



Statistical Policy
Working Paper 23

**Seminar on New Directions in
Statistical Methodology**

Part 1 of 3

Federal Committee on Statistical Methodology

Statistical Policy Office
Office of Information and Regulatory Affairs
Office of Management and Budget

June 1995

MEMBERS OF THE FEDERAL COMMITTEE ON
STATISTICAL METHODOLOGY

(June 1995)

Maria E. Gonzalez, Chair
Office of Management and Budget

M. Denice McCormick Myers, Secretary
National Agricultural Statistics Service

Susan W. Ahmed
National Center for
Education Statistics

Yvonne M. Bishop
Energy Information
Administration

Cynthia Z.F. Clark
National Agricultural
Statistics Service

Steven Cohen
Administration for Health
Policy Research

Lawrence H. Cox
Environmental Protection
Agency

Zahava D. Doering
Smithsonian Institution

Daniel Kasprzyk
National Center for
Education Statistics

Nancy Kirkendall
Energy Information
Administration

Daniel Melnick
Substance Abuse and Mental
Health Services Administration

Robert P. Parker
Bureau of Economic Analysis

Charles P. Pautler, Jr.
Bureau of the Census

David A. Pierce
Federal Reserve Board

Thomas J. Plewes
Bureau of Labor Statistics

Wesley L. Schaible
Bureau of Labor Statistics

Rolf R. Schmitt
Bureau of Transportation
Statistics

Monroe G. Sirken
National Center for
Health Statistics

Robert D. Tortora
Bureau of the Census

Alan R. Tupek
National Science Foundation

Denton R. Vaughan
Social Security
Administration

Robert Warren
Immigration and Naturalization
Service

G. David Williamson
Centers for Disease Control
and Prevention

PREFACE

The Federal Committee on Statistical Methodology was organized by the Office of Management and Budget (OMB) in 1975 to investigate issues of data quality affecting Federal statistics. Members of the committee, selected by OMB on the basis of their individual expertise and interest in statistical methods, serve in a personal capacity rather than as agency representatives. The committee conducts its work through subcommittees that are organized to study particular issues and prepare working papers presenting their findings. The subcommittees are open by invitation to Federal employees who wish to participate. This is the 23rd Statistical Policy Working Paper published under the auspices of the committee since its founding.

On May 25-26, 1994, the Council of Professional Associations on Federal Statistics (COPAFS) hosted a "Seminar on New Directions in Statistical Methodology." Developed to capitalize on work undertaken during the past fifteen 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 and on work on statistical standards undertaken by the Statistical Policy Office at OMB. The subjects covered at the seminar included:

- Economic Classification Revisions
- Disclosure Limitation Methodology
- Customer Surveys
- Advances in Data Editing
- Time Series Revision Policies
- Incentives in Surveys
- Computer Assisted Survey Information Collection
- Longitudinal Surveys
- Cognitive Testing and Self-Administered Questionnaires
- Statistical Uses of Administrative Records
- Small Area Estimation
- Nonresponse in 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 23, published in three parts, presents the proceedings of the "Seminar on New Directions in Statistical Methodology." In addition to providing the papers and formal discussions from each of the twelve sessions, the working paper includes Graham Kalton's keynote address, "Improving the Quality of Federal Statistics," and comments by Norman M. Bradburn, Robert M. Groves, and Katherine K. Wallman at the closing session, "Toward 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 but also prepared these materials for publication.

Table of Contents

Wednesday, May 25, 1994

Part 1

KEYNOTE ADDRESS

IMPROVING THE QUALITY OF FEDERAL STATISTICS	1
Graham Kalton, Westat, Inc.	

Session 1 - ECONOMIC CLASSIFICATION REVISIONS

ECONOMIC CLASSIFICATIONS IN THE NEW NORTH AMERICAN INDUSTRY CLASSIFICATION SYSTEM (NAICS).	17
Jack E. Triplett, Bureau of Economic Analysis	
REVISING THE UNITED STATES STANDARD OCCUPATIONAL CLASSIFICATION (SOC) SYSTEM.	29
Thomas J. Plewes, U.S. Bureau of Labor Statistics	
COMMENTS ON THE REVISIONS OF THE STANDARD INDUSTRIAL AND OCCUPATIONAL CLASSIFICATIONS.	34
Joel Popkin, Joel Popkin & Company	
COMMENTS ON ECONOMIC CLASSIFICATION REVISIONS.	37
Joe Matthey, Federal Reserve Board	

Session 2 - DISCLOSURE LIMITATION METHODOLOGY

RESTRICTED DATA VERSUS RESTRICTED ACCESS: A PERSPECTIVE FROM "PRIVATE LIVES AND PUBLIC POLICIES".	43
George T. Duncan, Carnegie Mellon University	
STATISTICAL DISCLOSURE LIMITATION METHODOLOGY.	57
Nancy J. Kirkendall, Energy Information Administration	
DISCUSSION OF PRESENTATIONS ON STATISTICAL DISCLOSURE METHODOLOGY.	68
Stephen E. Fienberg, Carnegie Mellon University	
SELECTED ASPECTS OF RESTRICTED DATA.	81
Tore Dalenius, Brown University	

Session 3 - CUSTOMER SURVEYS

QUALITY MANAGEMENT FOR CUSTOMER SATISFACTION SURVEYS.	87
Richard M. Devens, Jr., U.S. Bureau of Labor Statistics	
COMPARABILITY IN CUSTOMER SATISFACTION SURVEYS: PRODUCTS, SERVICES AND GOVERNMENT AGENCIES.	99
Michael D. Johnson, University of Michigan	
CUSTOMER SURVEYS: DISCUSSION.	121
Robert M. Groves, University of Michigan	
DISCUSSION.	126
Elizabeth Martin, U.S. Bureau of the Census	

Session 4 - ADVANCES IN DATA EDITING

IMPROVING OUTLIER DETECTION IN TWO ESTABLISHMENT SURVEYS.	137
Julia L. Bienias, David M. Lassman, Scott A. Scheleur, and Howard Hogan, U.S. Bureau of the Census	
TIME SERIES AND CROSS SECTION EDITS WITH APPLICATIONS TO FEDERAL RESERVE DEPOSIT REPORTS	152
David A. Pierce and Laura Bauer Gillis, Federal Reserve Board	
DISCUSSION.	172
Sandra A. West, U.S. Bureau of Labor Statistics	
DISCUSSION.	177
Brian V. Greenberg, U.S. Bureau of the Census	

Session 5 - TIME SERIES REVISION POLICIES

TIME SERIES REVISIONS: THE EFFECTS ON GROSS DOMESTIC PRODUCT.	183
Robert P. Parker and Teresa L. Weadock, U.S. Bureau of Economic Analysis	
RAISING THE NATION'S UNEMPLOYMENT RATE.	209
John E. Bregger, U.S. Bureau of Labor Statistics	
COMMENTS ON PARKER AND WEADOCK, TIME SERIES REVISIONS: THE EFFECTS ON GROSS DOMESTIC PRODUCT.	220
Murray F. Foss, American Enterprise Institute	

Part 2

Session 6 - INCENTIVES IN SURVEYS

TIME, DOLLARS, AND DATA: SUCCEEDING WITH REMUNERATION IN HEALTH SURVEYS.	225
Trena M. Ezzati-Rice, Andrew A. White, William R. Mosher, National Center for Health Statistics, Maria Elena Sanchez, Agency for Health Care Policy and Research	
THE USE OF INCENTIVES TO SURVEY "HARD-TO-REACH" RESPONDENTS: A BRIEF REVIEW OF EMPIRICAL RESEARCH AND CURRENT RESEARCH PRACTICE.	256
Richard A. Kulka, National Opinion Research Center	
DISCUSSION.	290
Diane K. Willimack, National Agricultural Statistics Service	
DISCUSSION.	295
W. Sherman Edwards, Westat, Inc.	

