts are made with the CATI version. This situation can and does happen, because the test version is implemented operationally with minimal revisions. It is time to re-evaluate CATI/CAPI technology in mixed mode designs. The modes must fit together into a single survey operation and produce compatible data. However, we need to look

for better ways of integrating existing technologies with the new so that total quality is optimized.

Reference

Dillman, Don A. and Tarnai, John, "Administrative Issues in Mixed Mode Surveys," in Robert M. Groves (ed.), Telephone Survey Methodology, New York, Wi1ey, 1988.

Session 1l

STATISTICAL DISCLOSURE - AVOIDANCE

DISCLOSURE AVOIDANCE PRACTICES AT THE CENSUS BUREAU

Brian Greenberg

U. S. Bureau of the Census

I. Introduction

The Census Bureau, as well as other statistical agencies, collects information about the Nation's population and institutions and releases this information to the public. The information is typically collected under pledges of confidentiality and agencies are required to release data in such a manner so as not to violate guarantees of non-disclosure either through design or neglect. At the same time, date collection agencies have the responsibility to make statistical information available for a wide range of uses that include policy decision making, program analysis, economic modeling, and many others. A data collection agency has the obligation to release as much information to the public as possible while adhering to pledges of confidentiality given to respondents. Broadly speaking, the objective is to release as much information as possible consistent with the requirement that the risk of disclosure is acceptably low.

There is no known way to quantify the amount of information

released or to quantify level of risk of disclosure. Finding methods to relate the levels of information and levels of risk is an area of very active research at the Census Bureau and at other statistical agencies in this country and abroad. In a recent paper and talk at the Census Bureau 1990 Annual Research Conference (Greenberg 1990), I discussed disclosure avoidance research activities at the Census Bureau. The report focused on the work to develop data release strategies through the use of tools of operations research, mathematics, and statistics. We discussed research efforts here at the Bureau, studies conducted under Joint Statistical Agreements, and other cooperative efforts with researchers on this topic. In that paper we describe the mathematical programming methods to design controlled rounding and suppression routines, the statistical techniques for data perturbation, and the more probabilistic analysis to attempt to evaluate risk. That paper contains an extensive bibliography and should be regarded as a companion to this one for the understanding of the underlying mathematics and methods. Although there will be some inevitable overlap between this report and the Annual Research conference paper the focus here will be on practical considerations in the design of a product for data release and a

description of current programs, planned products, and options which are available.

The overall theme of this Seminar is Quality of Federal Statistics. In addition to the notion of accuracy, other aspects

*of quality are timeliness and completeness. From the perspective of disclosure avoidance activities, we address the issues accuracy and completeness. We cannot release full and accurate detail on a public use file because that would exceed any reasonable level of disclosure risk. By taking measures to have acceptably low levels of risk we compromise completeness and/or accuracy. In designing a data release strategy we must evaluate the trade-off between completeness and accuracy and between completeness for one data attribute at the expense of completeness for another. To reduce levels of disclosure risk, one either suppresses information and collapses categories or introduces noise. Both these actions can be thought of as data masking. Under the first option we reduce

completeness while under the second we reduce accuracy.

Earlier, we introduced the idea of "amount of information" versus "level of risk" and indicated the need to Optimize amount of information while maintaining an acceptably low level of risk. We can think in terms of accuracy and completeness as components of level of information and evaluate the trade-off with acceptable levels of risk (which is much harder to characterize). This theme will run through-out the paper, and made explicit or not, this theme pervades the design of any data release strategy.

In Section II we discuss tabular data, including tables of amounts which are bated on our economic surveys and censuses and tables of frequency counts which appear in the Summary Tape Files (STF's) from the Decennial Censuses. In Section III, we discuss public use microdata. Public use microdata files are released as standard products from virtually all demographic programs and they are extensively used by researchers in many areas. In fact, the public use microdata files for the Survey of Income and Program Participation form the major data product from that survey.

Section IV consists of a brief summary.

II. Tabular Data

A. Frequency Counts of Demographic Characteristics

Cross-classified tables of frequency counts of demographic and housing characteristics constitute one of the major formats for release of data from the Decennial Censuses. For example, one such cross-classification can look like Table 1 below.

Table 1. Block Group 1 - Age by Sex

The major disclosure risk in the release of such tables occurs when a small value appears in a marginal position. If an investigator examines a table and knows the identify of the person or persons having marginal characteristics as indicated, the investigator could infer other characteristics of the respondent through the cross-classification. In so doing, the investigator would learn of information provided to the Census Bureau in confidence. The way to reduce this disclosure risk is to suppress cells with low marginal values or introduce uncertainty into cell counts.

Suppression was used for frequency counts from the 1980 Census

and for earlier Censuses. If a marginal value was below a specified cut-off, all cells summing to that marginal were suppressed. That is, if the cut-off were 10, Table 2 would have become Table 3.

In order to prevent deriving the third row in Table 3 by subtracting the non-suppressed rows from the totals row, at least

one more row must be suppressed. Suppressed values in Row 3 are

primary suppressions. Row 1 was chosen for the complementary suppressions, denoted by "C", to protect the primary as shown in Table 4. There were two problems with this method for disclosure avoidance. On one hand, due to the need for complementary suppressions, there were sometimes large values suppressed as complementary cells to protect small primary cells. This was considered a major draw-back for data users. The other problem with this procedure was that it was often difficult to guarantee geographic complementary suppressions. For example, if one or more data cells is suppressed for exactly one county in a state, then the suppressed value can be derived exactly by subtracting the value of all other counties from the state total. To avoid this from occurring, geographic complementary suppressions are required., It was clear that procedures to ensure complete complementary geographic suppressions would also take their toll in suppression

of even more information. This realization led to a recognized need to develop disclosure avoidance procedures for tables of cross-classified frequency counts along the lines of data distortion in order to introduce uncertainty into the data.

The method to be used as a disclosure avoidance measure on 1990 Census frequency count tables introduces uncertainty into the tables by changing some values. The basic idea is as follows and I quote, virtually verbatim, from (Greenberg 1990). For a subset of records, field values on a record will be replaced with field values on a different record having the same control characteristics so that the newly created records will be different on potentially all characteristics except the controls. This method has been called the Confidentiality Edit because of the use of a hot-deck similar in spirit to the hot-deck used in edit and imputation procedures. Given a target record on which some changes are to be made, based on specified control characteristics the system matches the target record to another record and "hot-decks" the remaining non-control variables. To be a little more specific, I paraphrase from (Griffin, Navarro and Flores-Baez 1989). The Confidentiality Edit selects a small sample of census household

records from the internal census data files and interchanges their data with other households which have identical characteristics on. a set of selected key variables but are in different geographic locations. The matching and interchanging operations are controlled on the key variables of number of persons in household; population characteristics of race, Hispanic origin and age; and on housing characteristics of units in building, rent/value and tenure.

Because of the controls described above, census counts for total persons, and totals by race, Hispanic origin and age, 18 and above. These counts provide information required for voting rights as outlined in Public Law 94-171. In addition, housing counts by tenure will not be affected by the Confidentiality Edit. The interchange of information on records will be accomplished on the detail file of records. The revised records will be used to

generate all tables so that there will be no inconsistencies

between tables and the revised records will also be used to produce other Census products. Three advantages of Confidentiality Edit include: (1) this procedure needs to be implemented only once on internal files to obtain protection for all Summary Tape File data products, (2) all data cells can be shown on Summary Tape Files so there is no interference with data aggregation by users, and (3) more data values will be available than in 1980. These procedures have been evaluated for their impact on data products and details of the analysis are contained in (Griffin and Thompson 1987) and (Navarro, Flores-Baez, and Thompson 1988).

For tables of frequency counts for the 1980 Census and earlier, there was a reduction in completeness through the use of suppressions to achieve an acceptable level of disclosure risk, Due to the need for complementary suppressions, the overall effect of a suppression pattern caused more loss in completeness than desirable. For the 1990 Census, by interchanging values on the detail record file, there will be a loss of accuracy through the interchange of information between records. The papers, cited above contain studies to show that loss of data utility due to this

reductions of accuracy is not significant.

B. Aggregate Economic Data

The primary method for releasing data from Census Bureau establishment surveys or censuses is in the form of cross-classified tables of amounts. For example, in a given state the total value of shipments may be cross-classified by SIC and by county. A cell is regarded as sensitive (i.e., having an unacceptable high disclosure risk) if the (N,K) -rule is violated, that is, if N or fewer respondents account for at least K% of the total cell value. Such cells are regarded as primary suppressions and they are not released. If only primary cells are suppressed, their values often can be derived exactly, or closely estimated through linear analysis using marginal totals. To prevent this, complementary suppressions are introduced, and one seeks a set of complementary suppressions which protects the sensitive cells yet suppresses as little additional information as possible.

We illustrate these ideas with a few (artificially) simple examples. Consider Table 5 in which cell (2,2) is considered sensitive because it failed the (N,K)-rule. We place a "P" in position (2,2) to indicate a primary suppression, and introduce a set of complementary suppressions, for example, as in Table 6.

Given a suppression pattern in a table, the values of all suppressed cells (primary or complementary) can be estimated. To indicate how this is done, we return to Table 6 which we rewrite as Table 6'.

From Table 6', we have the system of equations:

Note that Table 8 and Table 9 both display patterns of complementary suppressions. In Table 8 three complementary suppressions were introduced while in Table 9 four complementary suppressions have been introduced. The sum, of complementary suppressed values in Table 8 is 295 and the sum of complementary suppressed values in Table 9 is 135. For the 1982 and 1977 Economic Censuses, the criterion for selecting a set of complementary suppressions was to suppress as few complementary

cells as possible to protect the primary suppressions. For the 1987 Economic Censuses, we have implemented the criterion of suppressing the least total value. Thus, given Table 8 and Table 9 above, the preferred complementary suppression pattern under our current criterion will be as in Table 9 since less total value would be suppressed. In 1982 and before, the preferred pattern would have been as in Table 8 since fewer cells are suppressed.

The basic disclosure avoidance method for the release of cross-classified aggregate economic data at the Census Bureau is cell suppression. That is, reduction in completeness. This method seems to work well, especially as users can estimate the value of suppressed cells within acceptable limits. Whether we employ the criterion of minimizing the number of complementary suppressions or the total value that was suppressed constitutes a selection of methods within an overall strategy. We are currently investigating how procedures for finding complementary suppressions can be

improved for the 1992 Economic Censuses.

III. Microdata

Microdata records ate data records at the respondent level and the risk in the release of a microdata file is that someone may be able to discover the identity of a respondent. The risk can arise from the presence of highly visible and unique characteristics, or it may stem from the threat of matching public use microdata files to other files either privately or publicly held. For the latter threat of linking two files, some of the issues are: what data are available on both files, how comparably reported are the data, how up-to-date are they, and how easily accessed are the records? In particular, one must ask the cost to an investigator to carry out such a project. All these factors contribute to a picture of overall risk.

The basic strategy for the release of general purpose public use microdata files at the Census Bureau is to reduce completeness by restricting the level of detail on the file. Instead of

releasing exact date of birth, we can release month, or quarter, or year. Percentages can be grouped into deciles or quantiles. Income can be recoded into intervals of size, for example, $4,000 for income up to $100,000 and all income in excess of $100,000 can be topcoded to read as "$100,000 or more". Virtually all quantitative variables on public use microdata files are topcoded to obscure high visibility respondents and to reduce the likelihood of successful computer matching by removing outliers. In considering reduction in completeness or reduction in accuracy as disclosure avoidance practices, the Census Bureau tends to strongly favor reduction of completeness for the release of microdata files. By so doing, we are better able to maintain a broad range of utility for the files.

For some special purpose microdata files noise has been added to variables (in addition to topcoding and using categories) in order to further frustrate the ability for successful computer

matching, see (Greenberg 1990) for a further discussion. Such files can be created when we know in advance intended uses so we can design a noise introduction strategy to suit specified needs.

One of the most important fields oh a microdata record is the geographic identifier. Geography is the single identifier which cuts across all public use microdata files and is a field in which there is little error. Under current Census Bureau procedures, no area having fewer than 100,000 persons in the sample frame can be identified on a microdata record. This minimum can be raised for surveys which have a presumed greater disclosure risk. This was the case for the Survey of Income and Program Participation (SIPP) whose geographic cut-off was set to 250,000 by the Microdata Review Panel because of the fine level of detail on SIPP and the longitudinal nature of the survey.

Prior to 1981, each operating division had responsibility for the confidentiality of any public use microdata sample released by the division. At that time, no geographic area could be shown having fewer than 250,000 residents in the sampling frame. The

Microdata Review Panel was established in 1981 to review all proposed new microdata files for release. No new microdata file can be released by the Census Bureau without Panel approval. At that time, the geographic minimum was reduced to 100,000. The Panel is composed of representatives from Data Users Services Division, Program and Policy Development Office, Demographic Surveys Division, and representatives from the Associate Directors for Economic Fields, Demographic Fields, and Statistical Standards and Methodology. This Panel make-up reflects broad Census Bureau concern.

As part of the review process, survey staff seeking release approval must fill out a disclosure checklist which asks about identifiable geography, matching potential, topcodes, etc. The Panel typically meets with survey staff to discuss problems to seek a resolution. The Panel may request additional topcodes, deletion or recoding of some variables, and other actions to reduce disclosure risk. At times the Panel will request cross-tabulation frequency counts to observe if there are outlying combinations of values. The Panel may recommend changes; however, it is more

typical for the Panel to point out problems and leave it to the survey staff to find solutions based on their understanding of intended uses of the file. Survey sponsors attend Panel meetings to discuss options and assist in the determination of risk and resolution options.

There are often a number of options available to reduce risk on a file. For example, typically one of several variables can be recoded to reduce the possibility of matching to external files.

At times, and depending on perceived user needs, geographic specificity can be reduced. That is, one can provide more potentially identifying demographic characteristics on a national file (i.e., no subnational geography) than on a file that identifies a relatively small geographic locale. By and large, one must think in terms of trade-offs between the various data items and their relative completeness. In a public use microdata file, it is not possible to provide a very complete and accurate file due to an unacceptably high level of risk. Survey sponsors and data

users must contribute to the decision making process in identifying areas in which some completeness and/or accuracy can be sacrificed while attempting to maintain as much data quality as possible.

Below we list some options which are currently available to- enhance data utility with no increase of risk. If the topcode on some item, say income, is $100, 000, replace all values- over the topcode by the mean (or median, etc.) of the topcoded values. Thus, if the mean of the topcoded values were $130,000, replace any value in excess of $100,000 by $130,00O. This is in contrast to the current practice of replacing topcoded values by the cut-off (in this case $100,000). In fact, one can actually provide the exact distribution of all topcoded values.

Another option we have is for local topcodes. For the Metropolitan Sample of major cities from the American Housing Survey, each city has a different topcode for "home value" based on (roughly) a three percent upper tail cut-off for that city. Would state-level topcodes for such items as income, housing costs, etc. be desirable for other files? Would such a strategy provide more

useful data?

The Census Bureau is currently planning for the Public Use Microdatal Samples (PUMS) for the 1990 Decennial Census. Current plans call for a "standard" 1% file and 5% file as were produced for the 1980 PUMS. In addition, we are considering another file having only national geography but containing far more detail than the other files. For example, we are considering adding tract characteristics to each record. That is, we append to each record information about the tract of residence; information such as unemployment rate, percentage of minorities, median home value, etc. Such local detail would not be acceptable on a file with more specific geography, for fear one may be able to identify tract of residence based on tract characteristics. In addition, there is no-reason, a priori, that income topcodes, and other topcodes as well, cannot be raised to allow more detail for the respective variables on a national file. This also represents a trade-off between various kinds of reduction of completeness -- geographic detail verses demographic detail -- in which we provide less of the former to obtain more of the latter.

It is important that users of public release microdata files contribute to the discussion of methods for the design of such

files often options and choices are available, and to the extent that user priorities are known efforts can be made to accommodate them.

IV. Summary

In this report, we describe methods used by the Census Bureau to reduce disclosure risk in the release of data products. We discuss tabular data and microdata for which the issues are somewhat different. In a related paper (Greenberg 1990) we provide a detail discussion of Census Bureau research efforts in the area of disclosure avoidance.

In the design of a data release strategy many options are

typically available. The trade-off between loss of completeness and loss of accuracy is a theme that runs through much of the discussion. Plans are being made for the Public Use Microdata Samples from the 1990 Census. It is important that data users contribute to the planning process by contributing to the discussion of options and choices by indicating both needs and preferences.

References

Greenberg, B. (1990), "Disclosure Avoidance Research at the Census Bureau," Proceedings of the 1990 Annual Research Conference, Bureau of the Census, Washington, D.C. (to appear).

Griffin, R.A., Flores-Baez, L. and Navarro, A. (1989), "Disclosure Avoidance for the 1990 Census," Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, D.C., (to appear).

Griffin, R.A. and Thompson, J. (1087), "Confidentiality Techniques for the 1990 Census," presented at the Fall meeting of the American

Statistical Association and Population Statistics Census Advisory Committees.

Navarro, A., Flores-Baez, and Thompsono J. (1988), "Results of Data Switching Simulation," presented at the Spring meeting of the American Statistical Association and Population Statistics Census Advisory Committees.

THE MICRODATA RELEASE PROGRAM OF THE NATIONAL CENTER FOR HEALTH STATISTICS

Robert H Mugge, PhD

National Center for Health Statistics (retired)

My presentation will be in three parts: First I shall

describe the microdata release program of the National Center for Health Statistics (NCHS, or "the Center"); secondly I'll explain the rules and procedures followed by NCHS in attempting to insure the confidentiality of the subjects of our data; finally, I shall discuss some concerns I have for confidentiality protection for these NCHS data and some suggestions for meeting the problems that I see.

Let me make clear, that I am not speaking as a staff member of NCHS, but rather as one who retired from that staff nearly eight months ago after working in the confidentiality program of NCHS for quite a few years. So I am now speaking only for myself and not on behalf of the Center. I am told that there have been no important changes in the Center's data security program since I left, and Mr. Israel, Deputy Director of the Center, has kindly reviewed this paper for current accuracy. But all opinions and commentary are strictly my own and not necessarily those of the Center.

The NCHS Microdata Release Program.

The primary function of the National Center for Health Statistics is to develop and make available statistical information on the health of the U.S. population, on the vital statistics of the U.S., and related matters. This is clearly stated in the law authorizing the work of the Center (3). The Director of the Center decided many years ago that, carrying this mandate to its proper conclusion, the Center would make available its statistics in as full detail as possible for the use of scholars who wish to analyze these data. The covering policy statement is this: "Within prevailing ethical, legal, technical, technological, and economic restrictions, it is the policy of the National Center for Health Statistics to augment its programs of collection, analysis, and publication of statistical information with procedures for making available, at cost, transcripts of data for individual elementary units -- persons or establishments -- in a form that will not in any way compromise the confidentiality guaranteed the respondent (6)."

Implementing this policy, NCHS has now for a long time, and with only rare exceptions, made available quite detailed data sets, known as Public Use Data Tapes, on all of its finished surveys and data reporting programs, together with full printed documentation

(2). These systems include the National Health Interview Survey; the National Health and Nutrition Examination Survey; the National Hospital Discharge Survey; the National Ambulatory Medical Care Survey; The National Nursing Home Surveys; the National Survey of Family Growth; several follow-up surveys; annual vital statistics on births, deaths, fetal deaths, marriages, and divorces; and various others (4). Many years ago these files took the form of boxes of punched cards; now for a long time they have been on one or more reels of magnetic tape; recently they have been made available on tape cassettes; and now the Center is moving into a program of producing the files on CD-Roms. However, the material form of the data file is not relevant to the principles involved in

data release.

Confidentiality Protection.

As noted in the policy statement, the Center is very concerned that the confidentiality of data subjects in its surveys and reports be maintained. From its study of the problem NCHS has devised a set of rules for protecting the confidentiality of subjects -- persons and establishments -- whose information is included; these rules are stated in the NCHS Staff Manual on Confidentiality (5). The Center has a Confidentiality Committee, made up of high level staff, which reviews needs for policy changes and makes recommendations regarding them to the Director.

The rules followed in the Center for protecting confidentiality are of two kinds. One set of rules relates to data published in tabular form and provides limitations on the contents of published statistical tables; the other set of rules covers what may be included in the public use microdata files (5). But tables

published by the Center are generally limited to what may be tabulated from the public use files, and, when this is the case, the former set of rules may be ignored if the rules on the microdata files are first met.

The rules regarding tabular production of data are designed to insure that no single cross-tabulation, or combination of cross-tabulations, may permit disclosure of a confidential characteristic of any identifiable individual or establishment. This possibility may be substantially dismissed in the cases of the Center's large scale surveys of persons, involving samples representing usually far less than one one-thousandth of the relevant population, provided that data are not presented separately for any small areas, in which a unique individual might stand out. But special care must be taken in the reporting of establishments, since these often involve large proportionate samples, and reporting data for larger areas -- perhaps even census regions -- may serve to disclose data on particular institutions. But in any event, if the necessary care has been taken in restricting the contents of the microdata set, and only this microdata set is used in building

tables, it fo1lows that there should be no disclosures resulting from the publication of these tables.