Thursday, May 26

Session 7 - COMPUTER ASSISTED SURVEY INFORMATION COLLECTION

REDESIGNING A QUESTIONNAIRE FOR COMPUTER-ASSISTED DATA COLLECTION: THE CURRENT POPULATION SURVEY EXPERIENCE.	301
Cathryn Dipppo and Anne Polivka, U.S. Bureau of Labor Statistics, Kathleen Creighton, Donna Kostanich, and Jennifer Rothgeb, U.S. Bureau of the Census	
AN OVERVIEW OF THE 1992 CENSUS OF AGRICULTURE CATI SYSTEM.	311
Jeanette K. Mon, U.S. Bureau of the Census	
DISCUSSION: WHAT CAN CAI LEARN FROM HCI?.	363
Mick. P. Couper, Joint Program in Survey Methodology	
DISCUSSION OF TWO PAPERS ABOUT CASIC.	378
Sandra Sperry, Westat, Inc.	

Session 8 - LONGITUDINAL SURVEYS

PANEL DESIGN AND ESTIMATION STRATEGIES IN THE NATIONAL MEDICAL EXPENDITURE SURVEY.	381
Steven B. Cohen, Agency for Health Care Policy and Research	
METHODOLOGICAL ISSUES ENCOUNTERED IN FOLLOWING A COHORT OF EIGHTH GRADERS.	428
Steven J. Ingels, National Opinion Research Center, Jeffrey A. Owings, National Center for Education Statistics	

DISCUSSION.	446
Gary M. Shapiro, Abt Associates	
DISCUSSION.	449
James M. Lepkowski, University of Michigan	

Part 3

Session 9 - COGNITIVE TESTING AND SELF-ADMINISTERED QUESTIONNAIRES

LITERACY LIMITATIONS AND SOLUTIONS FOR SELF-ADMINISTERED QUESTIONNAIRES.	453
Judith T. Lessler and James O'Reilly, Battelle Memorial Institute	
THE LANGUAGE OF SELF-ADMINISTERED QUESTIONNAIRES AS SEEN THROUGH THE EYES OF RESPONDENTS.	470
Cleo R. Jenkins, U.S. Bureau of the Census and Don Dillman, U.S. Bureau of the Census and Washington State University	
DISCUSSION.	517
Jared B. Jobe, National Center for Health Statistics	
DISCUSSION.	524
Roger Tourangeau, National Opinion Research Center	

Session 10 - STATISTICAL USES OF ADMINISTRATIVE RECORDS

IMPROVING DATA QUALITY THROUGH INCREASED DATA SHARING.	531
Edward A. Trott, Bureau of Economic Analysis	
HEALTH REFORM INFORMATION SYSTEMS: GREAT EXPECTATIONS, UNCERTAIN PROSPECTS.	540
Edward L. Hunter, National Center for Health Statistics	
DISCUSSION.	549
Miron L. Straf, Committee on National Statistics, National Academy of Sciences	

Session 11 - SMALL AREA ESTIMATION

SMALL AREA ESTIMATION FOR THE NATIONAL HEALTH INTERVIEW SURVEY USING HIERARCHICAL MODELS.	555
Donald Malec, National Center for Health Statistics, J. Sedransk, State University of New York, Albany	

THE ROLE OF DESIGN BASED VARIANCES AND COVARIANCES IN SMALL DOMAIN ESTIMATION.569
Robert E. Fay, U.S. Bureau of the Census	
DISCUSSION.	595
Phillip S. Kott, National Agricultural Statistics Service	
DISCUSSION.	599
David A. Marker, Westat, Inc.	

Session 12 - NONRESPONSE IN SURVEYS

EXPLORING NONRESPONSE IN U.S. FEDERAL SURVEYS.603
Maria Gonzalez, Office of Management and Budget, Dan Kasprzyk, National Center for Education Statistics, Fritz Scheuren, Internal Revenue Service	
MODEL-BASED REWEIGHTING FOR NONRESPONSE ADJUSTMENT.	625
David A. Binder, Sylvie Michaud and Claude Poirier, Statistics Canada	
DISCUSSION.	646
J. Michael Brick, Westat, Inc.	
DISCUSSION.	649
Joseph L. Schafer, Pennsylvania State University	

TOWARD AN AGENDA FOR THE FUTURE

Norman M. Bradburn.	657
National Opinion Research Center	
Robert M. Groves.	660
The Joint Program in Survey Methodology	
Katherine Wallman.665
Office of Management and Budget	

May 25, 1995

Roger was so special. If you thought you had a good idea, it was smart to first pass it by Roger. He would think about your idea for a while, and then in his ever nice way tell you if he thought it was indeed a good idea. Then he would think about it some more, and come up with an even better idea.

This symposium is dedicated to Roger Herriot, a unique and wonderful person. We will all miss him.

Edward J. Spar
Executive Director
COPAFS

Part 1

KEYNOTE ADDRESS

IMPROVING THE QUALITY OF FEDERAL STATISTICS

IMPROVING THE QUALITY OF FEDERAL STATISTICS

Graham Kalton
Westat, Inc.

1. Introduction

This is the second seminar hosted by the Council of Professional Associations on Federal Statistics (COPAFS) related to the Statistical Policy Working Paper Series of the Federal Committee on Statistical Methodology (FCSM). In deciding on a theme for this talk, I reviewed the working paper for the previous seminar, entitled *Seminar on Quality of Federal Data* (U.S. Office of Management and Budget, 1991). As my title indicates, I have chosen a similar theme. Within the broad subject of "Improving the Quality of Federal Statistics", my main focus is on approaches to improving quality across the federal statistical system as a whole rather than in specific programs. I shall also pay particular attention to the role that the FCSM can best play in achieving the objective of quality improvement.

At the outset I should make it clear that my choice of topic is not to be taken to imply any criticism of the current quality of federal statistics. Indeed, I have a high regard for the federal statistical programs and the professionalism of the federal statistical workforce. Rather, my comments are made from the perspective that, however good the current situation, improvements are always possible.

Equally, my discussion of the role of the FCSM should be interpreted in the same light. Like Bob Groves, who gave the keynote address at the previous seminar, I believe that the FCSM and its working paper series perform a valuable service towards the goal of improving the quality of federal statistics. The working papers also make an important contribution to the general survey and statistical literature. For example, like Bob, I have employed the working papers for teaching purposes. Last fall, I used the excellent recent working paper on indirect estimators (U.S. Office of Management and Budget, 1993a) in my sampling course in the Joint Program in Survey Methodology. The suggestions made below for the Statistical Working Paper Series are offered in the spirit of seeking improvements in a series that has established itself as an extremely useful product.

2. Factors Involved in Quality Improvement

In discussing improvements in the quality of federal statistics, I am interpreting the term "quality" to include not only the production of accurate estimates, but also the production of relevant and timely statistics in a cost-efficient manner, and

the ready accessibility of statistics and data to users. The components of quality thus include:

- *Accuracy.* The estimates produced should have low bias and variance for the parameters being estimated.
- *Relevance.* The parameters being estimated should be the ones that are relevant for users. To ensure relevance, statisticians need to maintain regular and close contacts with users.
- *Timeliness.* The estimates should be up-to-date. The more out-of-date the estimates, the less relevant they are. Equally, other statistical products, such as public use tapes, need to be made available to users in a timely manner.
- *Accessibility.* Statistical products need to be accessible to users through such mechanisms as publications, public use tapes, CD-ROMs, and diskettes. Although restrictions on access may be needed to protect the confidentiality of survey respondents (using such techniques as cell suppression in tables and top-coding and suppression of variables in public use tapes), these restrictions need to be implemented in ways that minimize their consequences for the utility of the data. Good documentation of statistical products is needed to make them readily accessible.
- *Cost-efficiency.* The procedures used to collect and analyze statistical data should be ones that are most cost efficient, taking into account the need to satisfy the other components of quality.

Improving quality thus encompasses: using improved methods of data collection and processing to produce more accurate data; refining definitions of statistical concepts to better meet policy needs; instituting procedures to enable statistics to be produced more rapidly; developing ways to improve access to statistical data; and introducing more cost-efficient methods of data collection, processing and analysis. The broad definition of quality that I am using seems the appropriate one, and one that is consistent with the breadth of the Statistical Policy Working Paper Series. For instance Working Paper 11 deals with industry coding schemes, Working Paper 17 deals with survey coverage, Working Paper 19 deals with computer assisted survey information collection, Working Paper 21 deals with indirect estimates, and Working Paper 22 - an update of Working Paper 2 - deals with statistical disclosure limitation methodology (U.S. Office of Management and Budget, 1984, 1990a, 1990b, 1993a, 1994, and U.S. Department of Commerce, 1978, respectively). Gonzalez's (1994) description of the activities of the FCSM contains a useful review of the broad scope of the Statistical Policy Working Paper Series.

Improvements in the quality of federal statistics can come about in several ways. One is by improving the flow of communication between the user and producer of statistics. As noted above, relevance is a key component of quality and relevance requires the producer to fully understand user needs. Equally, users need to appreciate the inherent limitations in the capabilities of the statistical system that produces the statistics they employ. Regular contacts between users and producers are essential to keep producers aware of changing user priorities and of changes in the structure of society that need to be taken into account in producing relevant statistics in a changing world. The importance of user/producer communication is mainly one that needs to be addressed at the individual statistical program level. Since I am focussing on improving quality in the statistical system at large, I will not expand on this important issue here.

A second means for improving the quality of federal statistics is by increasing the use of what my colleague David Morganstein terms Current Best Methods (CBMs). He prefers this terminology to the more usual Standard Operating Procedures (SOPs) because it conveys the principle that the best methods are evolving over time. In this respect, improvements are produced by increasing the awareness of CBMs among those involved in producing federal statistics and by facilitating their use.

A third means of quality improvement is through research on statistical methodology. Such research can serve both to identify problems with existing methods and to suggest improved methods. The results of this research then feed into the evolving CBMs.

In the last issue of the *Survey Statistician*, Morganstein (1993) describes the application of the process of continuous quality improvement in the survey statistics group at Westat. He identifies three primary elements of the program: employee development; documentation and the use of CBMs; and improved technology. These same elements seem equally applicable for improvements in federal statistics.

The challenges of achieving quality improvement across the federal statistics system are, of course, far greater than they are within a single survey statistics department. Indeed, the challenges are much greater in the decentralized U.S. statistical system than they are in centralized systems such as those at Statistics Canada, and the Australian and Netherlands Bureaus of Statistics. This is not the place to discuss the relative merits of centralized and decentralized statistical systems. All that needs to be noted here is that the problems of communication of current best methods in a decentralized system are severe. The large number of U.S. government agencies that are conducting statistical work provides an indicator of the communication challenge. According to OMB's Annual Report to Congress, in 1994 there are around 80 different agencies that receive direct funding

for major statistical programs. The eleven principal federal statistical agencies receive less than two-fifths of the total major statistical program funding.

As discussed later, I see a prime function of the FCSM as being one of encouraging the use of CBMs. The committee can serve this function by developing working papers detailing CBMs and disseminating them to those engaged in federal statistical activities. In the decentralized environment of the U.S. statistical system, dissemination is a major challenge, a point to which I shall return.

3. Contributors to Quality Improvement

In considering the range of contributors to quality improvement in federal statistics, it is useful to distinguish between employee development on the one hand and the development of CBMs and methodological research on the other. A highly-skilled work force is critical for the production of high quality statistics. The essential components of a highly-skilled statistical workforce are, first, the recruitment of well-trained statisticians, with training appropriate to their job requirements and, second, continuing education over the course of their careers to keep them up-to-date with the many advances that are being made.

In response to a shortage in the numbers of trained survey statisticians, at the end of the 1980's members of the federal statistical system pressed for the establishment of a "Center for Survey Methods" to provide instruction and research training at a Washington-based university. I should like to note here the important contributions of Hermann Habermann and the agency heads of the Bureau of Economic Analysis, the Bureau of the Census, the Economic Research Service of USDA, and the Bureau of Labor Statistics, who worked on the proposal for the Center as part of a 1990 legislative initiative under the leadership of Michael Boskin, then chair of the President's Council of Economic Advisers. These efforts were successful, leading to the establishment of the Joint Program in Survey Methodology (JPSM) at the University of Maryland, a joint program of the University of Maryland at College Park, the University of Michigan and Westat, Inc. The program is now underway, with the first year of the MS program in Survey Methodology completed, and with a proposal for a Ph.D. program in progress. I am pleased to be a faculty member of the Joint Program, a program which I believe holds great promise for improving the quality of federal statistics through training. I should like to recognize the strong support given to the program by Kathy Wallman and all in the OMB Statistical Policy Branch.