The rules for protecting confidentiality in public use microdata sets, as set forth in the Manual, are as follows (5, P. 19):

1) All direct personal or establishment identifiers, such as name, social security number, or address are purged from the file.

2) The file must not contain any other detailed information about the subject that could facilitate identification and that is not essential for research purposes (such as the exact date of the person's birth). It is often found necessary to give certain numerical information-such as income, nursing home size, or costs and charges of

institutions -- only in broad class intervals in order to avoid disclosure.

3) Geographic places that have fewer than 100,000 people are not to be identified in the file. (In practice much larger places often cannot be identified, such as when a State is known to have primary sampling units totalling less than 100,000.)

4) Characteristics of an area are not to appear in the file if they would identify an area of less than 100,000 people.

5) Information on the drawing of the sample which might assist in identifying a data subject must not be released outside the Center. Thus the identities of primary sampling units are not to be made available outside the Center. (I must say in all candor that NCHS seems to have lost control of that one, as it turned out that for several reasons the PSU identifications have unavoidably been made public.)

6) Before any new or revised microdata files are published, they, together with their full documentation, must be approved by the Director or Deputy Director. (When I was there this responsibility was delegated to me, and I reviewed all plans for public use microdata files, to make sure they complied with the governing rules and that no other data had crept in which might compromise confidentiality. In the Census Bureau there is a high-level Microdata Review Panel that reviews plans for each public use microdata set release [1]; NCHS did not feel that such an expenditure of staff time was necessary in its particular situation.)

7) Finally, NCHS required that before anyone outside the Center may be provided a public use microdata file, that person would be required to sign a statement called a "Data Use Agreement." This statement points out the

legal requirement that no data obtained by NCHS under its mandate may be used for any purpose other than the purpose for which it was obtained, i.e., for statistical purposes (3, Section 308[d]), it notes that all appropriate precautions have been taken to keep the data safe from,disclosure, but that there may still be a way that the data could inadvertently be used for identification. In signing the agreement, the user states that he/she understands this and gives assurance that the data set he/she receives will not be misused, the security of the data will be protected, and no attempt will be made to reveal identities of data subjects, and, further, that if any subject is accidentally identified the user will work with the Center to make sure that this identification is not used, and procedures will be taken to assure that the identification cannot be repeated (2).

NCHS requires signing of the Data Use Agreement, even though it may have no force in law, because of its information value and its assumed effect in raising the

sensitivity of data users on the importance of protecting the files.

NCHS does not doctor the data in other ways to avoid disclosure. It does not substitute any false data, nor does it do anything like data swapping. It has been determined that any such procedure would lessen the value of the files for purposes of research, and it is felt that this is most undesirable to do, as the nation has an important stake in getting the best research possible using these health-related data. But primarily it has been considered unnecessary to doctor the data. The files contain many errors already -- normal errors in data collection and processing -- although the Center tries hard to minimize them. So the user cannot have absolute confidence in the information, especially as it relates to individual subjects. The Center is reluctant to add additional errors.

Is the system working? It seems to be. In all the years it's been operating the Center has never heard of a case in which a disclosure has been made through one of the survey files. That it

comforting, but the comfort is mitigated by the knowledge that the Center wouldn't necessarily ever hear about it if a survey file were compromised.

There is also another piece of evidence as to the effect of the confidentiality program. I don't think the Center has ever received a complaint from the public that it isn't protecting

confidentiality adequately in the data files program. But there have been many complaints from researchers that the Center isn't releasing enough information to them. Since we felt that, if anything, the Center erred in being too liberal in releasing data, the researchers' complaints encouraged me to feel that the Center is balance may have been about right.

Commentary

Is this confidentiality-protection program, then, good enough? There cannot be perfect confidentiality protection if any midrodata, are to be released. So each organization must seek to find the proper balance between the public's needs for data and the appropriate measures for protecting confidentiality (7, pp. 1-2). I have described the compromise position reached by NCHS.

In the large surveys it conducts NCHS obtains a great deal of information, which may be considered independent or dependent variables for data analysis, on each individual subject (4). This mass of data may constitute a fingerprint about an individual; there may be no one else with this particular set of characteristics. So if one had another source of such information about individuals (or establishments) then it would be easy to match them up and disclose all the new information in the survey on the identified individual. Fortunately, no such files exist about individuals on any mass basis. There may be similar files in other sample surveys, but the chance of overlap is so small that the likelihood can be dismissed. This is not true about establishments; there are lists of nursing homes, hospitals, and

clinics that could be used to identify them if the file contains the right kinds of characteristics. So great care must be taken in determining what can be published about characteristics of establishments included in data files; we've made some last minute discoveries on certain file-release plans which we hope enabled us to make these files disclosure-proof.

For those concerned about inadvertent disclosures there is this consolation: the vast majority of information in our data files is quite innocuous. Much of it is obvious, at least within certain limits, or already well known by individuals' associates, such as the person's sex, age, and weight, and various obvious health conditions. If you found a friend's file in one of the surveys you are not very likely to learn through it something you didn't already know, or, if you did learn something new it would probably be quite harmless. So chances are that data subjects are not likely to sustain any harm or embarrassment from having their data disclosed. This, however, would certainly not excuse an agency from not doing its utmost to keep its promise of confiden-tiality to the data subject.

Historically, however, there have always been some sensitive items in the data files -- information that could cause harm and embarrassment. There have been the early and late effects of venereal disease; there have been some diseases which at times have carried stigmas, such as leprosy and cancer; and there have been surveys on social behavior, such as sexual practices, which if the information got out could cause considerable harm. And now it seems that society needs to obtain new information which may carry threats of individual harm beyond that brought by any data in the past. I have two particular examples: one is AIDS or its precursor, the HIV virus; the other is information on the sexual practices of unmarried teen-age girls, as obtained in the latest cycle of the Family Growth Survey. (The latter information is obtained with the approval of parents, but the parents are not given the information obtained from the girls.)

There is so much concern about protection of the AIDS or HIV information that, so far, surveys to obtain it are only being done using procedures which guarantee the anonymity of subjects, even from the data collecting agency.

Now, with all of NCHS's efforts to protect the confidentiality of the data, there is one scenario that haunts me. That is this: If someone knows well a survey subject and knows the person was in a particular survey and at some time has access to that survey file, then he/she could easily use his/her knowledge of the person to locate the person's file in the data set. Then all the information about that person obtained in the survey will be laid out before them.

I don't know of any reasonable way to avoid this possibility, beyond what the Center is already doing, especially through the Data Use Agreement. I wouldn't like the idea of warning data subjects not to tell their friends and relatives of their being in the survey sample, when the Center is trying to get good publicity for the survey and urging people to cooperate. Of course, though,

if it is found that a data subject is advertising being in the survey, that person is out. (Last year a college professor wrote in to tell the Center staff how interested she was in having been included in the Family Growth Survey. She had told her class all about it. The Center wrote back to say that her interest was appreciated, but her record was being removed from the survey sample.)

There is, however, one policy response I think the Center should make to the scary scenario I alluded to. I think the Center should lean over backwards to assure that there is as little sensitive information as possible in the public use data files. I think that NCHS should, for example, make sure that there is no AIDS or HIV-positive information in any survey files published where there is any possibility of individual disclosures. By the same token, the Center should not release in microdata files the

Family Growth Survey data on sexual practices of unmarried women;

I personally feel that that entire survey is too sensitive to justify any microdata release from it. That will make the researchers angry with the Center, and the Center should arrange to do the special tabulations and analyses that the outside researchers want, within limits of practicality.

Eternal vigilance is the price of good data confidentiality protection. But a prime needed ingredient in a successful confidentiality protection program is clout! If the head of any agency producing statistical files is not keeping a close eye on protective procedures and lending her/his authority to maintaining a strict program of protection, that statistical program -- along with all the people depending on it -- could be in big trouble. I think the protection program in NCHS has been successful, and that success owes much to the continuing concern and support it has received from the Center's Directors.

References

(1) Gates, Gerald W., "Census Bureau Microdata: Providing Useful Research Data while Protecting the Anonymity of Respondents." U.S. Bureau of the Census, Program and Policy Development Office. Paper prepared for presentation at the annual meetings of the American Statistical Association in New Orleans, August 1988.

(2) National Center for Health Statistics, Cataloque of NCHS Public Use Data Tapes. DHHS Publication No. (PHS) 88-1213. U.S. Department of Health and Human Services, Public Health Service, Centers for Disease Control. Hyattsville, MD, July 1988.

(3)____, Current Legislative Authorities Enacted as of December 1989: Sections 304, 306, 307, and 308 of the Public Health Service Act. U.S. Department of Health and Human Services, Public Health Service, Center for Disease Control. Hyattsville, MD, April 1990.

(4)____, Data Systems of the National Center for Health Statistics: Programs and Collection Procedures, Series 1. No. 16. U.S. Department of Health and Human Services, Public Health Service, Office of Health Research, statistics, and Technology.

Hyattsville, MD, December 198l. More recent versions may be available in unpublished form from NCHS.

(5)____, NCHS Staff Manual on Confidentiality. DHHS Publication No. (PHS) 84-1244. U.S. Department of Health.and Human Services, Public Health Service. Hyattsville, MD, September 1984.

(6)____, Policy Statement on Release of Data for Individual Elementary Units and Special Tabulations. DHEW Publication No. (PHS) 78-1212. U.S. Department of Health, Education, and Welfare, Public Health Service. Hyattsville, MD, May 1978.

(7) Subcommittee on Disclosure-Avoidance Techniques, Federal Committee on Statistical Methodology, Statistical Policy Working Paper 2: Report on Statistical Disclosure and Disclosure-Avoidance Techniques. U.S. Department of Commerce, Office of Federal Statistical Policy and Standards. Washington, D.C., May 1978.

DISCUSSION

George T. Duncan

Carnegie Mellon University

Serving the public, federal statistical agencies must balance the respondent's need for privacy and the researcher's need for information. Ultimately, how the balance is struck should be the result of the political process, which in the U. S. is complex, indeed. What the agencies can contribute to this is the development of procedures, both administrative and statistical, that for a

given level of privacy protection maximize access to data and that for a given level of access maximize privacy protection.

Brian Greenberg, of the U.S. Bureau of the Census, and Robert Mugge, retired from the National Center for Health Statistics, have given us a clear perspective on the data dissemination policies and practices of two major federal statistical agencies. As chair of the Panel on Confidentiality and Data Access that is co-sponsored by the Committee on National Statistics and the Social Science Research Council, I find their work of special importance -- both to the panel and to all federal data users and providers.

Brian Greenberg, both here and in a recent paper (Greenberg, 1990) describes some disclosure limitation practices at the Census Bureau. He properly emphasizes the need in data dissemination for a tradeoff between "amount of information" and "level of disclosure risk". He identifies the open research problem of quantifying each to be meaningful for disclosure-limited data dissemination. While mathematical arguments are useful in this quantification, essentially the task is a decision-theoretic one that incorporates

the motivations of the stakeholders. Since these stakeholders are various -- including individual respondents, other government agencies, academic researchers, market researchers, commercial planners, the media, and lobbying groups -- the measures developed should be multivariate in nature.

He notes in Greenberg (1990) that

> The first general purpose public use microdata file released by the Census Bureau was the 1 in 1,000 sample from the 1960 Census of Population and Housing. This file was released in 1963. A few years later a public use microdata file from the Current Population. Survey was released. At present, public use microdata files are released as standard products from virtually all demographic surveys, and they are extensively used by researchers in many areas. In fact, the public use microdata files for the Survey of Income and Program Participation form the major data product from the survey. The Public Use Microdata Sample from the

Decennial Censuses are becoming increasingly important to users, especially researchers in the social sciences, and these files are gradually replacing the Summary Tape Files for many research applications.

Against this history, his paper focuses on statistical procedures that the Census Bureau uses to limit disclosure risk, rather than procedures -- whether statistical or administrative -- for expanding research access. While the disclosure limitation aspect is important, I would have liked to have seen more attention paid to the ways the Census Bureau has actively tried to make information available.

Much mote than Robert Mugge, Brian Greenberg emphasizes masking procedures, both for disclosure limitation in tabular data and for microdata. He draws a nice conceptual distinction between the effect of disclosure limitation on data utility through loss of accuracy and loss of completeness.

For tabular data disclosure limitation, he shows how cell suppression and the technique of a Confidentiality Edit can be employed. For microdata disclosure limitation, he stresses grouping of quantitative variables, including topcoding. In masking some special purpose files, noise is also added to lower the likelihood of a successful computer match with publicly available files having identifiers.

Greenberg describes the work of the Microdata Review Panel, which is broadly representative of key components of the Census Bureau. Commendably, the Census Bureau, through work of Gerald Gates and the Microdata Review Panel, has been open to suggestions of ways of expanding researcher access to data.

In thinking about Greenberg's paper, certain questions nagged me.

How useful will researchers find the data after it has been massaged by the various disclosure limitation procedures? A

research effort is needed on appropriate ways of analyzing masked data.

What do researchers need to be concerned about in analyzing data that has been "Confidentiality Edited"? Some researchers will ignore the fact that the data has been altered, and hence produce misleading conclusions from their analysis. Researchers need to be carefully informed about the limitations of standard analyses of the edited data. This requires a study of how researchers in practice respond to various caveats attached to the released data.

Do the thresholds on geography (100,000 persons or 250,000 for SIPP) have any basis in theory or do they just "feel right" to the Microdata Review Panel? In another paper to be presented at the

American Statistical Association meetings in August, Brian Greenberg and Laura Voshell relate the size of geographical units to the percentage of unit records. This is a good start but the

direct tie to disclosure risk is not yet made.

Are any special disclosure-limiting procedures used for longitudinal data?

How can data users best be brought into the decision making process? Specifically, how can agencies insure that data users help identify what in data accuracy can best be spent to buy disclosure limitation?

How do respondents view the level of disclosure limitation provided by these procedures?

Robert Mugge has described what is by all accounts a successful program in microdata release in an important area for public policy. As with Brian Greenberg's paper, I would like to highlight certain aspects of why I think the program is successful and then focus on some specific concerns that I have for the future.

I believe the program's success stems from the basic policy statement of the National Center for Health Statistics:

> Within prevailing ethical, legal, technical, technological, and economic restrictions, it is the policy of the National Center for Health Statistics to augment its programs of collection, analysis, and publication of statistical information with procedures for making available, at cost, transcripts of data for individual elementary units -- persons or establishments -- in a form that will not in any way compromise the confidentiality guaranteed the respondents

The three reasons I see are these:

First, NCHS has taken seriously its mandate to make microdata available to researchers. Its focus is not predominately on data collection but equitably on data dissemination. This balance ensures good stewardship, not the hoarding of quality data but rather its investment in the work of researchers who can advance the public good.

In thinking about how disclosure limitation related to researchers, I should point out that researchers come no just from academia, as faculty members at Carnegie Mellon, but also from the media, as reporters for the New York Times, and lobbying groups, as analysts for the American Association of Retired Persons, say.

Second, NCHS explicitly recognizes that there are constraints on microdata dissemination that have ethical, legal, technical, and economic dimensions. This cues where to look for potential problems.

Third, NCHS does not guarantee that identifiability is impossible but instead links to the implicit contract that it has established with the respondent. This is both realistic and responsible.

In implementing this policy, NCHS has made Public Use Data Tapes of key surveys available -- in accord with the directions of a Confidentiality Committee and under rules well stated in the NCHS Staff Manual on Confidentiality. So NCHS does in fact deliver the data, but makes sure that there-is well-identified administrative oversight and that the policies exist in written form for reference by agency staff, researchers, and the interested public.

Further, through the "Data Use Agreement" NCHS encourages the receiver of the data to assume some responsibility for proper use of the data. This agreement, while not legally binding, educates the researcher on the restriction to statistical use of the data, provides that no attempt will be made to identify data subjects, and provides for appropriate action in the case of accidental identification.

In pointing to the future, I would like to fix on a few concerns:

Why should NCHS not emulate the Census Microdata Review Board? And indeed go further by including representatives of both the respondent and the user communities? Through internal representation from various areas within NCHS it might further consistency of application of the policies. Through external representation it might foster responsible interaction with both respondents and researchers. It might help ensure that respondents got meaningful information about agency practices and intentions so that in authorizing how their responses are to be used, they would be properly informed. It might help ensure that researchers needs were addressed and that data quality was not unduly sacrificed in the name of confidentiality protection.

Why not have the Data Use Agreement be more binding on the researcher? Possible mechanisms for this might be legal requirements (such as legal sanctions or use of binding contracts) economic incentives (such as use of bonds or returnable license fees), or administrative practices (such as restrictions on further access)?

Why not use data masking in certain cases where the data might not be releasable otherwise? This requires that enough information be provided to the researcher that an appropriate analysis can be

carried out. It also requires that suitable techniques be developed for the analysis of masked data.

Might not advances in computer technology increase the prospect of linkage of a record with an identifier with a released record, even when only a sample is released? After all, in a recent JASA article, Bethlehem, Keller, and Pannekoek note that in a certain region of the Netherlands having 23,485 households composed of a father, mother, and two children with just a six-item key of ages (in years) and gender, 16,008 of the households were unique. Presumably, a plausible model could be constructed in which -- with a bit more detail -- a data intruder who matched a record in the sample uniquely would also place high probability that it is a unique match in the population.

Should not special administrative procedures be developed for establishment -- like nursing homes and hospitals -- that cannot be reasonably assured that they would not be identifiable? For example, large hospitals might be asked for authorization to include their data in a public use file.

Are many variables really innocuous? Marital status, age, and weight under certain circumstances are sensitive, let alone the details of sexual practices required in AIDS-related surveys.

Can an agency depend on naturally-occurring errors to provide confidentiality protection? Introducing noise with a distribution, known to the user may be effective. Research is ongoing in this area.

To sum up, I think with the attention of professionals such as Robert Mugge and Brian Greenberg that the growing tension between privacy concerns and demand for data access can better be mediated. To take a clue from Fritz Scheuren, we need DANTOTSU, Japanese for "choosing the best of the best". The federal statistical agencies

can then be better stewards of the data our citizens provide to further the public interest.

References

Bethlehem, Jelke G., Keller, Wouter, J., and Pannekoe Jeroen (1990) Disclosure Control of Microdata. Journal of the American Statistical Association 85, 38-45.

Greenberg, Brian (1990) Disclosure Avoidance Research at the Census Bureau. 1990 Annual Research Conference, Bureau of the Census, Arlington, VA, March 18-21.

Session 12

FEDERAL LONGITUDINAL SURVEYS

FEDERAL LONGITUDINAL SURVEYS

Daniel Kasprzyk

U. S. Bureau of the Census

Curtis Jacobs

U. S. Bureau of Labor Statistics

I. Introduction

During the 1960's and 1970's, panel surveys surveys in which similar measurements are made on the same sample at different points in time -- became a popular tool for social science and policy research. Boruch and Pearson (1985) indicate 64 national surveys of this kind were carried out during that period of time. The apparent popularity of such survey designs prompted the Office

of Management and Budget's Federal Committee on Statistical Methodology (FCSM) to form a subcommittee on "federal longitudinal surveys" during the Spring of 1983 under the chairmanship of Barbara Bailar and Daniel Kasprzyk. Maria Gonzalez, chair of the FCSM, provided organizational and staff support to the subcommittee. The subcommittee's goals were very general -- to identify the strengths and limitations of longitudinal surveys, and to propose some guidelines for using them more effectively.

The Subcommittee on Federal Longitudinal Surveys was composed of the following members: Barbara Bailar (co-chair, Bureau of the Census), Daniel Kasprzyk (co-chair, Bureau of the Census), Barry Bye (Social Security Administration), Dennis Carroll (National Center for Education Statistics), Robert Casady (Bureau of Labor Statistics), Steven B. Cohen (Agency for Health Care Policy and Research), Lawrence Ernst (Bureau of the Census), Maria Gonzalez (Office of Management and Budget), Catherine Hines (Bureau of the Census), Curtis Jacobs (Bureau of Labor Statistics), Inderjit Kundra (Energy Information Administration), and Bruce Taylor (Bureau of Justice Statistics).

This paper follows the general outline of the working paper developed by the OMB subcommittee. We discuss the advantages of longitudinal surveys, managing longitudinal surveys, some activities related to longitudinal survey operations, estimation, some persistent issues in longitudinal surveys, and data user issues.

II. Definitions

Terminology in this area of social science research has not been standardized. Kish (1987) describes longitudinal studies as a generic term referring to a wide variety of studies done over time. Duncan and Kalton (1987) prefer to use the word

"longitudinal" in the context of data; thus, permitting

longitudinal data to be collected in either a panel or cross-sectional (retrospective) survey.

The subcommittee chose to combine two components, design and data, into the definition of longitudinal survey adopted for the report. The distinguishing features of a longitudinal survey are: 1) repeated data collection for a sample of observational units over time; 2) the linkage of data records for different time periods to create a longitudinal record for each observational unit; and 3) the principal analysis was to be based on the data collected over time.