In addition to the JPSM at the University of Maryland, there have also been expansions to the programs of other universities in the Washington area that are of direct interest to those working

with federal statistics, such as the recently introduced Federal Statistics Certificate and Masters Degree programs at George Mason University and Masters Degree in Statistics for Policy Analysis at The American University. The numerous courses offered by the universities in the Washington area are generally made available to both degree seeking and non-degree seeking students, and they are often given at times chosen to fit in with full-time work schedules. They provide excellent opportunities for federal statisticians to obtain graduate training in a wide range of subjects. Many federal statisticians have, for instance, learned about such topics as variance estimation with complex samples and recent developments in survey methodology at evening courses at George Washington University and the USDA Graduate School. Through such offerings the universities in the Washington area make important contributions to the training of federal statisticians.

With the major advances taking place in all aspects of federal statistics, there is the need for continual updating and upgrading of the skills of the statistical workforce. One has only to reflect briefly on the advances in methods for questionnaire design, computerized data collection, variance estimation, handling nonresponse, small area estimation, and data disclosure limitation that have occurred within the past ten to fifteen years to realize that a substantial investment in continuing education is essential for keeping federal statisticians up-to-date on best current methods.

As well as through university courses taken on an ad-hoc basis, continuing education can be achieved through short courses, seminars and conferences. Perhaps in response to the recent methodological developments, there has been an impressive expansion of such offerings in recent years. An extensive array of continuing education opportunities is now available for federal statisticians. Moreover, many federal statisticians avail themselves of these opportunities, which I take to be a positive indication both of the desire of federal survey statisticians to upgrade and update their skills and of the strong support of the leadership of the statistical programs for continuing education.

For those in the Washington area, the Washington Statistical Society (WSS) has for many years been making major contributions to continuing education through its short courses and its extensive seminar series. The JPSM also now offers regular short courses at both introductory and advanced levels. In addition, continuing education short courses are regularly offered at the annual meetings of the American Statistical Association (ASA), the American Association for Public Opinion Research (AAPOR) and at the biennial sessions of the International Statistical Institute (ISI).

Conferences serve both as a form of continuing education and as a way to stimulate research work. The scientific programs of the ASA and AAPOR annual meetings and the ISI biennial sessions are

rich in contributions relevant to improving federal statistics. In addition, several series of more specialist conferences have been established in recent years, including the Bureau of the Census's Annual Research Conference, the Conferences on Health Survey Research Methods, the international conferences on survey methods, the Statistics Canada symposia, and now the COPAFS seminars.

Federal statistical programs have much to gain from the attendance and active participation of their staffs in such conferences. I would particularly single out the value of international conferences. We need to keep in touch with the statistical developments that are occurring throughout the world. Sometimes statisticians in other countries can benefit from research conducted in the U.S. and sometimes U.S. statisticians can benefit from research conducted elsewhere. In addition, the increasing interest in the production of comparable economic, social, and environmental statistics across countries points to the need for greater contact between, and collaboration of, government statisticians in different countries.

In the area of employee development, I should finally like to note the significant contribution made by the impressive program of research meetings run throughout the year by the Washington Statistical Society. In addition to three short courses, the WSS held as many as 57 meetings during the 1993-94 year, covering a wide range of topics of interest to federal statisticians. Many of the WSS presentations are made by federal statisticians, and the meetings are generally well attended.

Turning to quality improvement through promoting current best methods, there are again many contributors. Much of the work in this area is initiated and conducted by individual statistical programs, but there are important inputs from other bodies. For instance, many programs have advisory committees that provide expert advice on both substance and methods. In addition COPAFS provides advice, as does the Committee on National Statistics (CNSTAT) of the National Academy of Sciences. Panels of CNSTAT have conducted in-depth studies of specific programs and also of many aspects of federal statistical methodology. The latter include studies of missing data (Madow, Nisselson and Olkin, 1983; Madow, Olkin and Rubin, 1983; Madow and Olkin, 1983), surveying subjective phenomena (Turner and Martin, 1984), microsimulation modeling (Citro and Hanushek, 1991), and confidentiality and accessibility of government data (Duncan, Jabine and de Wolf, 1993).

Quality improvements also come about by improving current best methods. Improvements in CBMs arise out of methodological research, and once again there are many contributors. Much important methodological research is conducted by the federal statistical agencies. Much is also conducted in universities, in survey organizations and in other settings, in the U.S. and

elsewhere, and in the government statistical agencies in other countries. The challenge to maintaining CBMs as "current" and "best" is that of keeping abreast of the large volume of methodological research, and applying its results effectively in current practice. Networks of contacts are needed within the federal statistical system and between federal statisticians and those conducting methodological research elsewhere to keep CBMs up-to-date.

Given the many contributors to quality, what should be the role of the FCSM? Clearly, the FCSM is not well-positioned to contribute directly to quality improvements in programs on an individual basis. Rather, its prime role should be to provide a means for transfer of innovations across programs and for coordination of methodologies where called for.

In her contribution at the closing session of the 1991 symposium, Margaret Martin (1990, p.462) succinctly summarized four functions that the FCSM might perform:

- "(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 think that these four functions provide a good agenda for the Committee.

4. Activities of the FCSM

This section considers each of the functions Margaret Martin lists for the FCSM in turn.

4.1 Exchange of Knowledge, Techniques and Experience Among Subcommittee Members.

In forming a subcommittee to produce a working paper on a particular subject, the FCSM draws upon the expertise on that subject that is available throughout the federal statistical workforce. Membership of a subcommittee then potentially provides the opportunity for an individual to engage in discussions with

others working in the subject, often with different perspectives and experiences. Such a dialogue has the important benefit that the exchange of knowledge, techniques and experience can lead to improvements in the methods applied in the statistical programs from which the subcommittee draws its members. This benefit is particularly important when the subject is one that involves only one or two persons in any one program, so that there is little opportunity for within-program dialogue on it.

In practice, Margaret Martin's comments suggest that subcommittees often have little time for such productive dialogue. Rather, much subcommittee work is report drafting and reviewing, activities that are performed in evenings and at weekends. If this is the case, it is unfortunate: a valuable function of the subcommittee is being lost.

I appreciate that this may not be a good time to ask for additional resources. Nevertheless, the leadership of the statistical programs should recognize the significant rewards that can accrue to their programs and to the federal statistical system more generally from subcommittee activities, and they should seek to ensure that adequate resources are provided to enable the subcommittees to carry out their work as effectively as possible. In part, this means allowing subcommittee members sufficient time to fully perform their roles and in part it means providing each subcommittee with appropriate support staff to work efficiently. The latter could include administrative staff to organize meetings and maintain schedules; editorial staff to help with the production of the working paper; and junior statisticians to serve as research assistants to help with literature reviews and bibliographies if needed (an activity that can provide a valuable learning experience for the junior statisticians).

4.2 Production of Working Papers.

The FCSM has stimulated the production of 22 papers in the Statistical Policy Working Paper Series to date. As I have already remarked, these papers make an important contribution to improving the quality of federal statistics, and to the survey statistics literature more generally. I am therefore somewhat concerned that there appear to be possible signs of some slackening in the pace of working paper production in the last few years. I hope that this is not a true loss of momentum, because I believe there is much more that could usefully be done.

Most of the working papers that I have seen contain valuable descriptions of the applications of the methodology under study across a range of statistical programs. They thus provide a useful review of the state of current practice and help to foster cross-fertilization among programs. To the extent that the programs reviewed are employing current best methods, they document what those methods are. My concern is that the focus may be too narrow.

I think that the working papers would sometimes be improved by a broader perspective on current best methods, examining both the methods used in the government statistical programs of other countries and those used outside government. I acknowledge that some subcommittees attempt to go in this direction, but I think that a more systematic approach along these lines would enhance the value of the working paper series.

In his address, Bob Groves suggested the possibility of including members from outside the federal statistical system in the subcommittees. I note that this suggestion was adopted for the disclosure limitation working paper, with Tom Jabine serving as a member of the subcommittee. I think that this suggestion merits more widespread application. Other possibilities include inviting outside experts to make presentations at subcommittee meetings, arranging small workshops for subcommittee members and outside experts to discuss the issues, and inviting outside experts to review draft working papers. Individuals from outside the federal statistical system may even be asked to draft one or more chapters for a working paper.

If a working paper is to be viewed as a document of best current methods, then it should do more than simply review current practices. It should include recommendations for what are the best current methods, recognizing the variety of different circumstances in which the methods may be applied. To reach agreement on such recommendations may often be difficult, and clearly requires much discussion among the subcommittee members. Lack of sufficient discussion time may well be the reason that the recommendations in the working papers are often not as developed as would be desirable.

Another consequence of viewing the working papers as a means of promoting current best methods is that they should be seen as evolving documents that need to be updated as improved methods are developed. An example here is the latest working paper on statistical disclosure that updates a 1978 working paper to take account of the major advances that have occurred in the intervening period. Progress in recent years in other areas suggests the need to update other working papers, for instance those on developing questionnaires, telephone data collection, the use of microcomputers, and even the fairly recent working paper on computer assisted survey information collection.

The working papers should be prepared to meet the needs of their primary readership, which I take to be those working on federal statistical programs. They should aim to address the questions to which these readers would like answers. In this regard, I should like to recall the wide range of statistical programs that I have outlined earlier, many of which are relatively small. It is in fact the smaller programs that are likely to benefit most from the working papers, since they necessarily lack

the range of expertise that is often internally available in the large statistical agencies. The needs of the smaller programs should be borne in mind in preparing a working paper. To ensure their needs are met, it would be advisable to secure adequate representation of the smaller programs on each subcommittee.

The working papers are valuable only to the extent that they are read. Many able statisticians devote a great deal of effort to the production of each working paper. However, it is my impression that less effort goes into the distribution of the product. The working papers need to reach the desks of those for whom they were written, and mechanisms are needed to ensure that this is achieved. It is also valuable to have a widespread distribution outside the federal statistical system both in the U.S. and abroad. The papers have a great deal to offer to those involved in statistical work in many organizations and countries, and their exposure to a wide spectrum of readers opens up greater likelihood of future improvements. To achieve greater circulation of the working papers it may be useful to publicize them more extensively in appropriate newsletters and journals in the U.S. and abroad and to build up an international network of contacts to aid in the distribution. The recent article describing the working paper series by Gonzalez (1994) is helpful in this regard.

I am not in a position to suggest the best distribution system for the working papers in the federal statistical programs. One possibility might be to identify an individual in each program to serve as a liaison to the FCSM, and send copies of the working papers to that individual. The individual might also be asked to provide suggestions of topics for the FCSM to study. Another possibility is to organize a well-publicized workshop on the topic of a working paper as it is released. Since the working papers have become substantial documents, the workshop could provide a useful primer for those interested in its contents. To some extent, this COPAFS seminar serves such a role, but it is more general in nature spanning the contents of many working papers. The WSS may be able to play an important role in helping to achieve a wide dissemination of the working papers to statisticians in the Washington area.

4.3 Areas for Improvement and Directions for Research.

A number of the working papers indicate areas for improvement and for research, but these issues are not as fully developed as might be desirable. I attribute this situation to the limited discussion time available to the subcommittees. To identify needed improvements goes beyond describing current methods to pinpointing their weaknesses and coming up with ways by which the weaknesses may be addressed. Developing an effective research agenda requires a great deal of deliberation by the subcommittee.

Subcommittees of the FCSM are appointed on the basis of their technical expertise in the given subject area. As such, they are well-positioned to determine incremental research agendas for the given subject. They are, however, less suited to making proposals for major restructuring. In the last seminar, Fritz Scheuren (1993) and Steve Fienberg (1993) talked about the possibility of paradigm shifts in federal statistics. It would be useful to consider setting up federal committees of a different type, composed of individuals with wide experience and broad vision, to examine the possibilities of major changes in the ways federal statistics are produced. Users of statistical data have an important role to play in such committees. As an example, the possibility of continuous measurement in place of the Census long form, which is currently under discussion at the Bureau of the Census, raises a number of possibilities for substantial changes in other data collection efforts. Such committees may be separate from the FCSM, but they should maintain close contacts with it.

4.4 Developing Consensus Across Statistical Programs.

Margaret Martin notes that the objective of obtaining consensus on definitional, conceptual and classification issues has not been well met by the activities of the FCSM. Such consensus building requires lengthy discussions, and shortage of discussion time may again be the root of the problem. Also, different programs will have vested interests in preserving their own definitions, and that will make the attainment of consensus difficult. With a decentralized statistical system, the risk of definitional differences occurring when several programs overlap in their subject matters is high. Consensus building on definitions and methods across programs holds promise of significant advances in fields that cut across different agencies (e.g., aging, children, disability).

5. Topics for Future Working Papers

In concluding, I shall take the opportunity to put forward some specific suggestions for future working papers. Before giving them, I should however like to make two general points. First, I think that the FCSM should have a mechanism for generating suggestions from the federal statistical community at large. At an earlier point, I suggested that liaison persons be appointed in each program. If that suggestion were adopted, one role of those appointed could be to seek suggestions from their colleagues and to forward them to the FCSM. Another possibility is for the FCSM to convene meetings from time to time, perhaps in conjunction with the WSS, to discuss possible subjects for working papers.

My second point concerns the form of the working papers. With the needs of the statisticians working in the smaller statistical programs in mind, I suggest that the FCSM could usefully commission

some of the working papers to be prepared in a manual-style format, reviewing the given methodology in a relatively nontechnical and applied way, and giving practical advice on the implementation of the methodology (e.g., the availability of software). Manuals of this type could be extremely helpful to those inexperienced in the use of the methodology. They need not be lengthy documents; indeed the shorter the document, the more useful it might be. Working Paper 9 on *Contracting for Surveys* (U.S. Office of Management and Budget, 1983) is along these lines. Other illustrations are provided by the manuals on sampling errors (Butcher and Elliot, 1986) and on weighting for nonresponse (Elliot, 1991) produced at the U.K. Office of Population Censuses and Surveys.