The subcommittee's definition is more restrictive than that adopted by Duncan and Kalton or Kish, since longitudinal surveys are those in which the sample unit is followed, microdata assembled, and longitudinal analysis included as part of the estimation plan.

III. Advantages of Longitudinal Surveys

A longitudinal survey is usually needed to measure and study micro-level dynamics -- changes in attitudes, changes in prices, changes in economic well-being, for example -- or to improve the measurement of certain important concepts (Pearson, 1989). Some advantages for obtaining repeated measurements on the same sample unit over time are: 1) multiple interviews of the same sample unit reduce sampling variability on estimates of changes; 2) a matched longitudinal data set provides a better measure of components of individual change; that is, measures of gross change for the unit at two, points in time; 3) a longitudinal survey is capable of obtaining a wider range of variables from each sampled element than is possible from a repeated survey of cross-sections; 4) longitudinal surveys with relatively short reference periods may reduce telescoping errors that occur when respondents misplace the timing of the occurrence of events; 5) longitudinal surveys with relatively short reference periods can be used to produce aggregated data for a longer time period -- a year, for example. While longitudinal surveys are advantageous, they do not solve all data collection problems. In fact, they create some additional problems which will be discussed later.

IV. Managing Longitudinal Surveys

Managing large, complex longitudinal surveys has much in common with managing large, complex cross-sectional surveys. Successful project management techniques and the issues surrounding the successful execution of a project should not be related to the design of the project. There are, however, nuances in the case of longitudinal surveys that are important to recognize. They are:

1) inordinately high expectations for the project; 2) budget planning; 3) content, procedural, and methodological innovations; and 4) changes in the data collection organization.

Expectations associated with longitudinal data collections typically run high. The set of analysts interested in the data set

as a vehicle for answering-their own research questions is often broad and diffuse. The sum of these expectations as well as the project staff's expectations, almost by definition must exceed what is achievable in the short run. Grasso and Kohen (1978) make this point concerning The National Longitudinal Surveys (NLS); similarly, Duncan and Morgan (1984) admit that judged by the expectations for the Panel Study of Income Dynamics (PSID) the investment in the study could not have been profitable.

Long range planning and budget planning play an important role in the development of a longitudinal survey. A long range planning document laying out the budget, analysis plan, instrument development plans, staffing plans, survey procedures, and anticipated products is one way to assist senior agency officials in understanding the need for the project; it also provides a baseline document for the survey.

Another aspect of the management of longitudinal survey operations is the persistent tension in maintaining the status quo versus making corrections and alterations to the instrument and

processing system. A serious analysis of the trade-offs from the cost as well as analytic point of view ought to be made before making a change.

Long-term longitudinal surveys, such as the NLS and the surveys sponsored by the National Center for Education Statistics, can be spread over a decade or more. These, surveys when contracted to private sector survey research organizations usually have periodic recompetition for the contract. However, a change in data collection organization can be very traumatic to the longitudinal survey project if not properly planned for. A very detailed level of documentation of methods is required to ease the transition, if it should be necessary.

V. Longitudinal Survey operations

The differences between field and processing operations in one time cross-sectional surveys and longitudinal surveys are created by the time dimension. For example, time enters in the selection

of new units into the sample, in identifying and matching the same sample unit from round to round of the survey, in following sample units from one interview to the next, and in the way longitudinal products are released. We discuss these below.

A. Maintaining the Composition of the Sample

The composition of the sample may be expected to change across waves for many reasons. Respondents may refuse to participate, they may not be home, they may die or may be institutionalized, or may go abroad. To reduce the effects of these problems, some continuing panel surveys routinely introduce new sample units at certain points in time within a panel. These designs are called rotating panel designs.

Other longitudinal surveys, such as the Panel Study of Income

Dynamics (Duncan, Juster, and Morgan, 1986), argue that the representativeness of the sample of the entire population of families and individuals can be maintained over time through rules that allow families and individuals to enter the sample with known selection probabilities.

B. Following Individuals Over Time

The issue of whom to follow in a longitudinal survey and the intensity at which one follows individuals over time is directly related to the analytic uses of the data, the amount of time between interview rounds, and the budget of the survey. Analytic uses should drive the operational decisions of whom and how far to follow an individual. If the basic sampling unit and unit of analysis is the individual, then the following rules consist of following all individuals originally selected into sample. These are generally called cohort studies.

Another design, labelled by Cox and Cohen (1985) as a longitudinal household design, consists of the individual as the

basic sampling unit. The dwelling unit is sampled and all individuals living in the dwelling unit are selected into sample at the first round of interviewing and are interviewed in subsequent rounds whether or not they reside at the original sample address. In order to develop household and family estimates for the dwelling units, data are obtained from all individuals living at the address of the person originally identified as a sample individual.

Tracing is directed toward obtaining the current address of the survey respondent. Some people move great distances and are difficult to trace; others may not want to be traced. The operations of the survey organization must establish a set of information sources that are capable of providing current address information for individuals who move. Several surveys obtain information from the respondent at the end of each interview on the name and phone number of a person who will always know the sample person's whereabouts.

Other information sources can be developed by the interviewer through the sample person's friends, relatives, and other contacts

established through the respondent, such as neighbors employers, and directories (Burgess, 1989). The mode of interview may effect the types of tracking techniques used. Personal visit surveys will use mail or telephone tracking as well as place heavy reliance on the interviewer for creative solutions to finding respondents. Telephone surveys are likely to rely on the telephone for tracking, but rarely send staff into the field (Cantor, 1989).

An operational concern in tracing respondents is the additional costs incurred by field staff. White and Huang (1982) have estimated that during a one year time period (4 interviews) of the Income Survey Development Program (ISDP) Panel, the number of interviewing hours increased by 7% and the number of miles charged by the interviewer increased by 22%, due to the cost of following movers and interviewing additional households. However, NCES found that per-unit tracing costs for NCES' High School and Beyond Survey were approximately 20% less than the cost of base year sampling,

indicating the potential economies of longitudinal surveys (Office of Management and Budget, 1986).

C. Linking Analysis Units Between Waves

What is the point in conducting a longitudinal survey if data from successive interview rounds can not be successfully brought together for analysis? Obviously linking or matching variables must be created to permit the merging of data over time. Complications arise when consideration must be given to multiple units of analysis.

Surveys which are intended to follow individuals, regardless of their association with the sampled household location or address simply assign an independent and unique person identification number to each individual.

The Survey of Income and Program Participation (SIPP) is illustrative of surveys requiring linkage variables to allow analysis at various levels -- household, family, person, and event.

The SIPP has a complicated variable as described by Jean and McArthur (1984, 1987) which ensures that the identification number remains constant regardless of changes in address and household composition.

The National Medical Care Expenditure Survey (NMCES) took an altogether different approach by using the identification variable for internally matching the rounds of interviews and providing the public with matched rounds of data. As a consequence, round-to-round matching was unnecessary for the public.

D. Operational Changes over Time

Changes in the administration and operation of longitudinal surveys seem to be inevitable; these changes, however, are likely to make comparisons difficult to assess and interpret. One needs to recognize that aspects of the survey design and data collection that change during the course of the survey may influence results. This is not to say that one should never make any changes; rather, one needs to be aware of the consequences of actions taken and attempt to measure the effects of such changes.

E. Operational variations in Longitudinal Data Products

As with any complex study, many variations are possible in the processing and development of longitudinal data products. Three illustrations give a sense of the options available. The main Panel Study of Income Dynamics (PSID) data files contain information gathered since the beginning of the study in 1968 and are updated on an annual basis. Thus, each wave of data is released together with the data previously made available.

Because the length of the survey is predetermined, the

National Medical Expenditure Survey (NMES), prefers to wait to the conclusion of the panel -to develop its longitudinal products. This survey program uses the multiple interviews as a vehicle to revise data or fill in data not completely reported in earlier interviews.

The SIPP, on the other hand, first releases individual wave public data files to allow researchers the opportunity to analyze each wave as a separate cross-sectional data set and to provide the ability to develop their own multiple interview data set. A longitudinal data file for the entire panel (32 months) is released as a separate product only after all the individual wave products are released.

VI. Estimation

Three issues stand out in considering estimation from the longitudinal point of view: 1) defining the longitudinal universe; 2) defining longitudinal unit concepts; and 3) the treatment of missing data.

A. Defining the Longitudinal Universe

The target population of a longitudinal survey must deal with the consequences of birth, death, and mobility during the life of the study. Unlike cross-sectional studies that fix the population at a specific point in time and inferences made only about the time the sample was drawn, some longitudinal studies may be concerned

with drawing inferences about a nonstationary target population whose composition is changing over time.

Judkins et al (1994) describe three methods for defining a longitudinal universe. One method selects a specific time during the course of the study as the point that defines the universe. If the universe is defined at the time of sample selection, it is called a "cohort" study.

A second method of defining a longitudinal universe looks at more than one point in time. Several time points are selected,

each one defining a universe. The entire set of units defined by these different cross-sectional universes is included in the longitudinal universe.

A third method of defining a longitudinal universe includes only units common to all selected time periods; that is, in this approach one includes only those elements which were members of all cross-sectional universes. This universe contains only those units which did not enter or exit the survey universe and as a consequence is a static universe.

B. Defining Longitudinal Unit Concepts

Some longitudinal surveys, the prime examples being the NMCES and the SIPP, have undertaken the task of conceptualizing annual units of analysis using subannual data. Longitudinal analyses of a sample of households, families, or establishments must deal with the problems brought on by changes in the composition of these units. When a household or family splits up as a result of a divorce or separation, which of the two units is the same as the

original unit? Dicker and Casady (1982) and McMillen and Herriot (1985) discuss this topic for the NMCES and SIPP respectively.

Statistical estimation of longitudinal concepts is discussed by Ernst (1989) and Folsom, LaVange, and Williams (1989). Note that the acceptance of such concepts is not universal. Duncan and Hill (1985) argue that defining these concepts is unnecessary since all relevant analysis can be done at the individual level.

C. The Treatment of Missing Wave Data

In cross-sectional surveys, nonresponse is categorized in two ways: unit (total) nonresponse and item nonresponse. In longitudinal surveys, however a third type of nonresponse exists -- wave nonresponse. Wave nonresponse occurs when a sample unit does not respond in one or more waves of a longitudinal survey. In this situation, considerably more data are missing compared to the item nonresponse situation; however, considerably more data are available for use in nonresponse compensation strategies.

Solutions to the issue are not clear cut. Weighting and imputation, the methods use to compensate for nonresponse, have their own advantages and drawbacks (Kalton, 1986; Lepkowski, 1989).

Cox and Cohen (1985), Kalton and Miller (1986), and Mulvihill and Lawes (1980) have conducted empirical investigations into the relative quality of imputation and weighting as nonresponse compensation procedures in panel surveys. They found little difference when cross-sectional estimates were of interest. They show, however, that some forms of imputation are clearly inferior when longitudinal analysis is of interest. Singh, Huggins, and Kasprzyk (1990) advocate imputation for a restricted set of missing data patterns. This point of view is consistent with that expressed by Lepkowski (1989) where consideration is given to combined strategies of imputation and weighting.

VII. Persistent Issues in Longitudinal surveys

Longitudinal surveys theoretically offer the opportunity to measure change at the individual level as well as the opportunity to improve the overall measurement of data that are difficult to collect. In practice, two kinds of nonsampling error issues arise that play a significant role in longitudinal surveys: nonresponse and conditioning. A third issue, the role of nonsampling error in the measurement of gross flows, remains a complex and persistent problem for longitudinal surveys.

A. Attrition

A major concern as a longitudinal survey ages is the loss of representativeness of the sample due to nonresponse. Typically, the largest nonresponse occurs in the first several interviews with the wave-to-wave change in sample loss decreasing during the panel. Frequently it is not clear what the nature and character of sample loss is. See Kalton, Kasprzyk, and McMillen (1989) for illustrations of sample loss in selected longitudinal surveys.

The picture of nonresponse that we typically see is varied and likely dependent on factors such as, the frequency of interviews, difficulty in following or tracing respondents, sample composition, the length of the longitudinal survey, quality of the field staff, content of the questionnaire, and the efforts made to retain the sample. In general, nonresponse rates in longitudinal surveys increase over time as one would expect, but the rate of increase declines or stabilizes over time. However, we would not minimize the importance of the observation that cumulative overall nonresponse rates can be substantial over the length of a panel.

Analytic difficulties can occur if the nature of the nonresponse problem is not well understood. Too often, little is

done to describe the problem. Descriptive studies such as those done by McArthur (1988) and McArthur and Short (1985) for the SIPP provide some insight in understanding differences between respondents who participate in all waves with those who miss one or

more interviews.

Other studies aim to assess whether the current wave sample differs systematically from the original sample. See Rhoton (1986) for example. Another approach to understanding whether responses in later waves have a potential bias is to compare distributions of responses of subsequent respondents and nonrespondents to responses to questions asked earlier in the panel. This approach was taken by Petroni and King (1988) to study the effect of SIPP's cross-sectional nonresponse adjustment, variables in accounting for attrition in later waves. Finally, another approach was taken by the PSID. Becketti et al (1983) identified a particular analysis (e.g. regression analysis.) of data obtained from earlier waves and included variables indicating subsequent response status to provide evidence that nonresponse bias was not present in the PSID.

B. Time-In-Sample Bias

Time-in-sample bias refers to the concept that individuals' responses to the survey instrument may change due to the length of time an individual has been in the survey (that is, the number of times interviewed). Evidence of this bias has been found in estimates of unemployment, where higher rates of unemployment are observed among individuals in sample for the first or second time (Bailar, 1989). Other surveys have observed a time-in-sample phenomenon. See Neter and Waksberg (1964) and Woltman and Bushery (1975). In essence, this effect may occur because the early interviews in a longitudinal survey change either the respondents' behaviors or the way they answer the questions. Similarly the interviewers' behavior and approach to the respondents may change.

In practice, this bias is difficult to estimate because of a variety of changes taking place between rounds of a longitudinal survey, especially attrition. Unfortunately, even documentation of the existence of this bias is very difficult, requiring either a rotating panel design in which fresh replicate samples are added to the panel or an independent replicate sample implemented specifically to address this issue. Bailar (1989) and Kalton,

Kasprzyk, and McMillen (1989) review several studies.

C. Measurement Error

One of the presumed benefits of longitudinal surveys is their theoretical ability to measure change at the individual level. The difficulty, however, is that change measures are very sensitive to

individual measurement errors. Kalton, Kasprzyk, and McMillen (1989) identify aspects of panel surveys that may lead to measurement error: 1) simple response variability; 2) wave-to-wave changes in respondents; 3) changes in data collection mode; 4) wave-to-wave changes in interviewers; 5) wave-to-wave changes in questionnaires; 6) changes in the interpretation of questions over time; 7) wave-to-wave changes in coders; 8) imputation; 9) time-in-sample-bias; 10) matching interviews across time.

Any one of the above or all in some combination may make the

measurement of gross change problematic. A reporting error in the data at one point in time, corrected at another point in time, can lead to spurious measurements of change. Analytical difficulties in this type of analysis can be mitigated somewhat by sensitivity of the data collection organization to the problems; for example, detailed field edits, and proper documentation of amputations and identification of nonrespondents can help analysts in understanding their results.

VIII. Data User Issues

As discussed above, many statistical and measurement issues occur in the development of large-scale national surveys. In particular, many of these issues are exacerbated in longitudinal surveys in which repeated observations are taken on the same unit at several points in time. We believe the myriad of issues and their consequences places responsibility on the sponsors of such activities to provide substantially more documentation and guidance on the nature and extent of errors, both sampling and nonsampling

errors.

It is impossible to control and determine the effects of all the various sources of error; nonetheless, most of us, ourselves included, can make greater efforts at conducting evaluation of the quality of survey-data and documenting indications of nonsampling error that are likely to make a difference in longitudinal data analysis. Developing a "quality profile" summarizing in a convenient form what is known about the sources and magnitudes of errors in estimates should be done periodically for large multipurpose longitudinal surveys. See, for example, the SIPP Quality Profile (Jabine, King, and Petroni, 1990). Similarly, knowledge of the existence of both methodological and substantive research should be made available to the user community. The NLS has done this by publishing a bibliography of known research (Center for Human Resource Research, 1989).

Sponsors of longitudinal surveys should make available data quality evaluations in whatever form deemed appropriate. Duncan and Hill (1989) it a formal refereed article assessed the

representativeness of the PSID sample and compared their survey measures with program aggregates and aggregates from the Current

Population Survey. The SIPP now includes evaluations in the technical documentation of the file, if such evaluations are available when the file is released; otherwise, the evaluations are issued as "User Notes" after the release of the file.

Finally, several years ago the Social Science Research Council (SSRC) sponsored a research conference whose aim was to foster substantive analyses while providing information on the comparative strengths and weaknesses of several longitudinal data sets (U.S. Bureau of the Census, 1990). Conferences of this type, where program sponsors encourage comparative analysis of their data set, help engage the policy and research community to fully appreciate the strengths and weaknesses of each data base. One hopes that improved understanding of the data will result in better analysis. Sponsors of large complex surveys ought to be encouraged to foster

more of these kinds of exchanges.

Endnote

The Subcommittee enjoyed the benefits of the discussion of individuals who played active roles in several large Federal longitudinal surveys. Through these discussions, OMB Statistical Policy Working Paper 13 emerged. The paper we developed for the seminar on the "Quality of Federal Data" summarized in a rather lengthy fashion Statistical Policy Working Paper 13. Because of page constraints, the paper above is a considerably condensed version of the paper prepared for the seminar. The long version of the manuscript is available from Daniel Kasprzyk, Statistical Methods Division, U.S. Bureau of the Census, Washington, D.C. 20233.

References

Bailar, B.A. (1989), "Information Needs, Surveys, and Measurement

Errors," in Panel Surveys (D. Kasprzyk, G.J. Duncan, G. Kalton, M.P. Singh, eds.), John Wiley and Sons: New York, 1-24.

Becketti, S., W. Gould, L. Lillard, and F. Welch (1983), Attrition from the PSID, Santa Monica, California: Unicon Research Corp.

Boruch, R.F. and R.W. Pearson (1985), The Comparative Evaluation of Longitudinal Surveys, Social Science Research Council, New York.

Burgess, R.D. (1989), "Major Issues and Implications of Tracing," in Panel, Surveys (D. Kasprtyk, G.J. Duncan, G. Kalton, M.P. Singh, eds.), John Wiley and Sons: New York, 52-74.

Cantor, D. (1989), "Substantive Implications of Longitudinal Design Features,. The National Crime Survey as a Case Study," in Panel

Surveys (D. KasprZyk, G.J. Duncan, G. Kalton, M.P. Singh, eds.),

John Wiley and Sons: New York, 25-51.

Center for Human Resource Research (1989), NLS Annotated Bibliography: 1968-1989, Center for Human Resource Research, The Ohio State University, Columbus, Ohio.

Cox B.G. and S.B. Cohen (1985), Methodological Issues for Health Care Surveys, Marcel Dekker, New York.

Dicker, M. and R. Casady (1982), "A Reciprocal Rule Model for Defining Longitudinal Families for the Analysis of Panel Survey Data," Proceedings of the Social Statistics Section, American Statistical Association, 532-537.

Duncan, G.J. and D.H. Hill (1989), "Assessing the Quality of Household Panel Data: The Case of the Panel Study of Income Dynamics," Journal of Business and Economic Statistics, 7, 441-452.

Duncan, G.J. and M. Hill (1985), "Conceptions of Longitudinal Household: Fertile or Futile," Journal of Economic and Social

Measurement, 13, 361-375.

Duncan, G.J. and G. Kalton (1987), "Issues of Design and Analysis of Surveys Across Time," International Statistical Review, 55,97-117.

Duncan, G.J., F.T. Juster, and J.N. Morgan (1986), "The Role of Panel Studies in a World of Scarce Research Resources" in Survey Research Designs: Toward a Better Understanding of Their Costs and Benefits (R.F. Boruch and R.W.,Pearson, eds.). Lecture Notes in Statistics No.38, Springer Verlag, New York, 94-129.

Duncan, G.J. and J.N. Morgan (1984), "Behavioral Research with the Panel Study of Income Dynamics in Retrospect and Prospect," Vierteljahrshefte zur Wirtschaftsforschung, Dunker and Humblot, Berlin, 415-427.

Ernst, L.R. (1989), "Weighting Issues for Longitudinal Household and Family Estimates," in Panel Surveys (D. Kasprzyk, G.J. Duncan, G. Kalton, and M.P. Singh, eds.), John Wiley and Sons: New York,

139-159.

Folsom, R., L. LaVange, and R.L. Williams (1989), "A Probability Sampling Perspective on Panel Data Analysis," in Panel Surveys (D. Kasprzyk, G.J. Duncan, G. Kalton, M.P. Singh, eds.), John Wiley and Sons: New York, 108-138.