In addition to updating some of the existing working papers as discussed above, my specific suggestions for new working papers, undoubtedly blinkered by my own interests, are:

- *Quality profiles.* The error profile for the CPS (Brooks and Bailar, 1978), which was the third report in the working paper series, was an important advance in treating total survey error. Since then the SIPP Quality Profile (Jabine, King and Petroni, 1990) and the Schools and Staffing Survey Quality Profile (Jabine, 1994) have appeared, and other quality profiles are being developed. A subcommittee might usefully develop a blueprint of what such quality profiles should contain, and the methods that may be employed to produce the requisite data, based on the experience that has been gained to date.
- *Economic statistics.* At the previous seminar, Bob Groves commented that there is a distinct bias in the working paper series towards household surveys at the expense of economic statistics. I observe no change in that situation, and think that this should be remedied.
- *Customer surveys.* The requirement that government agencies conduct customer satisfaction surveys has brought many agencies with no prior experience of surveys in direct contact with survey research. In response to this situation, the U.S. Office of Management and Budget (1993b) has produced a resource manual on customer surveys and the JPSM has run a series of short courses to provide training in the conduct of such surveys. A detailed working paper on the subject would be extremely useful.
- *Evaluation research.* Large sums of money are spent by many agencies conducting experimental and quasi-experimental studies to evaluate and compare the effectiveness of various programs. A working paper on this subject could make a valuable contribution to this work.

- *Nonresponse adjustment methods.* Considerable advances have been made in methods of weighting adjustment for total nonresponse and imputation methods for item nonresponse since the late 1970s when the CNSTAT Panel on Incomplete Data studied the subject. Imputation methods are also being used more widely. A working paper on weighting and imputation could be particularly useful for those programs that have little prior experience in this area.
- *Variance estimation.* A working paper that examines the current methods and software for variance estimation, that considers the presentation of sampling errors in survey reports, and that deals with the use of generalized variance functions could be extremely useful, especially for those working in the smaller statistical programs.

6. Concluding Remarks

In concluding, let me restate that my suggestions for the Statistical Policy Working Paper Series are made in the spirit of continual quality improvements in what is a very successful activity. My particular plea is to the leadership of the statistical programs to make sure that this work is supported in the way it deserves. The quality of federal statistics derives considerable benefit from the Working Paper Series. The success of an endeavor such as this depends on the tireless support of those behind it. In this case, the FCSM is exceedingly fortunate to have Maria Gonzalez at the helm. Without her unstinting efforts over many years, it could not have succeeded as it has.

REFERENCES

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.

Butcher, B. and Elliot, D. (1986), *A Sampling Errors Manual*. London: HMSO.

Citro, C.F. and Hanushek, E.A. (eds.) (1991), *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling*. Vol. I Review and Recommendations, Vol II Technical Papers. Washington, D.C.: National Academy Press.

Duncan, G.T., Jabine, T.B. and de Wolf, V.A. (eds.) (1993), *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Washington, D.C.: National Academy Press.

- Elliot, D. (1991), *Weighting for Non-response: A Survey Researcher's Guide*. London: Office of Population Censuses and Surveys.
- Fienberg, S.E. (1991), Towards an Agenda for the Future. In U.S. Office of Management and Budget *Seminar on Quality of Federal Data*, pp. 455-461. Statistical Policy Working Paper 20. Washington, D.C.: Statistical Policy Office.
- Gonzalez, M.E. (1994), Improving Data Quality Awareness in the United States Federal Statistical Agencies, *American Statistician*, 48, 12-17.
- Jabine, T.B. (1994), *Quality Profile for SASS: Aspects of the Quality of Data in the Schools and Staffing Survey (SASS)*, NCES 94-340. Office of Educational Research and Improvement. Washington, D.C.: U.S. Government Printing Office.
- Jabine, T.B., King, K.E. and Petroni, R.J. (1990), *Survey of Income and Program Participation (SIPP): Quality Profile*. Washington, D.C.: U.S. Bureau of the Census.
- Madow, W.G., Nisselson, H. and Olkin, I. (eds.) (1983), *Incomplete Data in Sample Surveys. Vol. 1: Report and Case Studies*. New York: Academic Press.
- Madow, W.G., Olkin, I. and Rubin, D.B. (eds.) (1983), *Incomplete Data in Sample Surveys. Vol. 2: Theory and Bibliographies*. New York: Academic Press.
- Madow, W.G., and Olkin, I. (eds.) (1983), *Incomplete Data in Sample Surveys. Vol. 3: Proceedings of the Symposium*. New York: Academic Press.
- Martin, M.E. (1991), Towards an Agenda for the Future. In U.S. Office of Management and Budget *Seminar on Quality of Federal Data*, pp. 462-464. Statistical Policy Working Paper 20. Washington, D.C.: Statistical Policy Office.
- Morganstein, D. (1993), The Application of Total Quality Management (TQM) Within a Survey Statistics Department. *Survey Statistician*, July 1993, 5-7.
- Scheuren, F. (1991), Paradigm Shifts: Administrative Records and Census-taking. In U.S. Office of Management and Budget *Seminar on Quality of Federal Data*, pp. 53-65. Statistical Policy Working Paper 20. Washington, D.C.: Statistical Policy Office.
- Turner, C.F. and Martin, E. (eds.) (1984), *Surveying Subjective Phenomena. Vols. 1 and 2*. New York: Russell Sage Foundation.

U.S. Department of Commerce (1978), *A Report on Statistical Disclosure and Disclosure-Avoidance Techniques*, Statistical Policy Working Paper 2. Washington, D.C.: Office of Federal Statistical Policy and Standards.

U.S. Office of Management and Budget (1983), *Contracting for Surveys*, Statistical Policy Working Paper 9. Washington, D.C.: Office of Information and Regulatory Affairs.

U.S. Office of Management and Budget (1984), *A Review of Industry Coding Systems*, Statistical Policy Working Paper 11. Washington, D.C.: Statistical Policy Office.

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

U.S. Office of Management and Budget (1990b), *Computer Assisted Survey Information Collection*, Statistical Policy Working Paper 19. Washington, D.C.: Statistical Policy Office.

U.S. Office of Management and Budget (1993a), *Indirect Estimators in Federal Programs*, Statistical Policy Working Paper 21. Washington, D.C.: Statistical Policy Office.

U.S. Office of Management and Budget (1993b), *Resource Manual for Customer Surveys*. Washington, D.C.: Statistical Policy Office.

U.S. Office of Management and Budget (1994), *Report on Statistical Disclosure Limitation Methodology*, Statistical Policy Working Paper 22. Washington, D.C.: Statistical Policy Office.

Session 1

ECONOMIC CLASSIFICATION REVISIONS

Economic Classifications in the New North American Industry Classification System (NAICS)

Jack E. Triplett
Chief Economist, Bureau of Economic Analysis
Chairman, Economic Classification Policy Committee

Prepared for the Seminar on New Directions
in Statistical Methodology¹
Washington, D.C.
May 25, 1994

I. Why Has the United States Mounted this Effort?

The Economic Classification Policy Committee (ECPC), acting on behalf of the Office of Management and Budget (OMB), Statistics Canada, and Mexico's Instituto Nacional de Estadística, Geografía e Informática (INEGI), have agreed to create a new North American Industry Classification System (NAICS). Differences in the internal pressures on the three countries' respective statistical systems create corresponding differences in our motivations, and in the constraints we face. That the three countries have joined together in this effort suggests that the similarities among us may be more significant than the differences. Nevertheless, some of the pressures that have influenced classifications in the United States are not irrelevant to international as well as national discussions of classification systems.

I should first emphasize that the U.S. Standard Industrial Classification (SIC) system is quite old (it dates from the 1930's), it has been well tried in statistical agency collection programs, and it has often been revised (the last time in 1987) in an attempt to keep it up to date. Data users have had abundant experience with the industry data that this classification system provides, and they have had many years to learn of its strengths and shortcomings. Statistical agencies have had many opportunities to react to user experiences, within the parameters that have guided this classification system for the past 50 years.

Yet, increasing public dissatisfaction with the U.S. SIC system has been expressed through its last several revisions. Discussions of the adequacy of the SIC in the United States have occurred not merely or solely in professional exchanges between economists and statisticians, and have not occurred solely within the boundaries of narrowly technical dialogues. In the United States, focus on problems of the SIC has extended widely to the popular press and the business press.

¹ This paper was originally presented at a meeting of The Statistical Programme Committee (NACE), Statistical Office of the European Communities (Eurostat), Luxembourg, March 17, 1994.

Perhaps this public attention to industrial classifications is unique to the United States. To the extent that it is unique, we believe that, at least in part, it is a response to a mature industry classification system that has increasingly been viewed as inappropriate for generating the data that are needed for economic analysis.

The United States has experienced increasing private sector and government sector demands for data for the purpose of economic analysis. As one example, a major statistical initiative has in the last several years been launched [8], and a major priority within that initiative is to improve U.S. data for the measurement and analysis of productivity. We must, therefore, ask: Does the existing SIC system produce the industry data that are wanted for economic analysis, including the industry data needed for productivity measurement? These are exactly the questions that were posed for classification systems at the Williamsburg international conference on classifications [1].

II. The ECPC Examination of Classifications Starts from the Use of Economic Data

The ECPC was established by OMB in 1992 to conduct a "fresh slate" examination of economic classification systems. The ECPC began its investigation with an examination of the uses of economic data that are produced using classification systems. What is the purpose of classification systems? When economic data are produced from these systems, what are the uses of the data for which classifications are designed? The review of these issues is contained in ECPC Issues Paper No. 1, "Conceptual Issues" [3].

The ECPC's approach is a departure from the traditional approach to classifications, at least as it has developed in the United States. In the traditional view, there are many uses for data, and because there are many uses, it has been believed that the classification system must produce data for all the uses. This means, effectively, that the uses of data have relatively little ultimate role in the design of the classification system, because the requirements for different users tend to cancel each other. The traditional classification is a compromise between competing ends. The nature of these compromises is not dictated by the use of the data, nor do the designers of the classification system have a framework from which to examine the costs to the data user of the compromises incorporated into the system.

The ECPC, in common with the traditional view, also recognizes that there are multiple uses for data that are produced by classification systems. However, in contrast to the traditional view, the ECPC has concluded that the economic use of data must determine the design of the classification system; if it does so,

this will assure that the data produced by the classification system meet the intended use.

If there are alternative demands for data that must be grouped or classified, and if these alternative demands for data have different implications for the classification system, the ECPC concludes that different classification systems must be constructed. In the ECPC's view, when different requirements for classifications grow out of different data uses, this implies that different classification systems should be set up, each one designed to meet the intended need. In the traditional approach in the United States, a compromise has been sought to meet everyone's needs within a single classification system.

If tailoring classification systems to the needs for data implied a very large number of different classification systems, the ECPC's position might be impractical. We also have concluded, however, that the major analytic needs for classified data can themselves be grouped into two major classes of uses. This matter is discussed at some length in ECPC Issues Paper No. 1 [3].

Briefly, one class of uses requires that a classification system be erected on a production-oriented concept (which may also be called a supply-based concept). A second major class of uses requires that data be grouped according to a market-oriented concept (which may also be called a demand-based concept).

Thus, the major difference between the ECPC's position and the traditional one in the United States condenses to the following questions: Should there be two different classification systems, each designed to produce data for one of the two classes of uses? Or should there be only one compromise system for both classes of uses?

In international discussions of classifications, the situation is a little different from the U.S. tradition, because multiple classifications already exist. The United Nations systems include the International Standard Industrial Classification of all Economic Activities, Third Revision (ISIC), and the Central Product Classification (CPC), and Europe has Nomenclature des Activités économiques des Communautés Européennes (NACE), Classification des Produits associée aux Activités (CPA), and another system called PRODCOM.

Yet, the issues that have been debated are quite similar: What are the data uses for which different classifications are required? Should some of the classification systems (usually, the product classification system) be connected in some manner to the others (the industry classifications)? And whatever distinctions can be drawn in principle among the various systems (by reading their introductions and statements of principles, for

example), their implementations are marked by compromise. The connections between data use, classification concept, and construction of the classification system have been influenced by compromise among competing demands for data, rather than by the determination to tailor each classification system to a major economic use of grouped data.

Though seldom stated explicitly in this language, almost all of the literature discussing classification principles, whether in the United States or in international forums, can be understood as a conflict between designing a production-oriented, or supply-based, classification system and designing a market-oriented, or demand-based, classification system. That is, much discussion of classification problems concerns, at the root of the matter, the conflict between providing data for production-side economic analysis, on the one hand, and for market- or demand-oriented analysis on the other. The ECPC's issues papers, particularly ECPC Issues Paper No. 1, attempt to make more clear and explicit what has unfortunately remained implicit in much of the past discussion of economic classifications.

The distinction made in ECPC Issues Paper No. 1, however, is actually quite old. After this project was well under way, David Wharton of Statistics Canada called my attention to a very enlightening article on economic classifications published by R.H. Coats in 1925. Coats' pertinent observations for our time include the following:

"...the basic principle in classification is that mutually exclusive concepts may not be united on an equality in the same category.... It is precisely this elementary rule that statisticians too often ignore. Called upon for statistics of aggregates from many and diverse standpoints, they attempt to meet the demand within the limits of a single classification. This leads inevitably to confusion as between principles...."