Grasso, J. and A. Kohen (1978), "The National Longitudinal Surveys Data Processing Systems," in The Survey of Income and Program Participation: Proceedings of the Workshop on Data Processing (D.

Kasprzyk, ed.), Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services, II-33-II-53.

Jabine, T., and K. King, and R. Petroni (1990), Survey of Income and Program Participation: Quality Profile, U.S. Bureau of the Census, Washington, DC.

Jean, A. and E. McArthur (1987), "Tracking Persons Over Time," SIPP

Working Paper No. 8701, U.S. Bureau of the Census.

Jean, A.C. and E. K. McArthur (1984), "Some Data Collection Issues for Panel Surveys with Application to the Survey of Income and Program Participation," Proceedings of the Section on Survey Research Methods, American Statistical Association, 745-750.

Judkins, D.R., D.L. Hubble, J.A. Dorsch, D.B. McMillen, and L.R. Ernst (1984), "Weighting of Persons for SIPP Longitudinal Tabulations," Proceedings of the Section on Survey Research Methods, American Statistical Association, 676-681.

Kalton, G. (1986), "Handling Wave Nonresponse in Panel Surveys," Journal of Official Statistics, 2, 303-314.

Kalton, G., D. Kasprzyk, and D.B. McMillen (1989), "Nonsampling Errors in Panel Surveys," in Panel Surveys (D. Kasprzyk, G.J. Duncan, G. Kalton, M.P. Singh, eds.), John Wiley and Sons: New York, 249-270.

Kalton, G., and M. Miller (1986), "Effects of Adjustments for Wave Nonresponse on Panel Survey Estimates," Proceedings of the Section on Survey Research Methods, American Statistical Association, 194-199.

Kish, L. (1987), Statistical Design for Research, John Wiley and Sons: New York.

Lepkowski, J. (1989), "Treatment of Wave Nonresponse in Panel Surveys," in Panel Surveys (D Kasptzyk, G.J. Duncan, G. Kalton, M.P. Singh, eds.), John Wiley and Sons: New York, 348-374.

McArthur, E. (1988), "Measurement of Attrition through the Completed SIPP 1984 Panel: Preliminary Results," Internal Bureau of the Census memorandum to D. Kasprzyk, March 4, 1988.

McArthur E. and K. Short (1985), "Characteristics of Sample Attrition in the SIPP," Proceedings of the Section on Survey Research Methods, American Statistical Association, 366-369.

McMillen, D.B. and R. Herriot (1985), "Toward a Longitudinal Definition of Households," Journal of Economic and Social Measurement, 13, 504-509.

Mulvihill, J. and M. Lawes (1980), "Imputation Procedures for LFS Longitudinal Files," Statistics Canada Internal Memorandum.

Neter, J. and J. Waksberg (1964), "A Study of Response Errors in Expenditure Data from Household Surveys," Journal of the American Statistical Association, 59,18-55.

Office of Management and Budget (1986), Federal Longitudinal Surveys (Statistical Policy, Working Paper No. 13), National Technical Information Service, PB86-139730.

Pearson, R. (1989). "The Advantages and Disadvantages of Longitudinal Surveys," Research in the Sociology of Education and

Socialization, Vol. 8, 177-199.

Petroni, R.J. and K.E. King (1988), "Evaluation of the survey of Income and Program Participation's Cross-Sectional Noninterview Adjustment Methods," Proceedings of the Section on Survey Research Methods, American Statistical Association, 342-347.

Rhoton, P. (1986), "Attrition and the National Longitudinal Surveys of Labor Force Behavior: Avoidance, Control, and correction," IASSIST Ouarterly, 10(2).

Singh, R., V. Huggins, and D. Kasprzyk (1990), "Handling Wave Nonresponse in Panel Surveys," paper presented at the Conference on "Survey Design, Methodology, and Analysis," University of Essex, Colchester, England, July 4-7, 1990.

U.S. Bureau of the Census (1990), Individuals and Families in Transition: Understanding Change Through Longritudinal Data. Papers presented at the Social Science Research Council Conference in Annapolis, Maryland, March 16-18, 1988. U.S. Bureau of the Census.

White, G.D. and H. Huang (1982), "Mover Follow-Up Costs for the Income Survey Development Program," Proceedings of the Section on Survey Research Methods, American Statistical Association, 376-381.

Woltman, H. and J. Bushery (1975), "A Panel Bias Study in the National Crime Survey," Proceedings of the Social Statistics Section, American Statistical Association, 159-167.

THE ADVANTAGES AND DISADVANTAGES OF LONGITUDINAL SURVEYS

Robert W. Pearson

Introduction

Longitudinal surveys have existed for some time in the social sciences. A quick scan of research would find them employed at least as early as 1928, when Stuart Rice studied the changing presidential preferences of Dartmouth college students (Rice 1928). Perhaps more readily recalled are Theodore Newcomb's classic studies of the effects of a liberal environment at Bennington College on young women from conservative families (Newcomb 1943). Panel designs were further extended when Paul Lazarsfeld an colleagues studied the 1940 U.S. presidential campaign through a stratified random sample of about 2,400 Erie County, Ohio, citizens (Lazarsfeld, Berelson, and Gaudet 1944).

Longitudinal studies became especially prominent in the 1960s and 1970s in the United States as the federal government turned its attention and resources to a domestic public agenda in which

research and evaluation played an increasing part. As the technology of data collection, storage, and analysis developed, so too did the call for and subsequent investment in longitudinal surveys. In the United States, for example, some 13 national longitudinal surveys were conducted in the 1950s while 64 surveys of this kind were carried out in the following two decades (Taeuber and Rockwell 1982). Panel studies quite simply permitted the study of change that other study designs (principally, cross sectional surveys) could not. These surveys were asked to evaluate the effects of social programs and to unravel the processes by which individuals change. The surveys facilitated the development of several fields of inquiry, including -- but not limited to -- labor economics, developmental psychology, voting behavior, and evaluation research. Conversely, theoretical and conceptual developments within these fields called for the use of longitudinal surveys.

The love affair with longitudinal data appears to have been short lived, however. This earlier affection has been replaced with an increasing appreciation of the limits of longitudinal surveys. For example, the editors of a volume on longitudinal analysis of labor market data would begin the volume provocatively by saying,

> Longitudinal data are widely and uncritically regarded as a panacea. Given the substantial cost of collecting such data, it is surprising that so little attention has been devoted to justifying the expense. The conventional wisdom in social science equates "longitudinal" with "good," and discussion of the issue rarely rises above this level (Heckman and Singer 1985, p. xi).

Similar questioning can be found in other fields of research. For example, Hirschi and Gottfredson assert in their review of research on the relationship between age and crime that "Funding agencies seem convinced by researchers that the longitudinal study is necessary for the proper study of crime" (Hirschi and Gotttfredson 1983, p 582). They argue instead that the causes of

crime are similar across age cohorts and that cross-sectional designs are likely to produce more knowledge per dollar of research than are longitudinal designs, which Hirschi and Gottfredson believe to be relatively more costly to conduct (Greenberg 1985; Hirschi and Gottfredson 1983, 1985; Murray and Erickson 1987).

The recent concern with longitudinal or panel surveys stems in part from the substantial investment in such data made during the past 20 years. There is also a suspicion that several important panel studies have reached or have gone beyond their maximum usefulness. Members of the policy and research communities now discuss these limitations as well as their comparative advantages in the reflective mood that was catalyzed by reductions in the data collection and social science budgets of the early part of the Reagan administration (Pearson 1985).

The purpose of this chapter is to review several of the strengths and weaknesses of longitudinal surveys that have emerged from these discussions. The chapter will make special note of the manner in which these research designs have been oversold on one

hand and underused on another. The chapter will discuss the several advantages and disadvantages of these instruments of social observation and draw attention to several claims about these data collection strategies that appear to be not well established, even if widely believed.

The principal point of the chapter is a relatively simple one -- which survey design is most appropriate for a particular purpose is a complicated function of a large number of factors. These include, but are not limited to, the use to which the data are put, the cognitive capacities and interests of respondents, legal and

ethical restraints on the study of human subjects; the nature and quality of theories or assumptions about social processes and behavior; and the inferential abilities of different research designs. Unfortunately, these simple points are too often ignored. The research literature and the decisions concerning the choice of research designs appear to have become increasingly interested in

choosing one rather than another design. Too little attention is paid to their fruitful combination, both within the survey research tradition -- the focus of this chapter -- and between this tradition and more qualitative research approaches.

The Advantages and Disadvantages of Longitudinal Surveys

Discussions of the advantages and disadvantages of a particular research designs are difficult to conduct in the abstract. This is so for several reasons. First, the discussion needs to be framed in a comparative perspective. Is the question one of the relative advantages and disadvantages of one longitudinal panel vs. another? Is it the relative merits of longitudinal vs. other designs? The former question faces secondary analysts of existing survey data. The latter question -- the principal focus of this chapter -- confronts those who sponsor and design research.

These are often two distinct, though overlapping, levels of

concern. Users of such data may find several surveys that are ostensibly relevant to a given topic, but have few tools for judging the equivalence of their measures. It is difficult to confirm, validate, and replicate research results across surveys. Paralleling the users' concerns, those who fund or design surveys must consider which studies to initiate, maintain, or terminate, and for what reasons? What combination of ongoing data collection programs will meet the present and future needs of research and public policy? Clearly, legitimate replication may be hard to distinguish from unnecessary redundancy. Although reliance on a single data source invites biased or inconclusive results, investments in similar or equivalent data series are likely to yield diminishing returns. Put briefly, each additional instrument may not lead to an equally valuable increment in knowledge.

Second, discussions of the advantages and disadvantages of longitudinal surveys (and other research designs as well) are difficult because their evaluation depends on a variety of conditions. These conditions include:

- The questions one wishes to answer.

- The skills and analytic competences of the investigator or the "user friendliness" of the data.

- The sample size, target population, substantive content, and design of the survey.

- The timeliness of the survey.

- The quality of the information.

- The documentation and dissemination of the data.

The evaluation of longitudinal surveys as well as other survey research designs also depends on subtler factors. For example, there are substantial costs associated with gaining a working

knowledge of the structure (and anomalies) of a large data set, costs that are not entirely transferable to another survey. These impediments to use are frequently confronted by analysts because many data collection programs do not devote resources to the creation of adequate documentation, data-based management structures, or the creation and distribution of users' access utility programs or constructed variables (David 1980, 1985).

Many analysts use several different longitudinal data sets in their research. But when they do, they are often aided by students, research assistants, and computational facilities that minimize the costs of doing so. That is to say, the use (and usefulness) of these, and other relatively large data sets or instruments cannot be considered apart from a wider set of instrumentalities which include students, assistants, training programs, computational and analytical technologies, instructional materials, and the availability of research funds for secondary analysis.

Standards or guidelines for the conduct of longitudinal

surveys exist (Bailar and Lanphier 1978; Boruch and Pearson 1988). These guidelines cannot be used, however, a priori to compare one longitudinal survey to another because such evaluation relies heavily on the uses to which the results or findings of the studies are to be put. If the findings of a study are known before the data are collected, there would be little need to conduct the study. (For similar conclusions concerning the intractability of judging the relative value of different data, see David and Peskin 1984.)

Comparisons of longitudinal with other research designs are difficult because many of the advantages and disadvantages of panel designs are shared by other designs. Nonresponse, confidentiality, data access are problems or concerns that face each (Boruch and Cecil 1979). Moreover, some disadvantages or difficulties posed by longitudinal surveys are also part of their strength. For example" how one defines and measures such ever-changing phenomena as the "family" is a problem that accompanies the increased ability to conceptualize and measure these dynamic phenomena (Koo 1985; Citro and Watts 1985 Citro, Hernandez, and Moorman 1986).

Equally important, some problems or disadvantages can be avoided or minimized if anticipated and if appropriate quality control mechanisms are built into the technology. For example, sample attrition of panel members can be reduced if sufficient attention and resources are devoted to collecting information from sample-respondents about friends or relatives who are likely to know where a respondent may move between waves of an interview. The effects of attrition can be monitored and, through imputational or weighting algorithms, compensated for during the analysis of the data.

The comparison of different survey research designs are often inappropriate because they tend to criticize one research design while more or less explicitly extolling the virtues of an alternative, as if their discussion was part of a debate in which it was important that one type of research design "win", while others "lose". We should instead begin by agreeing that different designs can in principle be combined to take advantage of their

relative merits and to overcome their relative disadvantages. One ought to ask what combination is most effective or efficient for answering one's questions rather than which one research design is best.

Although comparisons of the relative advantages of longitudinal surveys should be made cautiously, research and experience suggest that longitudinal surveys have several generic advantages and disadvantages that are relatively well established.

The advantages of longitudinal designs include, for example:

- The development of reliable measures of individual change. (Retrospectively collected data are subject to telescoping, memory decay, etc.) Similarly, these designs permit the measurement of subjective phenomena as current states rather than as recalled states. (Consider the difficulty of asking a respondent to rate his or her health or happiness four years ago.)

o The development of concepts that are characteristically dynamic rather than static. (The burgeoning multidisciplinary research on life-course perspectives owes part of its vitality to the creation and distribution of panel studies. See, for example, Baltes 1979; 1983.)

o Better descriptions of the dynamics of change. (The typical episode of family poverty or welfare receipt has been shown htrough panel data to be considerably briefer than was assumed in studies using repeated cross-sectional surveys. See, for example, Dunca et al. 1984; Corcoran et al. 1985; and Duncan, Hill, and Hoffman 1988.) Similarly, longitudinal designs permit the

estimation of individual levels and rates of transition between states or conditions for which cross-sectional

data may only provide gross or aggregate measures of group change.

- The ability to conduct analyses that control for unmeasured attributes of individuals, thus improving the ability to distinguish between the influence of enduring individual differences (e.g., race and gender) and the influence of having previously experienced the condition that is under investigation (e.g., previous unemployment leading to current unemployment).

The disadvantages of panel designs include:

- Nonresponse bias (especially through panel attrition) may be high and analytically troublesome. (Respondents for whom subsequent interviews cannot be completed may differ in analytically important ways from those who remain in the survey.)

- Response and learning effects (i.e., "panel effects") may

prejudice responses. (People who are interviewed about their voting behavior tend to vote more frequently thereafter.)

- Errors in the measurement of variables (and the correlation of these errors) and changes in the accuracy, reliability, and validity of such measures may spuriously create the appearance of change.

- Panel data, unless regularly refreshed or augmented, may provide useful or accurate estimates of the population from which the original sample was drawn, but not from the current population, which may be of interest.

- Panels always involve a moving target. Panel surveys of families, for example, must cope with movement into and out of families, the formation of new ones, and the dissolution of old.

Let us consider in more detail the first of the listed advantages and discuss several features often included in such

lists that are not well established. (For several discussions of the strengths and weaknesses of longitudinal surveys see, for example, Ashenfelter and Solon 1982; Boruch and Pearson 1988; Duncan, Juster, and Morgan 1984; Duncan and Kalton 1985; Fienberg and Tanur 1987a; and Subcommittee on Federal Longitudinal Surveys 1986.)

The limits of retrospection. The repeated observations of longitudinal surveys permit an investigation of change in phenomena

that can be measured in the present. They rely less than, say, a single cross-sectional survey design on the memory of respondents' prior conditions. This principal limitation of the ability of cross-sectional research designs to assess individual change is one of the major relative advantages of longitudinally designed studies.

Increasing evidence and recent theoretical developments in

cognitive psychology and survey methodology question, in more sophisticated ways than in the past, the trustworthiness of retrospective -- or memory-based -- responses to survey questions. Some research has found that certain kinds of memory-based data are flawed not only by temporal confusion and forgetting, but are systematically influenced by the respondent's current emotional state and beliefs about life and self.

Memory is basically reconstructive (cf. Bartlett 1932). And this reconstruction often involves the "top down" processing of the past that includes the development or use of scripts and narratives about the self or society, as well as the organization of details about the past. These scripts, schemata, self narratives, or, stereotypes define more or less coherent sets of beliefs around which more detailed images are actively (although not necessarily consciously) organized or distorted. If by virtue of sharing a. common culture, the respondent's schemata or theories of self or society are the same as those of the questioner (e.g., that adult mental distress follows from childhood problems), then research that relies on retrospective questioning techniques typically found

in cross-sectional surveys may be systematically biased in the direction the questioner expects. The resulting "theory validation" of retrospective studies may simply be the result of widely -- even if only implicitly -- shared cultural stories, narratives, stereotypes, or folklore whose accuracy is unknown and unprovable (Dawes and Pearson 1987).

Several studies substantiate-this conclusion. In two separate but similar experiments, for example, Conway and Ross (1984) examined randomly selected participants in a program designed to improve study skills and a control group of nonparticipants who indicated a desire to participate in the program but who were placed on a waiting list. Participants and control group members were questioned both before the beginning of the study skills program and at its conclusion. At both times, they were asked to assess their own study skills (e.g, how much of their study time was well spent, how satisfactory were their note taking skills, etc.) and the amount of time they studied. At the second interview they were also asked to recall what they reported during the first session concerning skills and study time.

At the initial interview, participants and control group members did not significantly differ on any measure of skill, study time, or on additional information about grades on a psychology

examination taken prior to the study skills program. Nor did these two groups differ in their recall of hours spent studying; there was a slight tendency for subjects in both conditions to recall studying less than they initially reported.

Recall of skills produced marked differences, however. Program participants recalled their study skills as being significantly worse than they initially reported. On the average, waiting list subjects recalled their study skills as being approximately the same as those they reported initially (p. 743). Participants in the study skills program appeared to exaggerate their improvement in a direction consistent with their theories of what ought to be -- taking a course should improve skills -- but they did so by retrospectively derogating their initial status.

They did not exaggerate their current skills, but reconstructed their memory of the past to combine: (1) a theory that they should have improved because of the instruction and (2) a relatively accurate assessment of their current level of skills. In both studies, the study skills program did not have a significant effect on academic performance, as measured by subsequent psychology examinations or average grades for the semester. The recall of past events and conditions were in error, and these errors were in a direction that was consistent with what the students thought that the past should have been as a result of their current conditions and prior participation in a study skills program.

Survey research has become increasingly aware of the distortions and misrepresentations of the past that are engendered by retrospective questions (cf., Turner and Martin 1984, p. 296; Sudman and Bradburn 1982, pp. 43-51; Schuman and Kalton 1986, pp. 644-647). In a recent validation study of employment-related information, for example, Mathiowetz (1986) and Duncan and Mathiowetz (1985) found that when a firm's employees were asked (in July 1983) whether they had been unemployed at any time during 1981

and 1982, 15 percent were in error concerning 1981 and seven percent were in error concerning 1982. Validation studies of retrospective reports have also observed substantial error in the recall of hospitalizations (Cannell, Fisher, and Baker 1965) and of victimizations (Turner 1972).

Obviously, longitudinal, research designs themselves may rely upon recall, as well as other cognitive processes. Their dependence on retrospective accounts is in part a function of the length of time between waves of a panel study and the need to measure experiences prior to the first interview. Panel designs have the advantage of providing opportunities to employ bounded recall techniques (asking respondents to recall events since the last interview) and to use information provided in a previous interview to reinstate prior context and to provide cues to facilitate their recall. The relative advantage of longitudinal surveys in this regard is often grounds for choosing this design over cross-sectional surveys.

Costs: A red herring. The list of strengths and shortcomings of longitudinal research provided above is not exhaustive. But it excludes their cost as a relative disadvantage. One can find numerous references to the expensiveness of longitudinal designs (cf., Murray and Erickson 1987, p 109), a belief that appears to be widespread. Unfortunately, this belief is not well established And the limited attempts to empirically assess the relative costs of longitudinal and cross-sectional surveys have shown under certain assumptions that longitudinal surveys may be less expensive than repeated cross-sectional surveys (Duncan, Juster, and Morgan 1984). No one can argue that surveys such as the PSID, HS&B;, and NLS72 are relatively expensive instruments to create and maintain. But these costs are largely a function of the number or special character of sample members required by the study; not necessarily their longitudinal design.

Surely, longitudinal surveys require the added expenses of tracing and tracking respondents as they move between waves of interviews, costs that are unique to a research design that follows

subjects through time. Locating and securing the cooperation of sample respondents during an initial interview, however, reduces the costs associated with drawing new sampling frames or screening households for the desired universe of sample members. These features also permit the use of relatively less expensive modes of administering subsequent waves of the, survey (e.g., phone, mail back questionnaires) than may be required in cross-sectional national samples.

Evaluating the relative costs of longitudinal surveys depends a great deal on what one chooses to compare them to. In this regard, we are faced with the difficulties posed by the proverbial comparison of apples and oranges. It is only suggestive -- but nonetheless in opposition to the belief about their expense -- that the average field cost of completed interviews of the 1987 General Social Survey of NORC was $400. The average costs of each completed interview of the 10th wave of the Youth Cohort of the National Longitudinal Survey of Labor Market Experience was $333 (Carter 1987). Similarly, the total cost of the first year of interviews of the National Post Secondary Student Aid Study of 1987 was $7.2 million; its first year follow-up is currently estimated

to cost $3.0 million (Carroll 1987).

The comparative costs of different survey designs compound the difficulty of simultaneously weighing their advantages and disadvantages. A relatively ambiguous attitude toward panel surveys in assessing the effects of job training programs, as suggested for example by Heckman and Robb (1985), could be turned on its head by altering assumptions about the comparative costs of these different designs. Indeed, if one assumes relatively equal expenses, or cost advantages to panel designs, it would be prudent to select panel rather than cross-sectional designs (holding a great many other factors constant) because panel designs permit the use of a wider variety of statistical and theoretical assumptions. The application of a wider range of assumptions provides a useful means of testing how sensitive conclusions are to different assumptions. That is to say, one can Analyze a panel study as if it was a repeated cross-sectional design, but not vice versa.

The largest costs of such studies lie in the creation and maintenance of the organization that is required to collect the data and in the burden which such surveys impose on their respondents. This commitment of resources is largely fixed and shared with other large survey-based designs; whether panel, experimental, or cross-sectional. Ongoing instruments of data collection are more likely to present opportunities for linkage or augmentation with side studies, experiments, topical modules than are "one-shot" data collection programs, which single cross-sectional surveys can often be, an advantage to which we return below.

Causal Inference: An oversold advantage. Longitudinal survey designs do not, however, as is often incorrectly claimed, permit unequivicable inferences about causation. Surely, the temporal dimension of longitudinal surveys provides strong priors for assuming that a leads to or causes b if a is observed to occur before b. But there are several dangers in making such strong causal inferences from panel designs. One's anticipation of future

events can influence current behavior, for example. And selection biases (i.e., people found in a program often differ in unmeasureable or unmeasured ways from nonparticipants) invariably trouble the estimation of program effects.

Heckman and Robb (1985), for example, examined three survey designs and associated econometric techniques to determine whether one was "better" than the others in assessing the consequences of a public policy interventions, e.g. a, youth employment training program. They compared (1) single retrospective cross-sectional, (2) repeated cross-sectional, and (3) panel designs and their corresponding analytical techniques. Heckman and Robb showed that each design and corresponding analytical technique requires untestable assumptions in evaluating the earnings effects of participation in training programs. Their research argued that many of the assumptions of cross-sectional analytical techniques were no more or less justifiable than those upon which the panel

designs were based, although some assumptions could be -- although too infrequently are -- the object of independent study.

Although panel studies permit one to trace spells and transitions and to order conditions in sequences that suggest causation, longitudinal surveys cannot do so without the aid of assumptions. This point is forcefully illustrated by Lord's paradox (1967, 1968, 1973) and its discussion by Holland and Rubin (1986).

Attempts to understand how different models generate the same data, or how similar models can generate or represent different data, have produced a greater sensitivity to the problems of making causal inferences from the research designs of panel studies (and other observational designs such as matched comparison groups and cross-sectional surveys). Fraker and Maynard (1985, 1987), for example, analyzed data from several sources to compare the estimated earnings effects from participating in an employment and training program. They compared estimates of training effects derived from (1) control groups of the National Supported Work Demonstration program that were selected in accordance with

experimental research designs and (2) comparison groups constructed from the Current Population Survey. Matched comparison designs involve the creation of samples of respondents (typically drawn from such surveys as the Current Population Survey) who are similar in important respects to the participants in a program that one seeks to evaluate. These research designs are common in program evaluations in part because information about program participants is regularly collected at their enrollment or discharge from the program. Experiments in which a number of eligible individuals are randomly precluded from participating in a program (and later compared to those who are allowed to enroll) on the other hand are at times difficult to conduct or proscribed by ethical and legal considerations.

Fraker and Maynard's comparisons of experimental versus nonexperimental estimates of training program effects on annual earnings showed that comparison group procedures and analytical models produced estimates of large negative effects on the earnings of youth both during the program's employment period and after. The experimental design revealed estimates of program earnings for

youth that were modestly positive during the program, and negligible thereafter. Comparisons of the effects of training on AFDC recipients revealed similar positive effects between the experimental and matched comparison designs. The differences in results between unemployed youth and AFDC recipients suggest that the greater earnings and employment variability of youth may result in more biased selection into the employment program, which in turn makes the task of defining a comparison group and an analytical model more difficult. Corroborative evidence to this work can be found in LaLonde (1986).

The implication of these marked differences in results is that the longitudinal and cross-sectional designs (or other nonexperimental designs) alone do not permit one to unravel the many causes and consequences of social and economic change or of program interventions. Perhaps more disturbing to those who must rely on such data, these research designs may produce the wrong answer when the behavior of the population under study is

undergoing considerable change (as are the employment activities of youth). Experimental designs are superior to panel designs in making causal inferences.

Rarely is the use of or investment in data made "on purely statistical grounds" alone, however. In addition to costs, choices are constrained by legal, ethical, and administrative considerations (Riecken et al. 1974). Considerable experience (much from studies of states and municipalities) has produced a greater appreciation of many of the difficulties of importing laboratory-oriented experimental designs into the field. It is often difficult, for example, to sustain the separation of treatment and control groups in the field. Moreover, some of the problems associated with the design and implementation of panel studies, such as attrition, apply equally to experimental designs (Betsey, Hollister and Papagiorgiou 1985). Experiments, are useful for assessing the relative differences among program variations on a common set of outcome variables. But experimental designs have their own scientific and administrative shortcomings. For example, treatments are often limited to a narrow set of variables and to

specialized samples, and so their results may be of limited generalizability. Moreover, they are of ten difficult to administer and require substantial managerial skills to conduct. These limitations, among others, have retarded the use of experimental designs in the social sciences and in the evaluation of government programs. On the other hand, the design, implementation, and analysts of field experiments is possible, and some evidence exists of a renewed interest in them (c.f. Maynard 1987; Bloom, Borus, and Orr 1987; and Cottingham and Rodriguez 1987).

Coupling experimental and longritudinal designs: a not fully realized potential. Longitudinal surveys are often an appropriate technology for describing the timing, duration, and sequence of individual change. And they are often better in this regard than alternative nonexperimental observational research designs because of the problems these alternatives confront when relying on retrospective measures of past conditions. Under certain conditions, longitudinal surveys appear to be no more expensive to conduct than repeated cross-sectional surveys. But their ability to draw causal inferences has in general been overdrawn. Although temporal order provides prima facie evidence for causation, it is

insufficient.

Increasingly, the research community is considering the fusion of longitudinal and experimental survey designs in which randomly

assigned treatments or interventions are given to some members of an ongoing longitudinal survey. Coupling experiments and longitudinal surveys capitalizes on the strongest merits of each design. That is, one obtains both the information produced by national probability samples -- often conducted over a considerable length of time -- and the information produced by smaller comparative experiments in which causal inferences are more appropriately deduced. Insofar, as the experiments can be adjoined systematically, their generalizability will be enhanced.

Joining experiments to ongoing longitudinal surveys also permits one to use the experiments to calibrate estimates of program effects that are derived entirely from the longitudinal

survey. That is, the biases engendered by using estimates that are based on longitudinal data can be assessed, and periodically corrected, through controlled experiments. Thus, longitudinal studies are likely to be more policy-relevant and less ambiguous with respect to biases in estimating program effects. Experiments are likely to benefit from their greater generalizability, lower costs, and more manageable administration.

There is no doubt about the need for social experiments in understanding change (Berk et al. 1985). The National Academy of Sciences' Committee on Youth Employment Programs, for example, examined major studies to understand whether one could draw firm conclusions about program effects from earlier research. The committee concluded, among other things, that longitudinal surveys ate no substitute for randomized experiments when the object is to estimate the effectiveness-of youth employment programs. Moreover, they urged the use of randomized experiments for this purpose (Betsey, Hollister, and Papagiorgiou 1985).

Coupling randomized designs to longitudinal surveys can also

be traced to a technical advisory committee for employment program evaluation appointed by the U.S. Department of Labor. The DOL sought to learn whether analyses of manpower programs based on conventional longitudinal surveys lead to adequate estimates of program effects. Adequacy was assessed, for example, by comparing estimates of effects based on longitudinal surveys against estimates based on randomized trials. The conclusion of this exercise was that the two estimates are not always in accord. Indeed, they differ remarkably depending on what population is the subject of inquiry (Fraker and Maynard 1985, 1987).

Obviously, changes in standard practices that are suggested here would introduce costs and difficulties, at least until their implementation permitted organizations to identify and remedy the problems that naturally arise with any new technology. And surely, there are a number of programs that could not be evaluated through such a coupling of designs because of the nature of the intervention or the limited number or location of possible respondents even in relatively large national longitudinal surveys.

Randomly varying policy responses to violent domestic disputes could not be comfortably grafted onto High School and Beyond (HS&B;), for example. Unfortunately, the development and Application of this general strategy has yet to adequately tested.

Summary and Conclusion

Longitudinal surveys are an important technology for the measurement of individual change and development. Considerable resources have been devoted during the last two decades to their creation and maintenance. These instruments of social observation have contributed a great deal to the development of several fields of inquiry, and promise to continue to do so.

Recent years have seen a growing restlessness with these research designs, however. Their limitations especially those related to causal inference -- are increasingly recognized, although their relative strengths have continued to argue for their use as important new instruments of data collection. Their support

and criticism is the healthy consequence of the continual scrutiny that a principal tool of social analysis should undergo.

Their relative strengths, however, have not yet been systematically and regularly coupled with the strengths of another research design -- experiments. The promise of combining these methods and of moving beyond the discussion of the strengths and weaknesses of a particular research design still lays before us.

References

Ashenfelter, O. and G. Solon. 1982. Longitudinal labor market data: Sources, uses, and limitations. A paper presented at a conference sponsored by the National Council on Employment Policy, An Assessment of Labor Force Measurements for Policy Formulation, Washington, D.C. (June).

Bailar, B. A. and C. M. Lanphier. 1978. Development of Survey Methods to Assist Survey Practices. Washington, D.C.: American Statistical Association.

Baltes, P. B. 1979. Life-span developmental psychology: Some converging observations on history and theory. In Life-Span Development and Behavior. Vol. 2. eds. P. B. Baltes and 0. B. Brim, Jr. New York: Academic Press.

Bartlett, F. C. 1932. Remembering: A Study in Experimental and Social Psychology. Cambridge: Cambridge University Press.

Berk, R. A. et al. 1985. Social policy experimentation. Evaluation Review 9:387-429.

Betsey, C., R. Hollister, and M. Papagiorgiou. 1985. Report of the Committee on Youth Employment Programs. Washington, D.C.: National Research Council.

Bloom H. S., M. E. Borus, and L. L. Orr. 1987. Using random assignment to evaluate an ongoing program: The National TTPA Evaluation. A paper presented at the annual meeting of the American Statistical Association, San Francisco (August 17-20).

Boruch, R. F. 1975. Coupling randomized experiments and approximations to experiments in social program evaluation. Sociological Methods and Research. 4:31-53.

Boruch, R. F. and H. W. Riecken. eds. 1975. Experimental Testing of Public Policy. Boulder, Colorado: Westview Press.

Boruch, R. F. and J. S. Cecil. 1979. Assuring the Confidentiality of Social Research Data. Philadelphia, PA: University of Pennsylvania Press.

Boruch, R. F. and R. W. Pearson. 1988. Assessing the Quality of Longitudinal Surveys. Evaluation Review In press.

Cannell, C. F., G. Fisher, and T. Bakker. 1965. Reporting of hospitalizationin the Health Interview Survey, Vital and Health Statistics, Series 2, No. 6. Washington, D.C.: U.S. Government Printing Office.

Carroll, D. 1987. Personal communication, December 8, 1987.

Carter, W. 1987. Personal communication, December 7, 1987.

Citro, C.F., and H. W. Watts. 1985. Patterns of Household Composition and Family Status Change. Paper presented to the American Economic Association, New York, New York.

Citro, C. F., D. J. Hernandez, and J. E. Moorman. 1986. Longitudinal household concepts in SIPP. Paper presented to the American Statistical Association, Chicago, Illinois, May 30.

Conway, M. and M. Ross. 1994. Getting what you want by revising what you had. Journal of Personality and Social Psychology 47:738-748.

Corcoran, M. E., G. J. Duncan, G. Gurin, and P. Gurin. 1985. Myth and reality: The causes of persistence of poverty. Journal of Policy Analysis and Management.

Cottingham, P. and A. Rodriguez. 1987. The experimental testing of the Minority Female Single Parents Program. A paper presented at the annual meeting of the American Statistical Association, San Francisco (August 17-20).

David, E. L. and H. M. Peskin. 1984. Theory of an optimal database. Review of Public Data Use 12:45-53.

David, M. 1980. Access to data: The frustration and utopia of the researcher. Review of Public Data Use 8:327-337.

Dawes, R. M. and R. W. Pearson. 1987. The efffect of the present on retrospective data: Measuring the then, now. New.York: Social Science Research Council, mimeo.

Duncan, G. J., F. T. Juster, and J. N. Morgan. 1984. The role of panel studies in a world of scarce research resources. In The Collection and Analysis of Economic and Consumer Behavior Data: In Memory of Robert Ferber, eds. S. Sudman and M. A. Spaeth. Champagne, Illinois: Bureau of Economic and Business Research.

Duncan, G. J., R. Coe, M. E. Corcoran, M. Hill, M. S. Hoffman, and J. M. Morgan. 1984. Years of Plenty, Years of Hope. Ann Arbor:

Survey Research Center, University of Michigan.

Duncan, G. J. and G. Kalton. 1985. Issues of design and analysis of surveys across time. A paper presented at the centenary session of the International Statistical Institute, Amsterdam.

Duncan, G. J. and N. A. Mathiowetz. 1985. A validation study of economic survey data. Ann Arbor, MI: Institute for Social Research, mimeo.

Duncan, C. J., M. S. Hill, and S. D. Hoffman. 1988. Welfare dependence within and across generations. Science. 239:467-471.

Fienberg, S., B. and J. Tanur. 1986. From the inside out and the outside in: Combining experimental and sampling structures. Technical Report No. 373, Carnegie Mellon University (December).

_____. 1987a. The design and analysis of longitudinal surveys: Controversies and issues of costs and continuity. In Designing Research With Scarce Resources, eds. R. F. Boruch and R. W. Pearson. New York: Springer-Verlag.

_____. 1987b. Experimental and sampling structures: Parallels diverging and meeting. International Statistical Review 55:75-96.

Fraker, T. and R. Maynard. 1985. The use of comparison group designs in evaluation of employment related programs. Princeton, N. J.: Mathematica Policy Research, mimeo.

_____. 1987. The Study of comparison group designs for evaluations of employment-related programs. The Journal of Human Resources 22: 194-227.

Greenberg, D. F. 1985. Age, crime and social explanation. American Journal of Socilogy 91, 1-21.

Heckman, J. J. and R. Robb, Jr. 1985. Alternative methods for evaluating the impact of interventions. In Longitudinal Analysis of Labor Market Data, eds. J. J. Heckman and B. Singer, 156-246. New York: Cambridge University Press.

Heckman, J. J. and B. Singer. eds. 1985. Longitudinal Analysis of Labor Market Data. New York: Cambridge University Press.

Hirschi, T. and M. Gottfredson. 1983. Age and the explanation of crime. American Journal of Sociology 91, 359-374.

____. 1985. Age and crime, logic and scholarship: Comment on Greenberg. American Journal of Sociology 91, 22-27.

Holland, P. W. and D. B. Rubin. 1986. Research designs and causal inferences: On Lord's Paradox. In Survey Research Designs: Towards a Better Understanding of Their Costs and Benefits. eds. R. W. Pearson and R. F. Boruch, 7-37. New York: Springer-Verlag.

Koo, H. 1985. Short-term change in household and family structure. Paper presented to the American Statistical Association, Las Vegas, Nevada.

LaLonde, R. 1986. Evaluating the Econometric evaluations of training programs with experimental data. American Economic Review

76 (4):604-20

Lazarsfeld, P. F., B. Berelson, and H. Gaudet. (1944) 1960. The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign. 2nd ed. New York: Columbia University Press.

Lord, F. M. 1967. A paradox in the interpretation of group comparisons. Psychological Bulletin 68:304-305.

_____. 1968. Statistical adjustments when comparing preexisting groups. Psychological Bulletin 72:336-337.

_____. 1973. Lord's paradox. In Encyclopedia of Educational Evaluation. Anderson, S. B. et al. San Francisco: Jossey-Bass.

Mathiowetz, N. A. 1986. Episodic recall and estimation:

Applicability of cognitive theories to survey data. Paper presented at a Seminar on the Effects of Theory-Basedf Schetas on Retrospecitve Data, June 26-28, New York: Social Science Research Council.

Murray, G. F. and P. G. Erickson. 1987. Cross-sectional versus longitudinal research: An empirical comparison of projected and subsequent-criminality. Social Science Research 16, 107-118.

Newcomb, T. M. (1943) 1957. Personality and Social Change: Attitude Formation in a Student Community. New.York: Dryden.

Pearson, R. W. 1985. The changing fortunes of the U.S. statistical system, 1980-1985. Review of Public Data Use 12:245-269.

Rice, S. A. 1928. Quantitative Methods in Politics. New York: Knopf.

Riecken, H. W. et al. 1974. Social Experimentation. New York:

Academic.

Schuman, H. and G. Kalton. 1986. Survey methods. In The Handbook of Social Psychology, 3rd ed. eds. G. Lindzey and E. Aronson, 635-697. Reading, MA: Addison-Wesley.

Subcommittee on Federal Longitudinal Surveys, Federal Committee on Statistical Methodology. 1986. Federal Longitudinal Surveys. Washington, D.C.: Office of Management and Budget.

Sudman, S. and N. M. Bradburn. 1982. Asking Questions. San Francisco,. Jossey-Bass.

Taeuber, R. and R. C. Rockwell. 1982. National social data series: A compendium of brief descriptions. Review of Public Data Use 10:23-111.

Turner, A. G. 1912. The San Jose methods test of known crime victims. Washington, D.C.: National Criminal Justice Information and Statistics Service, Law Enforcement Assistance Administration, U.S. Department of Justice.

Turner, C. F. and E. Martin. eds. 1984. Surveying Subjective Phenomena, Volume 1. New York: Russell Sage Foundation.

LONGITUDINAL ANALYSIS OF FEDERAL SURVEY DATA

Patricia Ruggles

Joint Economic Committee

I. Introduction

Longitudinal panel data provide a unique opportunity to examine patterns and sources of economic and demographic change at

the individual and family level. These data are relevant to a host of policy issues, from the assessment of welfare program participation to an understanding of patterns of health care usage or of the determinants of retirement. Many policy issues require some understanding of the factors that lead up to a particular event, or of the consequences that stem from it. Without repeated observations of the individuals concerned, however, such factors and consequences can only be inferred. Thus, our increasing store of longitudinal panel data holds the potential for major breakthroughs in our understanding of the basic determinants of economic and demographic change as they affect individuals and families over time.

Unfortunately, however, many of our longitudinal data sets have been somewhat under-used by researchers so far, especially compared to similar cross-sectional surveys. To some extent this under-usage may simply stem from the fact that many of these data sets are still fairly new -- researchers need a chance to become familiar with the opportunities offered by these new sources of information. A more fundamental problem, however, is that to an

analyst whose primary research experience is with cross-sectional microdata, a longitudinal panel of microdata on families and individuals can be rather intimidating.

The purpose of this paper is to provide some guidance to users and potential users of longitudinal data sets who are trying to sort out appropriate approaches to the problems of analyzing longitudinal panel data. This paper does not attempt to offer any new insights into the methodologies available to estimate the determinants of change (or stability) in a given variable or set of variables over time, nor are the theoretical issues underlying these methodologies addressed in any detail. Instead, the paper is designed to be a much more basic "how to" guide, focusing on the most fundamental choices that must be made by the analyst in undertaking a project involving the use of longitudinal data to examine the economic circumstances of families and individuals.

The major focus of this paper is on specific methods of making comparisons across time, with emphasis on matching the outcome measures and statistical techniques chosen to the basic research

question being asked. For many policy issues fairly simple outcome measures may be perfectly appropriate, but it is important to understand the measurement implications of alternative choices in order to avoid misinterpreting one's results.

II. Making Comparisons Across Time

The major purpose of a longitudinal research file is of course to facilitate the analysis of change over time. There are three major types of time-related analysis that are commonly carried out with such files, and there are some specific methodological issues that pertain to each.

Comparing Two Points in Time

The simplest type of time-related analysis the comparison of data from two discrete points in time -- does not actually require a complete longitudinal data file at all. The major advantage of this type of analysis is that it is relatively simple to implement and can often yield a great deal of useful information, particularly for questions that focus on rates of turnover in a specific variable. This method is very commonly used with many different longitudinal data sets -- several examples of such analyses can be found for PSID data in the Institute for Social Research's volume of PSID research results entitled Years of Poverty, Years of Plenty, for example. Other examples include Alan Fox's study using RHS data which examined income changes at retirement, and the SIPP-based study produced by Jack McNeil and his colleagues at the Census Bureaus that considered how many of those poor in 1984 were still poor in 1985.

The major drawback of this method of making comparisons across time is that the outcome variables are sometimes quite sensitive to the specific time periods chosen-for analysis, and there is no way

for the analyst to determine this if only two points in time are examined. Further, such comparisons are valid measures of change among those who already have a given characteristic, but cannot be used to determine the distribution of durations of a particular state among all those who enter it.

For example, using this method we can tell what the total remarriage rate for all divorced women is over a given period of time, but we cannot determine the average amount of time that women spend between marriages, because we do not know when those who were already divorced at the time of the first observation got divorced, and we have no distribution of remarriage probabilities by duration of divorce to use in forecasting future remarriage rates for those who have not yet remarried. Indeed, we cannot even determine if

the remarriage rate is sensitive to the amount of time that has elapsed since the divorce. In other words, to the extent that the determinants of changes in state are themselves time-related, they may be difficult to observe if one must rely on simple "before and after" comparisons.

Examining Transition Events

A second approach to making comparisons across time, therefore, is to examine transitions between two states directly. By focusing on the transition itself one can more closely examine its association with other factors that may not be observable in a simple before and after comparison. This is helpful both in considering the effects of the transition on other variables and in estimating a causative model of the determinants of the transition itself.

To illustrate this point, let us reconsider the analysis of divorce discussed briefly above. If the analyst is interested not

only in the determinants of the divorce transition, but also in its impacts, a simple comparison of two points in time may be doubly misleading. For example, family income may dip temporarily at the time of divorce as the family changes from one household to two. Eventually, however, as the two households make post-divorce adjustments in employment and living arrangements, income is likely to recover at least somewhat. Estimates of the impact of the divorce on income and poverty status for the various family members may be quite sensitive to both the unit definition used to compute income (as discussed in the last section) and, to the specific timing of two income observations compared to the divorce itself.

In a case like this, examination of income or poverty status over a longer period leading up to and then following the transition will give a better picture of its actual impacts. For this type of examination it is necessary to have a longitudinally linked file with the transition, flagged, but if such a file is available a descriptive analysis of this type is quite straightforward to perform. Similarly, the transition flags

themselves can be used as explanatory variables in a larger model of change over time as it affects some other variable. The recent paper by Suzanne Bianchi and Edie McArthur on the impacts of marital disruptions on children's economic status illustrates a transition analysis of this type.

Considering the determinants of a given transition is also facilitated by the availability of a linked longitudinal file. For example, probit-type regression models can be used to examine the probability that a given transition will take place, subject to the various other characteristics of the cases in question. In analyzing divorce, for example, one might want to consider the impacts of the spouses' employment statuses in the period before the divorce on the probability that they will become divorced. In other cases, a broader set of dependent variables may be necessary -- those leaving a given state may have more than one alternative

option. The work by Alan Gustman and Thomas Steinmeier on retirement probabilities as observed in the RHS offers a good example of a fairly complex application of this type of transition analysis.

With a linked longitudinal file, the conditional probability of a given event such as divorce or retirement can be calculated fairly easily for specific population subgroups, and/or conditioned on specific events, using readily available software packages such as SAS. Again, however, such an approach can be misleading if the determinants of the transition in question are themselves time-related -- if for example, the previous duration of the marriage or even the length of the unemployment spell are important determinants of the probability of divorce.

These duration-related issues, then, are potentially problematic with either a straightforward comparison of data from two points in time or with a more sophisticated analysis of specific transitions. Although it is sometimes possible to shoehorn duration-related information into one's transition

analysis - one could create separate dummy variables for short and long unemployment spells in the above example, for instance - this is a rather ad hoc approach that is likely to leave many unanswered questions. In addition, in many cases one is interested not only

in the transition event itself, or even in its impact on other events, but also in the expected duration of the new state that it creates. One wishes to know, for example, how long someone who enters poverty may be expected to remain poor, or how long someone who loses a]ob may be expected to remain unemployed. Questions of this type require some type of duration analysis.

Analyzing Data on Duration

There are many possible approaches to questions of duration,

and alternative approaches can produce quite different and even seemingly contradictory statistics. The confusion generally results from differences in the population to which the duration estimate applies. The two major possibilities are cohort-based estimates, which typically apply to all those observed in a given state at a point in time, and spell estimates, which apply to all those observed to enter the state within a given span of time.

To illustrate these possibilities, consider the case of welfare program participation. A point-in-time or cohort-based estimate of welfare durations will ask a question like "How long have those who are currently receiving welfare been on the program?" This question has been phrased retrospectively, but it can also be put in a prospective form: "How long are those currently on the program likely to remain on in the future?" In either case, the base population being considered is all those on the program at a given point in time. Such estimates are therefore relatively easy to line up with cross-sectional estimates of the total population on welfare, which are of necessity also point-in-time estimates. Estimates of this type are very useful for a number of purposes -- for example, estimating the future costs of

the current welfare caseload (although obviously to get total costs one would also have to account for new welfare entrants).

One useful way to think about estimates of this type is as an examination of the experiences of a particular cohort -- a group that all happened to be in a given state at a given point in time. The NLS, for example, is designed with just such applications in mind. It is possible to use these data to examine the subsequent experiences of several distinct demographic cohorts selected at specific points in time -- teenagers, men nearing retirement, women in their middle years. It is even possible, with the new youth cohort to link up families across generations, and to relate young women's experiences to those of their mothers, as Peter Gottschalk has done recently for welfare recipients, for example. A similar type of application using PSID data is Frank Levy's path-breaking 1977 paper on the "underclass," which traced the subsequent experiences of a cohort of those in poverty in 1967.

Cohort-type analyses are very useful for many policy questions, but it is important to be aware of their limitations in applying them to policy analyses. Specifically, because they apply only to those in the state at a given time, such analyses are sometimes difficult to generalize to the population as a whole, or even to the experience of all those who may pass through the state over a period of time.

What a point-in-time estimate cannot do, in other words, is answer questions like "How long will a typical person entering welfare stay on the program?" Such a question refers not to the population on the program at a point in time, but rather to the population entering the program. Although that may seem like a subtle distinction, in fact these two populations are likely to be very different if there is any significant variation at all in spell durations within the population as a whole. Those who are on welfare at a point in time are likely to have much longer spell

durations, on average, than the typical entrant, because those with longer spell durations are more likely to be in the welfare population at any particular point in time.

To see this point, consider a very simple example. Suppose the population of interest consists of 13 people, one of whom is in the state under consideration for one year, and twelve of whom are in that state for one month each. Further suppose that these twelve one-month spells are distributed so that one occurs in every month of the year. At any given point in time, therefore, the total population in the state being considered will consist of two people, one who is in a one-month spell, and one who is in a twelve month spell. A point-in-time analysis conducted any time after the first month will therefore conclude that 50 percent of the observable population reports a spell of more than one month. An analysis based on all entrants observed during the year, however, will find that only one-thirteenth of the population reports a spell of more than one month. Clearly, if the reasons for these differences in estimates are not well understood, they could lead to very different conclusions about the prevalence of long spells.

Many of the most useful and interesting questions that can be addressed using a longitudinal database are questions that relate to duration. In any type of duration analysis, however, it is

necessary to be sensitive to the issue of censoring. Inevitably, there will be some spells that start before the beginning of the observation period or that end after the panel has come to an end. Further, there will be some cases that join the panel with a spell already in progress or leave the panel before one has ended. These spells cannot simply be ignored, since of course longer spells are more likely than short ones to be censored and ignoring this problem will therefore produce biased estimates.

An alternative approach that unfortunately is fairly often used by analysts who have not completely thought through the

problem of spell censoring is to mix together all one's observations over a given span of time, whether they apply to completed spells or to those that are only partially observed. This produces results that are confusing and even potentially misleading, since it is easy to misclassify spells that are only partially observed as short spells, producing misleading estimates of average spell durations.

The measure of the "persistently poor" produced by Duncan et al. using PSID data is an example of this approach, and illustrates some of its problems. In this study, the base population was defined as all those in the population during the ten year observation period -- not just those in poverty in a particular year, as in Levy's study. Duncan et al. then defined the "persistently poor" as those poor for at least eight out of the ten years. They went on to calculate the proportion of the total population that was "persistently poor" simply by dividing the number of people observed in poverty for at least eight years by the total population observed.

The problem with this approach is that some people who are poor for less than eight years during the observation period are nevertheless in the midst of spells of poverty that will total eight or more years -- but unfortunately some of those years happen to fall outside the observation-period. Thus the true number of individuals in the sample who were actually poor for at least eight out of ten years (at least some of which fell in the sample period) cannot be estimated using these data. Estimates of the proportion of those observed who experience long poverty spells will be understated, because some spells that appear short are in fact longer, but they simply haven't been completely measured. At the same time, however, because these estimates mix together people who were poor in different years, they also cannot be used to predict, say, what proportion of those poor in a given year will still be poor eight years later.

A preferable approach to the problem of estimating spell durations when some observations are censored is to use some sort

of survival analysis technique. Under this methodology, a survival function for a given type of spell is estimated based on the cumulative distribution of observed spell durations. In other words, in order to compute the probability that a spell of welfare participation, for example, will end in its sixth month, conditional on its having lasted for the first five months, one must include all cases known to have lasted at least five full months, whether or not their eventual disposition is known.

By including all spells -- even those whose endings will eventually be unobserved -- for as long as information on their status is available, systematic biases related to spell duration will be minimized. At the same time, censored spells are essentially treated as if they had the same distribution of durations as spells with otherwise similar characteristics whose endings are observed. Under this methodology, censored spells do not pull down the estimated median spell duration, for example, as they do when the problem of censoring is not recognized. It is worth noting, however, that this approach assumes that censored spells are not systematically different from uncensored spells (except in ways fully captured in the X vector of explanatory variables), and that spells that occur at the beginning of the observation period are not systematically different from those starting nearer the end. To the extent that external events -- for example, legislative changes or changes in the state of the economy affect spell durations over time, analysis techniques that pool spell observations across the period at a whole may be misleading.

This approach does allow the contribution of a variety of factors -- either fixed (e.g., sex and race) or time-varying (e.g., employment status) -- to the conditional probability of exit (or of survival) to estimated -- these factors are simply included in the X vector of explanatory variables described above. This approach is very popular as a general method of analyzing spell durations and their determinants, and models of this type can be implemented in SAS as well as in other easily-obtained statistical packages (although typically the analyst is required to assume some specific underlying form for the distribution of exit probabilities). Only data sets that provide a reasonably continuous record for a reasonably large sample of individuals entering the state being examined can be used with this approach, however, which limits its usefulness with smaller or less focused data sets or those in which data has been collected in an intermittent pattern.

III. Conclusions

In summary, the many new sources of longitudinal data on incomes and family structures that have become available in the last decade offer exciting research opportunities to the policy analyst, but they bring with them their own unique measurement problems. Because these data sources are both, more complex and less familiar than are cross-sectional databases covering such topics, analyzing them can present some challenges. For analysts willing to address these challenges, however, there are useful solutions, and these data can be used to provide important new insights into the processes underlying economic and demographic change.

Indeed, as discussed briefly in the various examples of measurement problems and their solutions given throughout the paper, important applications of longitudinal analysis to policy issues have already been carried out in many areas. A few examples include Bane and Ellwood's analysis of poverty spells and of AFDC

participation using the PSID; the work by Bruce Vavricek and Ralph Smith of the Congressional Budget Office on spells of unemployment insurance recipiency as observed in the SIPP; several Social Security Administration-sponsored studies on retirement behavior as observed in the RHS; and Peter Gottshalk's work on intergenerational transmission of dependency as observed in the NLS. Projects are now underway to address a whole host of additional issues, including patterns of health insurance coverage, multiple program participation for low-income beneficiaries, and earnings and employment patterns for the working poor.

The work that has been done so far and the work that is now underway represent major advances in our understanding of these issues, but there is much further analysis that could be done with our existing longitudinal survey data. To some extent, this expansion will simply take time analysts need to become more

familiar both with the surveys themselves and with appropriate

techniques for analyzing and interpreting these data. Already, however, there is beginning to be a large literature on the applications of duration analysis, in particular, to economic and demographic data, and this literature can only be expected to grow over the next several years as additional data become available and additional issues are explored.

What can statistical agencies, and data producers in particular, do to help the analyst undertaking this type of study? In my view, these agencies could support longitudinal analysis efforts in two major ways.

First, data producers do not always produce files that are highly amenable to longitudinal analysis, even when such analysis is the primary mission of a particular data-collection effort. Understandably, when a new survey such as the SIPP comes out a great deal of effort is devoted to the early cross-sectional files, since analysts are anxious to see how these new data line up with data from famillar cross-sectional surveys. In addition, the early waves of any survey will be ready for analysis long before the

survey itself has been completed and edited longitudinally, and data producers are understandably anxious to get these first products to the users as fast as possible.

Once a survey has been in regular production for some period of time, however, it would make sense to lessen the emphasis on cross-sectional files and to increase efforts to produce reasonable longitudinal data in a reasonably timely fashion. We already have excellent cross-sectional data on family incomes and labor force status, and unless the survey in question is clearly adding to our store of available cross-sectional data on a particular topic, cross-sectional applications should receive less attention. In particular, the level of effort devoted to activities such as cross-sectional imputation that have no application in the longitudinal context should be reduced. instead, greater research efforts should be devoted to continuing problems like longitudinal editing and the development of reasonable longitudinal imputation procedures.

The second way in which statistical agencies could support

longitudinal analysis would be to undertake more of it themselves. Data producers typically publish at least some cross-sectional information from the files they produce, and in some cases -- the CPS publications in the Census P-60 series, for example, come to mind -- these tables themselves provide important information on which policy-makers come to rely. It ought to be possible for the Bureau of the Census and other data producers to publish similar information, but of a longitudinal nature, using the longitudinal databases that they now produce.

The assumptions underlying survival analyses might be difficult to explain in such a context, but basic information on the experience of a given cohort, for example, is fairly easy to explain and to interpret. For instance, one could look at how many of those becoming unemployed in a given period were still unemployed one, two, or more months later; how many of those on welfare or in poverty at a given point in time were still in that state x months (or years) later; and so forth.

Similarly, one could examine the transitions between states more directly, along with the characteristics of those experiencing the transitions. One could ask, for example, what proportion of those leaving unemployment in a given year find jobs, and what proportion leave the labor force? Does it differ for men and women, blacks and whites, old and young workers? For that matter, one could ask who becomes unemployed, and how does the incidence differ by demographic characteristics? Or, for example, what about those who enter welfare programs in a given year -- what is the incidence of entry for those in different categories? What happens to those who leave welfare in that year? Do they get married? Do they get jobs? How many of those gaining jobs are still employed six months later, or a year later? Similar questions could be asked about the incidence and impacts of many other transitions, from divorce to retirement to the birth of a child.

The longitudinal analysis issues outlined above represent only a small proportion of those that could be undertaken -- but the point here is that there is a great deal of fairly straightforward

longitudinal analysis that would be very helpful to policy-makers, and that is not now being done in any systematic way.

Some very useful reports have been issued, of course -- for example, the Census Bureau's P-70 series includes some longitudinal analysis from the SIPP, although so far such applications have been relatively limited in both quantity and scope. Again, many of these surveys, especially the SIPP and the NMCES, are still fairly new, so perhaps it is not surprising that their producers have not yet developed a complete, systematic schedule of reports examining basic longitudinal issues. Nevertheless, devoting more attention to their own longitudinal analyses would probably be the most important step data producers could take to support this type of research, and could also increase substantially the useful information that we are able to obtain from these surveys.

References

Allison, P.D. "Discrete-Time Methods for the Analysis of Event Histories," in S. Leinhardt (ed.), Sociological Methodology 1982, San Francisco: Jossey-Bass, 1982.

Bane, Mary Jo and David T. Ellwood. "The Dynamics of Dependence: the Routes to Self-Sufficiency." Report prepared for the U.S. Department of Health and Human Services. Cambridge, Mass.: Harvard University, 1983.

Bane, Mary Jo and David T. Ellwood. "Slipping Into and Out of Poverty: The Dynamics of Spells." Journal of Human Resources, Winter 1986, 21(l), pp. 1-23.

Bianchi, Suzanne, and Edith McArthur. "Family Disruption and Economic Hardship: The Short-Run Picture for Children." Paper presented at the annual meeting of the Population Association of America, May 1989.

Blank, Rebecca. "How Important is Welfare Dependence?" Working Paper No. 2026. Cambridge, Mass.: National Bureau of Economic Research, Sept. 1986.

Citro, Constance F., Donald J. Hernandez, and Roger A. Herriot. "Longitudinal Household Concepts in SIPP: Preliminary Results." SIPP Working Paper Series No. 8611. Washington D.C.: U. S. Bureau of the Census, 1986.

Cox, B. and S. Cohen. Methodological Issues for Health Care Surveys. New York: Marcel Dekker, 1985.

Duncan, Greg J. (ed.). Years of Poverty, Years of Plenty. Ann Arbor, Mich.: Institute for Social Research, 1984.

Duncan, Greg J., Richard D. Coe, and Martha S. Hill. "The Dynamics of Poverty," in G. Duncan, ed., (op. cit.) 1984, pp. 33-70.

Ernst, L., D. Hubble, and D. Judkins. "Longitudinal Family and Household Estimation in SIPP." Proceedings of the Survey Research

Methods Section. Washington D.C.: American Statistical Association, 1984.

Fox, Alan. "Work Status and Income Change, 1968-72: Retirement History Study Preview." Social-Security Bulletin, 1976.

Gottschalk, Peter. "The Intergenerational Transmission of Welfare Participation: Facts and Possible Causes." Paper presented at the annual meeting of the Association for Public Policy Analysis and Management, November 1989.

Gustman, Alan L. and Thomas L. Steinmeier. "A Structural Retirement Model." Econometrica, May 1986, pp. 555-584.

Levy, Frank. "How Big is the American Underclass?" Working Paper 0090-1. Washington, D.C.: The Urban Institute, 1977.

McMillen, David B. and Roger A. Herriot. "Toward a Longitudinal Definition of Households." SIPP Working Paper Series No. 8402. Washington DC: U.S. Bureau of the Census, 1984.

McNeil, John, Enrique Lamas and Cynthia Harpine. "Moving Into and Out of Poverty: Data from the First SIPP Panel File." Proceedings of the Social Statistics Section. Washington DC: American Statistical Association, 1988.

Office of Management and Budget, Statistical Policy Office. Federal Longitudinal Surveys. Statistical Policy Working Paper No. 13. Washington DC: OMB, May 1986.

Ruggles, Patricia. Drawing the Line: Alternative Poverty Measures and Their Implications for Public Policy. Washington DC: Urban Institute Press, 1990.

Ruggles, Patricia. "Welfare Dependency and Its Causes: Determinants of the Duration of Welfare Spells." Paper presented

at the annual meeting of the American Economic Association, Dec. 1988.

Ruggles, Patricia and Roberton Williams. "Longitudinal Measures of Poverty: Accounting for Income and Assets Over Time." Review of Income and Wealth, Sept. 1989, 35(3), pp. 225-244.

Ruggles, Patricia and Roberton Williams. "Transitions In and Out of Poverty." Paper presented at the annual meeting of the American Economic Association, Dec. 1986.

Short, Pamela Farley, Joel C. Cantor, and Alan Monheit. "Dynamics of Medicaid Enrollment." Inquiry, Winter 1984, 25(4), pp. 504-516.

Tuma, Nancy B. and Michael T. Hannan. Social Dynamics: Models and Methods. New York: Academic Press, 1984.

Vavrichek, Bruce and Ralph E. Smith. Family Incomes of Unemployment Insurance Recipients and the Implications for

Extending Benefits. Washington DC: Congressional Budget Office, 1990.

Williams, Roberton. "Poverty Rates and Program Participation in the SIPP and the CPS." Paper presented at the annual meeting of the American Statistical Association, August 1986.

Williams, Roberton and Patricia Ruggles. "Determinants of Changes in Income Status and Welfare Program Participation." Paper presented at the annual meeting of the American Statistical Association, August 1987.

DISCUSSION

Michael Brick

Westat, Inc.

Pearson

Pearson's paper is an excellent guide to federal agencies on the merits of choosing between various alternatives in designing a survey to meet specific policy relevant objectives. He argues quite persuasively that the important design question is not whether cross-sectional or longitudinal is better, but which combination of designs is most effective to answer the policy questions. Another important issue that Pearson raises is the underutilization of experimentation in longitudinal surveys. I strongly agree with him in that experiments are needed if causal modelling is a goal.

Along these same lines, longitudinal surveys offer a rich environment for experimenting with a wide variety of other issues such as memory and recall. Some items could be collected in the baseline of a longitudinal study and then ask the respondent to recall this information in a later followup. Some examples that might be interesting are income from previous years, grades while in school, even opinion and attitudes. These types of experiments might help support some of the cognitive research theories or open

the door to new and more realistic theories.

Pearson's listing of the advantages and disadvantages of longitudinal files is very useful which can and should be used to help improve design decisions. I have a few quibbles about the list that may offer a slightly different perspective. The first issue is the placement of nonresponse as a disadvantage. Although it is true that attrition is typically a bigger concern in longitudinal files, the availability of additional covariates to reduce nonresponse bias may partially offset this disadvantage. However, until we devise and implement these methods effectively in large-scale longitudinal files, the nonresponse problem will remain a disadvantage.

In many ways comparing cross-sectional and longitudinal nonresponse problems is fraught with many of the same difficulties associated with cost comparisons. If you are accomplishing something that cannot be done reliably in any other way, then you do have an "apples and oranges" comparison. When the cost of survey or the problem associated with nonresponse is discussed, the

alternatives that satisfy the same objectives must be clearly specified. Pearson is correct that general statements or conventional wisdom can lead to poor design decisions. More complete models of the errors and costs for longitudinal

alternatives to cross-sectional surveys are needed to help unravel these questions.

My main complaint with the list of advantages and disadvantages is that the discussion of response errors is too limited. If response errors create spurious estimates of change, then the major advantages (the first three of the four advantages he lists) of longitudinal files are reduced or eliminated. I'll return to this point after commenting on the paper by Kasprzyk and Jacobs.

The paper by Kasprzyk and Jacobs is a welcome insight into the many and varied issues that are peculiar to longitudinal survey design, operations and analysis. Their even-handed treatment of the differences that are encountered in large-scale longitudinal surveys obviously reflects many hours of wrestling with the real problems in this setting.

In their discussion of the advantages and disadvantages of longitudinal surveys, they mention that the net change can be estimated more precisely because of the positive correlation that can often be expected in the variables over time. While this is true, the practice in many federal longitudinal surveys has not taken advantage of this correlation properly. In some cases only cross-sectional estimates of variances are ever computed. In other cases, correlations are estimated only for a very few statistics and then a generalized correlation is proposed for all other variables.

Since the more precise estimation of net change is really

probably the greatest advantage that a longitudinal survey has, this practice needs to be re-examined. If generalized correlations are to be used, then it is important to put greater efforts into their production and distribution. For example, in a recent survey Westat conducted for the National Science Foundation estimated correlations over a two year period that ranged from -0.10 to +0.65. sampling errors for estimates of net change are not difficult to measure and should be included as a routine product in a federal longitudinal survey.

On a different issue, Kasprzyk and Jacobs note that in some longitudinal surveys efforts are made to avoid presenting data containing obvious errors. While this seems like a reasonable objective, it can actually result in poorer quality data. For example, if an error is made in the baseline period and all later data are verified against it for consistency, then new problems could be created. "Correcting" the errors in an edit program could be simply a way to suppress the problem so that users do not "see" it. It is still a real problem. The presence of earlier data may

encourage "over-editing" longitudinal survey data, creating false impressions of data quality, and increasing errors in computed statistics.

In the discussion of longitudinal weighting and imputation, Kasprzyk and Jacobs review a number of important statistical issues. As they note in their discussion there is not a universal agreement on these issues. Re-iterating a previous comment, I think that imputation should play a much larger role than weighting in estimation from longitudinal files. The information obtained in different data collection waves should be used for more efficient estimation than is possible from simple weighting adjustments. Of course, the imputation of longitudinal files is also much more complex, and methods for handling imputation in large longitudinal surveys are not very advanced. This is a challenge for producers and analysts of longitudinal files.

The Importance of Measurement Error in Longitudinal Surveys

There are four concerns about measurement errors that I think are very important to designers and analysts of longitudinal surveys. These concerns are:

- Measurement errors are the most crucial problem facing longitudinal surveys

- Measurement errors result in biased estimates of gross change and the ability to measure gross change is a prime

goal in many longitudinal surveys

- Measurement errors are a much greater problem in longitudinal surveys than in cross-sectional surveys

- Changes in survey processes are required if the potential of longitudinal surveys is to be realized

The concern over measurement errors in longitudinal surveys is not new. Errors in estimates of gross change have long been recognized as having biases which reduce their usefulness. Efforts have been made to address these problems from both the design and the analytic perspective.

A simple hypothetical example may help to understand the problem. Figure 1 shows values of a characteristic (e.g.,

participation in a program, unemployment, health coverage) for a

sample of units. The two extreme columns show the true values at times 1 and 2. Measurement error results in the values shown in the adjacent columns being actually observed. The observed values then give rise to the observed change or transition values shown in the center column.

First, notice that measurement error has not greatly distorted the cross-sectional estimates for either time 1 or time 2 (the values in error are shown in bold). Therefore the estimate of level and the estimate of the net change between times 1 and 2, which could be measured with either a longitudinal or cross-sectional survey, are not greatly affected by the measurement error.

On the other hand, the impact of measurement error on the gross change is dramatic. Ten units are observed to have changed, while the true number that changed is only 4. One of the most important and distinguishing features of many longitudinal surveys is the ability to produce estimates of gross change, but measurement error can seriously distort these estimates.

Measurement error can have a profound impact on estimates of transitions, spells, durations, and flows.

It is instructive to examine the reasons why measurement error causes so many more problems in longitudinal than cross-sectional surveys. The truth-by-survey table for a cross-sectional survey is a useful way of working with measurement error for qualitative variables. (See Table l.a and l.b)

The net bias is the difference of two margins from the table (a+b) - (a+c), or simply b-c. The goal is to have zero or at least

a small net bias. The conditions for zero net bias (i.e., when b=c) are given in the Appendix of the Bureau of Census (1985). Using their notation, let

Pr(observed value = No / true value = Yes) = q

and

Pr(observed value = Yes / true value = No) = f.

Then the net bias equals zero if Pq = (1-P)f, where P is the true proportion of the population with the characteristic.

If the two error rates are approximately equal and P>0 and q>0, then the net bias will get smaller as P approaches .50. If P=.02, then the ratio of q:f must be 49:1 for the net bias to equal zero. This merely points out the inter-relationship between the net bias and the size of the estimate. For estimates of rare

characteristics, measurement error is likely to be more problematic. Of course, the distribution of the two error rates is also of great importance.

If we extend the example to a second observation time, we encounter the same problem but the impact is larger. First, note that the net bias for the two observation periods is equal if the probabilities of error are the same between times 1 and 2 and the proportion with the characteristic does not change.

The point of this simple exercise is to show that in a longitudinal setting the net bias for gross change involves the sum of four differences, while for estimates of level there is only one difference. The problem is naturally greater in trying to measure gross change, which is often one of the main objectives of a longitudinal survey.

As I noted earlier, the problems of response errors in longitudinal surveys have been addressed from both a design and analysis perspective. The work of Bye and Schechter (1986), Chua and Fuller (1987), and Poterba and Summers (1984) are some excellent examples of the analytic approach. Marquis and Moore (1989) offer additional insight using data from records, and highlight the need for designing the surveys and instruments better.

I suspect that if Dr. Deming were to become involved in this issue he might say that longitudinal surveys offer new challenges and we must change the way we do business. In longitudinal surveys we can no longer accept the errors and expect others to buy our

products. We must concentrate on the survey process, identify the major sources of variability, and take steps to eliminate them from the system. If we fail to take these types of actions, then it is

likely that it will be harder and harder to support longitudinal surveys in the future.

References

Bye. B.V. and Schechter, E.S. (1986), "A Latent Markov Model Approach to the Estimation of Response Errors in Multiwave Panel Data," Journal of the American Statistical Association, 81, 375-380.

Chua, T.C. and Fuller, W.A. (1987), "A Model for Multinomial Response Error Applied to Labor Flows," Journal of the American Statistical Associatign, 82, 46-51.

Marquis, K.H. and Moore, J.C. (1989), "Some Response Errors in the SIPP- With Thoughts About Their Effects and Remedies," Proceedings of the Section on Survey Research Methods of the American Statistical Association.

Poterba, J.M. and Summers, L.H. (1984), "Adjusting the Gross Changes Data: Implications for Labor Market Dynamics," Proceedings of the Conference on Gross Flows in Labor Force Statistics.

U.S. Bureau of the Census (1985), "Evaluating Censuses of Population and Housing," Statistical Training Document, ISP-TR-5.

DISCUSSION

Marilyn E. Manser

U. S. Bureau of Labor Statistics

The papers by Patricia Ruggles and Robert Pearson, on which I was invited to comment, both provide helpful insights into the usefulness of longitudinal data. I find myself in agreement with them, for the most part. What I primarily will do is reinforce points which I think are particularly important and discuss other points on which my perspective may be a little different.

Let me begin with a fundamental question suggested by this session's papers: what is the definition of a longitudinal survey

vs. a cross-section survey? Although I do not have a clear answer to this I want to raise it for thought. Pearson defines a longitudinal survey as one "in which repeated observations are made of the same individual subjects." In his paper with Robert Boruch (1988), the Current Population Survey (CPS) is included in the description of longitudinal surveys. In contrast, the OMB Statistical Policy Working Paper 13, "Federal Longitudinal Surveys," excluded rotating panel surveys such as the CPS, the Consumer Expenditure Survey, and the National Crime Survey because there was no explicit plan for longitudinal analysis incorporated. Ruggles never explicitly defines what she means by a longitudinal survey, but is clearly using CPS as an example of a cross-sectional survey. Alternative definitions have of course been considered elsewhere. One possible design-based definition could include a requirement that in order to be called longitudinal a survey must follow movers -- on this basis, CPS would not be called longitudinal even if a specific plan were developed to make use of its longitudinal aspects. (CPS permits, for example, a variety of longitudinal studies of labor market situations, although at present there are other problems with the quality of longitudinal estimates based on CPS besides the fact that movers are not

followed.)

It is important to note also that a purely design-based criterion would be less than fully satisfactory -- problems could prevent the design from being implemented. For instance, budget cuts could prevent any follow-up after the first round. Less drastically, it is important to ask, if following movers is viewed as important, what proportion of movers are actually found? It would be useful to have this information produced regularly on longitudinal surveys.

To my knowledge, few major ongoing program efforts depend on truly one-shot surveys, which seem to be what Pearson is calling a cross-sectional survey. Most statistical surveys are used to produce at least aggregate estimates of change, even if that was not the primary purpose in mind when they were designed. For

example, for many analyses of the economic situation one is really interested in whether the unemployment rate is high or low compared to other periods. Uses of data to construct aggregate measures of change and arguments for improving aggregate cross-sectional estimates can both justify a statistical design including a rotating panel, even when no explicit longitudinal analyses are planned. But in addition there may be cost implications to one-shot surveys, making them less cost effective than what I will call "mixed surveys", rotating panel surveys which fail stringent definitions of a longitudinal survey but are not truly one- shot. In any case, for a relatively small additional effort to improve longitudinal aspects of mixed surveys such as CPS, it may be possible to improve analytic possibilities enormously.

Both the Pearson and the Ruggles papers focus on household surveys. But establishment surveys such as the Census's Annual Survey of Manufacturers and BLS's 790 survey typically go back to the same sample units repeatedly. Such surveys offer tremendous opportunities for increasing understanding of economic phenomena if the problems in making use of their panel aspects can be overcome.

I. Advantages and Disadvantages of Longitudinal Surveys

A major focus of Pearson's paper is on weighing the advantages and disadvantages of longitudinal surveys. This is a valid and useful discussion, but note that the disadvantages all center on measurement problems. No one has successfully argued, to my knowledge, that non- experimental cross-section surveys are more useful for the analytic purposes for which non-experimental longitudinal surveys are designed. Further, given that it is too burdensome to obtain the needed information with retrospective questions, which in any case would entail severe recall problems, there is really no alternative to longitudinal surveys for many types of analyses. Longitudinal surveys are extremely important and should be given greater use than they have received in the past but much more research is needed on measurement problems.

One section of Pearson's paper argues for coupling longitudinal and experimental designs. Surveys conducted to collect and analyze data on social experiments have typically been

longitudinal. The point made here is a recommendation to conduct policy-related experiments with individual respondents to a general-purpose longitudinal survey. Clearly this could be a cost-effective way to collect data if the individuals on whom the experiment were being conducted were no longer to be counted as part of the original survey. But otherwise I would be extremely

troubled by doing this. The whole purpose of policy-related experiments is often to influence outcomes, and even if that is not the intent outcomes are still likely to be influenced. If this occurs then responses of the sample members to a wide range of survey questions are no longer representative of the population as a whole. In contrast, conducting survey methodological experiments which can be assumed not to have a measurable impact on outcomes can be useful or necessary in some instances.

Pearson also makes the related point that joining additional questions to an ongoing survey can be valuable. I am in wholehearted agreement with this. One thing that the Department of

Labor's (DOL's) National Longitudinal Survey (NLS) program has done that has, in my view, been very beneficial to a variety of government agencies as well as to the outside research community is to accept funding from other agencies to collect information of interest to them which also enhances usefulness of the data for analyzing labor market behavior. For example, the National Institute for Child Health and Human Development has added blocks of questions on child care use to the NLS Youth survey. Because of the importance to us and to others of joining data collection needs from other agencies to our survey, we have developed a general policy to preserve the integrity of the basic data.

II. Longitudinal Analysis

As implied by her title, Patricia Ruggles' paper focuses on analysis of longitudinal data, primarily econometric analysis using the micro data. It has been in the micro area that the major use of longitudinal survey data has occurred. For instance, as documented by Frank Stafford (1986), much of what we know about

labor economics has come from longitudinal surveys, primarily NLS and the Panel Study of Income Dynamics. Research using these data continues, on topics such as the impact of private sector training on future earnings, low wage jobs and their impacts, and the labor supply behavior of women during pregnancy and shortly after birth of the child. Similarly, as more experience with the Survey of Income and Program Participation (SIPP) accumulates, we are likely to see many useful micro studies that will impact the way the research community thinks about issues such as spells of dependence on various programs.

But where longitudinal household surveys have not made a large contribution yet, it seems to me, is in short-term analysis of current data: for instance, a series of reports on a topic such as how have the transition rates out of poverty changed since last year. However, I understand that the Census Bureau has recently released two P-70 reports using SIPP to analyze transitions. I believe that development of a series of current analytic reports in addition to long-term econometric research studies from NLS is an extremely important goal for this program, but the program has

never in the past included this dimension. In general, I think that carrying out both long-term econometric studies and shorter-term, more current, tabular analyses is important and that they are complementary. However, as all three papers in this session note, further research on weighting problems is greatly needed.

Much of the Ruggles paper considers the complexity of use of longitudinal data sets, for which they are often criticized. She recognizes, and this is important, that these complexities necessarily come along with the richness of the data sets that are responsible for the "exciting research opportunities" that they provide.

While it is true that existing longitudinal data sets are hard to use, this is the case because of the vast amount of information they contain, particularly after several rounds of interviews have taken place. (For instance, she notes that SIPP provides a choice

of accounting periods for income measures and this choice can make a difference to the analysis. This is in contrast to CPS where income is available for only one accounting period. With SIPP, richness of choice creates complexity. Unless there is widespread agreement about what measure to use, summary income measures provided on its files by a statistical agency would presumably not suit some of its users.)

Another related point in Ruggles' paper is the recommendation that agencies append a myriad of transition flags to person records. While this is feasible, again, in general, unless there is a large set of users with a particular need or an ongoing agency use of the data for a particular purpose it will probably be the case that a particular user will not find all aspects of a publicly available file ideally suited to his or her particular use. Limited resources can often be used, however, to respond to needs affecting a number of users. For instance, many users of NLS data use the event history information on jobs -- very rich data that were initially very difficult to use. As a result, a Workhistory data tape was developed which contains weekly arrays of labor force status, usual hours worked per week, and dual job information.

This Workhistory data tape makes analyses using this information considerably easier.

As she notes in her introduction, Ruggles' focus is "almost exclusively on the application of longitudinal analysis to questions concerning patterns of family income, expenditures, and/or demographic change." Because of this focus, she devotes considerable attention to efforts to construct a longitudinal family definition. This is a major problem for analyses of topics such as transition on and off of means-tested government programs. This problem, too, exists because of the richness of this type of longitudinal data. It is well-known that examining income-levels by family type using CPS is plagued by the fact that the family structure information relates to a different period than the income

measure. Because a longitudinal data source entails problems for analysis does not necessarily mean that analysis of a similar topic would be preferable using other, easier-to-use data.

Note also that this problem of family status definition is not a central problem in analysis of many types of longitudinal issues -- it is the focus here that makes it one. For example, studies of labor supply behavior, work experience, earnings growth, and so on focus on the individual. NLS follows individuals primarily because of its focus on labor force related information for people in groups of particular interest to DOL. Similarly, the Department of Education focuses on the individual in its longitudinal studies which focus primarily on educational experiences and outcomes for youth.

Ruggles' point that just reweighting the way it is typically done does not necessarily solve a problem due to nonrandom attrition is an important one. This suggests one reason, among others, for why a microeconometric analysis may be preferable to looking only at tables: as she points out, it is possible to include people who are in the sample only for some of the periods in a micro study. But in general, use of tabular and econometric analyses can be complementary.

In her section on analyzing data on duration, Ruggles provides a useful discussion of some of the pitfalls in this type of study. Restricting an analysis to a particular age/sex cohort is not problematic if the interest is in a particular group. It is using a variable that represents a choice to select a sample for analysis that causes all the types of problems she discusses in this section.

In her conclusion, she makes two recommendations for statistical agencies. One is to lessen the emphasis on cross-sectional files. Her point that it is important not to use data with cross-sectional amputations for a longitudinal analysis is an important one. But surveys may have multiple purposes so that avoiding cross-sectional amputations entirely, especially given needs for issuing timely data, would not be a possibility in many cases. The second recommendation is for more longitudinal analysis by statistical agencies. I agree that this is important, even though longitudinal studies and tables may be difficult to explain in many cases as she notes.

In conclusion, let me note that as part of the major ongoing joint BLS/Census CPS Redesign effort, attention to longitudinal issues is planned. Tables of gross monthly flows between labor force states are presently produced regularly but are not officially published because the estimates are not of sufficient quality. Efforts are planned to improve longitudinal aspects of the survey and to research adjustment techniques to improve the gross flows tables. In addition, if funding permits, plans are to

conduct a separate CPS-like longitudinal survey which would follow movers and keep people in the sample longer. This survey, in addition to supporting improved analysis of short-run changes in labor force behavior, would permit research on a multiplicity of survey-related topics.

References

Boruch, R. F. and R. W. Pearson, "Assessing the Quality of Longitudinal Surveys," Evaluation Review, Vol. 12, February 1988, pp. 3-58.

Stafford, F., "Forestalling the Demise of Empirical Economics: The Role of Microdata in Labor Economics Research," in 0. C. Ashenfelter and R. Layard, eds., Handbook of Labor Economics, Vol. 1. Amsterdam: North-Holland, 1986, pp. 387-423.

TOWARDS AN AGENDA FOR THE FUTURE

Stephen E. Fienberg

Carnegie Mellon University

My remarks this afternoon will focus on a few key themes that emerged in various sessions over the past two days. I will attempt to use these themes to point towards elements of an agenda for the future of the federal statistical system, not just the future of

the Federal Committee on Statistical Methodology that oversees the OMB Statistical Policy Working Paper Series around which the seminar has been centered.

On Quality

George Hanuschak in the session on survey quality profiles recalled the words of one of the present-day quality gurus, to the effect that we should build quality into the system, not just inspect for the lack of it after the fact. A variant on this is the theme that we need to build quality and evaluation into our data collection processes. The traditional notion of coming back several months later to check on the answers provided by a survey respondent seems at odds with the notion of ongoing change and improvement. For example, consider two components of the 1990 Census, the group quarters censuses of college and university campuses and the special homeless component - - S-night - - program. In neither case can one expect to return a month or so after the enumeration to check on information recorded. Thus a careful census quality program would have some built-in evaluation

mechanism for these components.

At Carnegie Mellon University we have a new Statistical Center for Quality Improvement which we operate jointly with the statisticians at the University of Pittsburgh, and my colleagues associated with this center are fond of referring to the three generations of statistical approaches to quality. The first of these is the basic univariate control chart generation of technology associated with the names of Shewart, Deming, and others and based on ideas that were found in the literature in the 1920s and 1930s. The second generation was linked to the introduction of careful experimentation specifically designed for the industrial setting, e.g., response surface methodology and EVOP, and introduced in the 1950s and 1960s. The recent interest in Taguchi methods is rooted in large part in basic fractional factorial design ideas. we are just beginning to see the emergence of the third generation of quality techniques which focus on statistical methods for the analysis of complex multivariate data using high speed computation and computer graphics.

Based on what I know about quality efforts in the federal statistical agencies, and what I heard described at this seminar, I would describe the current state-of-the-art as being focussed on the first generation of quality ideas, univariate in approach, lacking careful and systematic experimentation, and devoid of techniques rooted in the modern world of computing. Yet there are ample opportunities for moving quickly into the second generation by utilizing ideas on the embedding of experiments in surveys (e.g., see Tanur and Fienberg, 1988, 1989). The simplest of the embedded designs (the split ballot experiment) is often recommended for use (as it was in the session here on questionnaire design) but rarely analyzed properly. Indeed, as we look to the widespread exploration of ideas and concepts coming out of the cognitive laboratories, the federal agencies must take seriously the second generation ideas of embedded experiments.

On the Need for Integration

Some of the recent advances in methods for data collection and analysis appear as add-ons, off to the side of the main enterprise. In the spirit of the Total Quality Management movement we have heard in several sessions about the need for Integration of the components of survey design, and in a larger sense for the integration of thinking across agencies.

I am reminded of an academic story. As a dean at Carnegie Mellon, I sit on the university promotion and tenure committee and get the opportunity to review cases from diverse disciplines. A few years ago, we were reviewing the case of a physicist for tenure and his file contained a number of letters describing his experimental work as brilliant, innovative, or outstanding. As we looked over his curriculum vitae, we noted that he had no individually-authored papers but only appeared as one of a cast of thousands on each paper. Finally, one committee member asked the presenter of this case, what was so distinctive about the candidate's work in high energy physics that merited the laudatory comments. The response was: "He focuses the beam."

Now many of you have roles in federal data collection that are akin to that of the physicist's beam-focussing. These are important and often crucial roles, but their value needs to be understood in the broader integrative setting, both by you in your work and by those who are looking towards quality improvement more broadly.

On the Statistical Policy Working Paper Series

While many of the sessions at this seminar were based directly on papers from the OMB Statistical Policy Working Paper Series, others have been on collateral advances in methods and data quality

assessment. Bob Groves began the seminar by noting three perspectives on the goals that the series should have. These were to serve as:

(a) reports in the "state-of-the-art" of federal practice,

(b) vehicles for agency cross-fertilization,

(c) prods to new developments.

Many of the nineteen papers issued to date have succeeded admirably in categories (a) and (b), and they have changed how work is done across agencies. Others have had only limited impact. But I think that we could agree with Groves that few of the papers were prods to major new methodological developments. Perhaps the Federal Committee on Statistical Methodology that oversees the OMB Statistical Policy Working Paper Series needs to be more daring in its choice of topics in the future. New topics need not be rooted in ongoing work in specific agencies nor do they need to be ones on which the committee agrees. For how else can we achieve a major shift or revolution in methods and quality?

At the same time I should note the need for attention to and support for the committee's activities on the part of senior administrators in the statistical agencies. If staff do this work

only in their spare time we can expect to see few major methodological advances.

Shifts of Paradigm for Federal Statistics

Fritz Scheuren has been talking both at this seminar and in recent years about the need for a paradigm shift in how we do federal statistics. I believe that he is correct in this claim although I do not think that many people understand what he and the philosophers of science mean by paradigm shifts. I commend those of you who have not read Thomas Kuhn on scientific revolutions to do so as his ideas often get mangled in the translation.

Kuhn talks about the day-to-day orderly change and incremental knowledge approach to science which gets radically altered and reorganized by the introduction of a new set of ideas and a new paradigm such as that associated with the work of a Newton or an Einstein. Now when a paradigm shift occurs, the past tools and perspectives are not all discarded. Rather they are looked at in

a different way and accorded a different place in the hierarchy of importance. What we also see is the introduction of dramatically different measurement methods, with markedly changed error profiles.

Up through the present day the federal statistical system has been based in large part on tools developed many years ago, more often than not in the 1930s and 1940s. This is especially true in survey design and census taking. With the technological revolution

of the 1970s and 1980s, one might have expected to see a paradigm shift in statistics in the agencies, but the computer and its effects have been forced into the old paradigm instead of being the trigger to a reorganization of our thinking.

The last decade has been a difficult one for statistical agencies, but perhaps the problems that the agencies have

encountered during this period should spur us to rethink what we do and how. We should be asking if tools like CAPI, CATI, distributed computing networks, major new analytical statistical methods, and the cognitive-statistical laboratory may be the vehicles to major changes.

Impediments to Major Change

Perhaps the biggest impediment to change is the bureaucracy in which most of you work. A piece of this is the attitude: "We've always done it that way." This is related to the theme I would label as "The Agency is the Data." The purpose of collecting statistical data is not an end unto itself, but rather a means to a social or policy goal. The aim of the federal statistical agency then should be to serve these broader goals well, rather than to collect data insulated from outside input and protected from outside scrutiny. We need to move towards making our data relevant; to measure what is of importance, albeit poorly, instead of measuring what current methods are designed to be good at, even

if it is of marginal interest.

I'd like to tell a parable about of the National Goodness Survey which was mandated late one night in conference by Congress as an amendment to a foreign aid bill. The federal methodology coordinating committee was asked to propose a design for this new survey at one of its meetings, and each of its members was asked to come back to the next meeting with a proposal for the design:

(a) The representative from the Bureau of the Census returned with a household survey design that resembled the Current Population Survey, and she noted that surely goodness resided in household locations, just as unemployment does.

(b) The representative from the Energy Information Administration noted that goodness was likely to flow from reservoirs in the group and thus proposed a design modelled on their survey of natural gas reserves.

(c) The Bureau of Labor Statistics representative suggested

that we couldn't ignore the component of goodness that was due to business establishments, and proposed a separate survey based on their new establishment list.

But she also offered the auspices of the BLS cognitive laboratory for testing ideas on goodness consumption.

(d) The Bureau of Justice Statistics representative noted that his agency didn't actually conduct its own surveys and referred the committee to the representative of the Census Bureau for how this should be done.

(e) The representative from the National Center for Health Statistics suggested that goodness was a manifestation of physical well-being and urged that the new survey be a supplement to the National Health Interview survey.

(f) Finally, National Center for Education Statistics

proposed that we ask the state superintendents for public schools to report on the fostering of goodness in the educational process, and that we develop a new standardized test that could be administered annually to measure the acquisition of goodness skills.

(I leave it as an exercise for the reader to describe how the representatives from BEA, DoD, IRS, and NASS responded.)

Now part of the problem with my parable lies with the approach taken by each of the agency statisticians who, instead of asking what the concept of goodness is all about and how could one measure it, looked to analogues close at hand and let the standard methods he or she was familiar with frame all of the answers to the crucial unasked questions. Perhaps a survey is the wrong tool for the task of measuring goodness. The other problem arises from the fact that no agency has a monopoly on statistical methods or the ability to design new surveys, not the Bureau of the Census, not BLS, not even the small band of statisticians in OMB who must approve the design. New projects the federal statistical system is likely to face in

the next decade likely will require innovative thinking and true interagency collaboration.

The example given by Judy Lessler of measuring the quality "of Flowing Waters" is illustrative of the point I am trying to make. The Research Triangle Institute (RTI) statisticians put this problem of measuring the quality of the nation's flowing waters back into the traditional survey domain of a frame with a population of units (river reaches) to be sampled. The approach was ingenious and some might even call it innovative. But in so describing the problem of measuring the quality of the nation's waters Lessler missed the opportunity to note a point that I am sure the RTI statisticians discussed, namely that many radically different frameworks are possible for looking at this issue, and only some of these fit neatly into a traditional sampling approach.

Thus one of the messages I bring you today is that we all must learn to question the appropriateness of traditional statistical frameworks and institutional dogma. This is especially true as we move into some of the more fascinating new domains of federal statistics, e.g., related to the environment, as well as in considering different ways of collecting data, for censuses, and especially for longitudinal surveys.

What we do know about longitudinal surveys is that they should not be a simple pasting together of waves of cross-sectional surveys. What we do not know is how to design such surveys except by faulty analogy to traditional cross-sectional methods. Traditional concepts of frames and survey coverage suddenly become elusive, shifting over time. For longitudinal surveys we need to rethink what data we collect, when, and how. And we need to have a more flexible set of analytical tools that allow the data to be viewed from multiple perspectives. Technology may well help here with problems associated with sample attrition and the followup of movers.

Some Advice

I'd like to end with a bit of advice and encouragement about what you can do to improve the quality and appropriateness of the statistical work in your own settings. Your challenge is to keep yourself from being isolated, to prevent yourself from accepting as infallible the data collection and analysis methods you currently use in your job, and to look beyond the walls of your organization.

(a) Ask "why" more often that you have in the past.

(b) Dare to have new ideas or suggest the exploration of someone else's new ideas. Innovative ideas have a long gestation period and only a small fraction of them actually work in practice.

(c) Insist on careful evaluation and documentation of what you are doing.

(d) Don't be afraid to say that you don't know or you don't understand. Such statements are often not a sign of

ignorance but rather indicators of wisdom.

(e) Hang in there. Your jobs are difficult and most of you are doing them well. The nation depends on your efforts.

References

Fienberg, S. E. and Tanur, J. M. (1988). From the inside out and the outside in: combining experimental and sampling structures. Canadian Journal of Statistics, 16, 135-151.

Fienberg, S. E. and Tanur, J. M. (1989). Combining cognitive and statistical approaches to survey design. Science, 243, 1017-1022.

Kuhn, T. (1970) The Structure of Scientific Revolutions. (Second

Edition, Enlarged) University of Chicago Press, Chicago.

TOWARDS AK AGENDA FOR THE FUTURE

Margaret E. Martin

I have been given a "Where do we go from here?" assignment to help in focussing the experience of the Federal Committee on Statistical Methodology (FCSM) on future directions.

So who is "we" and where is "here"? I have chosen to consider "we" as something broader than the FCSM itself -- perhaps the coordinating role of the Statistical Policy office, perhaps that amorphous entity, the federal statistical system in general.

Where is "here"? It seems to me "here" is an amazingly distant and productive way from the starting point when the FCSM was founded--19 "state of the art" reports ago. The productivity of the FCSM's part-time, interagency subcommittees has been outstanding. Much credit belongs to Maria Gonzalez.

Some notion of expectations is essential in order to assess past progress and future progress. What can such committees accomplish? We do not usually look to such groups to produce major breakthroughs in statistical theory, nor to engage in detailed technological applications or experiments. Rather, it seems to me, an interagency committee might be expected to perform one or more of the following functions:

1) exchange knowledge, techniques or experience among committee members to enhance the quality of the member

agencies' own operations;

2) provide "state of the art" reports to encourage best practice among a broader group;

3) recommend areas for improvement and needed directions for research; and

4) obtain consensus on such issues as -- defining problems and the priorities among them, developing or changing classifications or other concepts, and setting statistical standards.

I am uncertain how much the various subcommittees have served to fulfill the first objective -- that of exchanging knowledge among the subcommittee members -- especially upon hearing informal comments that much subcommittee work is report drafting and criticizing undertaken by individuals on evenings and weekends, rather than exchanges at committee meetings. In such circumstances, the interplay among participants that sometimes

leads to unexpected and happy outcomes is not encouraged. Perhaps this is a point that needs more consideration and possible

development in the future. Suggestions that arose in the opening session yesterday for more followup on subcommittee reports might lead to more continuing and profitable interactions among subcommittee members.

The FCSM has fulfilled admirably the second objective I listed, that of providing state of the art reports to encourage best practice among a broader group of agencies both within and outside the Federal Government. Robert Groves reported yesterday, for example, that he has used some of the reports in training future survey statisticians. The record of the FCSM in meeting this objective is outstanding.

Many of the reports meet the third objective of recommending areas for improvement and needed directions for further research --

although the record here is more spotty.

The fourth objective, that of obtaining consensus on broad definitional, conceptual and classification issues, has not been well met. Indeed, the FCSM apparently does not deem such issues to be within its purview. Very well, but such issues are of immediate concern to OMB's Statistical Policy Office and at this time pressures are increasing to re-examine basic concepts and classifications in both economic and social areas and to establish more extensive statistical standards for the federal data collecting agencies.

For example, many of our economic statistics depend for classification purposes on the concept of the establishment; many of our social statistics are collected about and from families -- yet both of these concepts are becoming more difficult to apply and possibly less relevant. Changes in either would have major impacts on the uses of the resulting data; they might also have major methodological repercussions.

Although the FCSM has developed an admirable program of sponsoring, reviewing and publishing reports on specific topics, it has not been so forthcoming about its own operations. I am especially curious about how it selects the areas for subcommittee operation. The most important problems? By what criteria? The problems most likely of immediate solution? Or some interface between these criteria? Here I would only note that the Committee has not yet tackled the most difficult problem of all those facing federal statistical agencies, that of setting statistical priorities. It is possible that chances of a successful subcommittee outcome are too remote to warrant effort on this issue. It is now fifteen years since a panel of the Committee on National Statistics issued a preliminary report* and recommended additional research on costs, and, especially, the benefits of statistical activities. To my knowledge, there has been little if any follow-through by federal agencies. The time may be ripe for another look at this issue.

Fritz Scheuren spoke yesterday about possible paradigm shifts

in the taking of the Census. I have my own pet paradigm shift to recommend. Back in the 1940's and 1950's when I was being educated in statistical methodology by Morris Hansen and others in the federal statistical agencies as part of the process of their obtaining Bureau of the Budget approval for forms, I learned the paradigm that the sponsor of the form (the subject matter specialist, the scientist, the policymaker) specified the subjects to be covered and the accuracy desired, the statistician provided the statistical design and methodology and estimated the costs of alternatives. The description is over-simplified but not unfair, I think. Yet how far from actual practice. The economic or social theorists seldom specify an operationally feasible concept. It is the applied survey economists, demographers and other specialists, together with the statisticians, interacting, who develop concepts, designs and methodologies in a succession of approximations.

As a case in point, take the definition of employment. In classical economics, employment is not defined, nor even mentioned. An undifferentiated mass known as "labor" was identified as one of the three factors of production, and when current information on

the demand for labor was wanted in the late nineteenth century, it was determined to collect data on employment from employers. The result was that employment was defined to be something obtainable from employer records, a concept approximating filled jobs. It excluded the self-employed, but one person could be counted on several payrolls by holding a number of part-time or part-period jobs. The source and method of data collection thus determined the effective definition. Later, when a new survey attempted to measure the unemployed, it proved necessary to go to persons in households for the information and to identify the employed to differentiate them from those not at work and seeking work. A quite different count of employment resulted, again reflecting the basic survey methodology. Some of the differences between the two series in level and change can be explained by known differences in the concepts, some remain intractable.

I think it is high time to shift the paradigm more towards one centered on survey methodology broadly defined. This would argue for either expanding the scope of the FCSM or establishing additional coordinating committees under OMB auspices to work on

developing consensus on critical conceptual and classification issues.

Reference

Setting Statistical Priorities, Report of the Panel on Methodology for Statistical Priorities, Richard Savage, Chair, Committee on National Statistics, National Academy of Sciences, Washington, DC, 1976.

TOWARDS AN AGENDA FOR THE FUTURE

Hermann Habermann

Office of Management and Budget

The keynote address at this seminar given by Bob Groves

presented the goals for the working papers prepared by the Federal Committee on Statistical Methodology. The goals are documentation of Federal statistical practices, cross-fertilization among agencies, and to prod new developments. The address suggested the need to establish a reward structure for the Federal staff that work on the FCSM projects.

The Federal statistical system needs to periodically examine itself to determine if we are meeting our goals. Some of the areas of work that the statistical system must now investigate are listed below.

o Cognitive laboratories

o The National Academy of Sciences' Committee on National Statistics is studying:

  Trade data

  Disclosure-avoidance techniques

o The Bureau of Economic Analysis is moving away from the

present system of National Accounts towards the United Nations National Accounts System.

- Census 2000 is 10 years away so we need a fresh look at the methods used to collect census data.

- Private data bases are now burgeoning. The Federal government no longer has a monopoly on data bases.

There are many changes in public attitudes. Some of the questions that we need to consider in the context of the Seminar on Quality of Federal Data follow.

1. What is the purpose of the decennial census?

2. What is the relationship between the "10 year ceremony" and intercensal data collection?

3. Where are we going on disclosure-avoidance techniques?

4. Where are we going with Federal-State statistical program? How can we evaluate the multiple models used in these programs?

5. What is the best strategy to take care of the increasing difficulties that agencies have with recruitment and training of technical personnel?

Reports Available in the

Statistical Policy

Working Paper Series

1. Report on Statistics for Allocation of funds (Available through NTIS Document Sales, PB86-211521/AS)
2. Report on Statistical Disclosure and Disclosure-Avoidance Techniques (NTIS Document Sales, PB86-211539/AS)
3. An Error Profile: Employment as Measured by the Current Population Survey (NTIS Document Sales PB86-214269/AS)
4. Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics (NTIS Document Sales, PB86-211547/AS)
5. Report on Exact and Statistical Matching Techniques (NTIS Document Sales, PB86-215829/AS)
6. Report on Statistical Uses of Administrative Records (NTIS Document Sales, PB86-214285/AS)
7. An Interagency Review of time-Series Revision Policies (NTIS Document Sales, PB86-232451/AS)
8. Statistical Interagency Agreements (NTIS Document Sales, PB86-230570/AS)
9. Contracting for Surveys (NTIS Document Sales, PB83-233148)
10. Approaches to Developing Questionnaires (NTIS Document Sales, PB84-105055/AS)

11. A Review of Industry Coding Systems (NTIS Document Sales, PB84-135276)

12. The Role of Telephone Data Collection in Federal Statistics (NTIS Document Sales, PB85-105971)

13. Federal Longitudinal Surveys (NTIS Document Sales, PB86-139730)

14. Workshop on Statistical Uses of Microcomputers in Federal Agencies (NTIS Document Sales, PB87-166393)

15. Quality in Establishment Surveys (NTIS Document Sales, PB88-232921)

16. A Comparative Study of Reporting Units in Selected Employer Data Systems (NTIS Document Sales, PB-90-205238)

17. Survey Coverage (NTIS Document Sales, PB90-205246)

18. Data Editing in Federal Statistical Agencies (NTIS Document Sales, PB90-205253)

19. Computer Assisted Survey Information Collection (NTIS Document Sales, PB90-205261)

20. Seminar on the Quality of Federal Data (NTIS Document Sales, PB91-142414)

Copies of these working papers may be ordered from NTIS Document

Sales, 5285 Port Royal Road, Springfield, VA 22161 (703) 487-4650

"1" A copy of the complete paper which details the sample selection and matching procedures used in ERUMS is available from John Pinkos, Bureau of Labor Statistics, GAO Building Room 2913, 441 G Street, NW, Washington, DC 20212, Telephone (202)523-1636.

"2" Acknowledgement: This chapter was partially supported by a grant from the National Science Foundation's program of Measurement Methods and Data Improvement, Grant # SES-8511609. The chapter benefitted from the helpful comments and suggestions of Robert F. Boruch, Calvin C. Jones, and Nancy A. Mathiowetz. A more extended version of this chapter appears in Krishnan Namboodiri and Ronald G. Corwin, editors, Research in the Sociology of Education and Socialization. Volume S. Greenwich, Connecticut: JAI Press, 1989, pages 177-199, and is reprinted here in part with the permission of the publisher.

"3" This survey of approximately 2,000 respondents was a face-to-face stratifed probability sample of the adult noninstitutionalized population of the United States, which included a special supplemental sample of minorities for that year.

"4" This survey of approximately 10,000 respondents is a cohort of youth (age 14-21 during the first year of the survey in 1979) which included oversamples of females and minority youth and a special military sample.

"5" A longer version of this paper that discusses several other topics in longitudinal analysis such as designing a longitudinal file, dealing with attrition, imputation and weighting issues, and the choice of an accounting period is available from the Census Bureau as SIPP Working Paper No. 9007.

"6" See Duncan (ed.) (1984) and McNeil et al. (1988).

"7" Applications illustrating the use of this technique to analyze

income change can be found in Ruggles and Williams (1986) and Williams and Ruggles (1987).

"8" See Bianchi and McArthur (1989).

"9" See for example Guatman and Steiruneier (1986).

"10" Additionally, if rates of divorce are changing rapidly over time, the use of pooled data on transitions from A long-term sample such as the PSID may give misleading estimates of, transition probabilities. See for example Tuma and Hannan (1984) for more discussion of this point.

fnote>"12" Mary Jo Bane and David Ellwood's classic paper on poverty spells makes this point very well, and provides a good example of spell analysis as applied to the PSID. (See Bane and Ellwood (1986)). For a similar example using SIPP data, see Ruggles and Williams (1989). Other useful applications include the work by Pamela Parley Short and her colleagues on spells of Medicaid participation and Rebecca Slank's imaginative use of longitudinal data from the

Seattle and Denver Income Maintenance Experiments to examine spells of welfare program participation. See Short et al. (1988) and Blank (1986).

"13" See Duncin et al. (1984).

"14" This discussion is aimed at the analyst trying to decide whether this approach is appropriate for the particular application he or she has in mind. Anyone attempting to implement such an analysis should bf course review some of the more technical literature on this topic. Tuma and Hannan (1984) provide a good basic an overview of these methods. In addition, the treatment in Allison (1982) may be helpful to analysts who are completely unfamiliar with event history analysis techniques.

"15" The author is Assistant Commissioner, office of Economic Research, Bureau of Labor Statistics. The views expressed herein are those of the author and do not necessarily reflect those of the Bureau of Labor Statistics.