"...To state as was stated in the resolution originally tabled at the Geneva Conference on Labour Statisticians that a combination of principles must be adopted is surely to abandon the issue prematurely...." [2, emphasis in original].

Allowing for changes in economic language over the past 70 years, a subsequent passage of Coats' article can be interpreted as discussing production-oriented and market-oriented principles as alternative concepts for classification systems. Coats proposed also a third principle--distinguishing the stage of process in the hierarchy of the classification system.

III. Why Was the Production-Oriented Concept Chosen for NAICS?

A statement adopted by the ECPC, Statistics Canada, and Mexico's INEGI reads, in part:

"The uses of industrial statistics which include measuring productivity, unit labor costs, and the capital intensity of production require that information on outputs and inputs be used together. Moreover, statistical agencies in the three countries expect to be called upon to produce information on inputs and outputs, industrial performance, productivity, unit labor costs, employment, and other statistics in order to analyze the effects of the North American Free Trade Agreement. An industry classification system erected on a production-oriented, or supply-based, conceptual framework will assure maximum usefulness of industrial statistics for these and similar purposes. Therefore, the three countries agree that the new North American Industry Classification System should conform to a production-oriented economic concept" [5].

The reasoning behind the three countries' decision may be summarized as follows. An industry is grouping of economic activities. Though it inevitably groups the products of the economic activities that are included in the industry definition, it is not solely a grouping of products.

Put another way, an industry groups producing units. Accordingly, an industry classification system provides a framework for collecting the variables that describe production--inputs and outputs--together on a consistent basis. The industry system thus groups data for analyses for which it is important that inputs and outputs be used together.

What uses of economic data require that inputs and outputs be used together, and be collected on the same basis? Such uses include production analyses, productivity measurement, studying input usage and input intensities, and so forth. For these uses, producing units should be grouped together by similarities in their production processes, which is exactly the production-oriented concept discussed in ECPC Issues Paper No. 1.

Thus, the North American countries have chosen the production-oriented concept as the framework for industry statistics (a) because important production analysis uses of data require groupings of producing units, and (b) because these uses are the ones that require that inputs and outputs be collected together on a comparable basis. The production-oriented concept for classification systems is discussed at greater length in ECPC Issues Paper No. 1

IV. Planning for the Alternative, Product Grouping System

A classification system that groups or aggregates products is a very different system from an industry classification system. An example of a product grouping system is the Central Product Classification (CPC) prepared by the United Nations Statistical Office.

A product grouping system² satisfies a different need--a different use of data--from the one served by an industry classification system. A product grouping system is used for analyses from the demand side--to define markets to study market power or to conduct marketing studies, for demand estimation, for determining the extent of substitution among commodities, and so forth. One does not want a product grouping system for studying productivity; an industry classification system produces the data for productivity analysis.

A product grouping system has the following characteristics:

- (a) It should incorporate, and facilitate the analysis of, relationships among products--demand relations, substitution relations, marketing relationships, uses by consumers or by other ultimate purchasers.
- (b) For demand and market analyses, the inputs to production generally do not matter for the intended data use. As a

² The term "product system" has been used to encompass at least three different ideas, only one of which is the product grouping system. The first, which might be termed the "product enumeration system," provides a list of all the products (goods and services) that exist. For example, the Harmonized System (HS) of the Customs Cooperation Council provides in principle a listing of all the products that move in international trade. A product enumeration system can also contain a grouping system, and frequently does so for organizational reasons, though the enumeration system's grouping system is not necessarily constructed to facilitate economic analysis, and is usually not suitable for analytic purposes. The listings in the product enumeration system provide building blocks for the product grouping system, which is described in the text. Finally, one often needs to list the outputs in each industry of the industry classification system. Such a listing has sometimes been given its own name; other times it has been referred to simply as the "index items" or "indexes"--for example, in the United States, the products in the "alphabetical index" of the *Standard Industrial Classification Manual, 1987* [9]. This "index system" also uses, with certain exceptions, the listings in the product enumeration system.

consequence, only the outputs matter in a product grouping system, no information on inputs need be collected.

- (c) Accordingly, product groupings may cut across the producing relationships in establishments, or other producing units. Establishment outputs may be separated and assigned to different product groupings, as the principles of the product grouping system dictate.

Because it satisfies a different data use, a product grouping system is appropriately constructed on a different economic concept from the one that is used for an industry system. A product grouping system requires a market-oriented, or demand-based, economic concept. The market-oriented, or demand-based, concept for economic classifications is discussed at greater length in ECPC Issues Paper No. 1.

Moreover, there is no reason to integrate a product grouping system with an industry classification system, and there is every reason to avoid linking the two where they are in fact different. A product grouping system is intended to meet its own needs, and should meet those needs independently of the industry classification system, which is properly designed to serve a different purpose.³

The three North American countries have agreed that product grouping systems should be established on their own merits, as indicated in the following paragraph from their joint statement:

"The statistical agencies of the three countries also agree that market-oriented, or demand-based, groupings of economic data are required for many purposes, including studies of market share, demands for goods and services, import competition in domestic markets, and similar studies. Each country will provide product data compiled within the framework of its respective statistical system, to meet the need for such information. Recognizing the increasing international trade in goods and services, each country will work cooperatively to help improve commodity classification systems, including the Harmonized System (HS) of the Customs Co-operation Council and the United Nations provisional Central Commodity Classification (CPC) system for services, by coordinating efforts and keeping each agency informed of proposals for changes" [5].

³ Some controversy exists on this point, which I believe has arisen out of failure to distinguish between the purposes for which a product grouping system is needed and the quite different functions of a system that lists the index items in the industry classification system (see footnote 1).

V. Is a Conceptual Classification System Practical?

The approach followed in constructing NAICS involves: (a) taking the economic uses of industry data as a starting point, and (b) deriving an economic concept for industry classifications, making use of the economic theory that underlies the economic analyses that use industry data. Though it is clearly desirable that data be constructed in accordance with the implications of economic theory, and constructed so that the data meet the requirements for economic analysis, there has been some justifiable concern about the practicality of such an endeavor. Can we analyze pragmatically and empirically our present economic classifications with respect to the theoretical requirements for a conceptually based classification system? Can we design new and improved classification systems making use of the theory?

The ECPC and Statistics Canada have produced a number of studies that suggest that the task we have set ourselves is indeed practical.

A. The matrix papers

In two separate studies [6] [10], U.S. and Canadian 4-digit SIC industries were reviewed. Teams in each country asked whether individual industries embodied a production-oriented economic concept, or a market-oriented economic concept.

As explained in ECPC Report No. 1, "Economic Concepts Incorporated in the Standard Industrial Classification Industries of the United States" [6], these two reviews combined understanding of the economic concepts, as developed in ECPC Issues Paper No. 1, with informed judgments about the technologies and the markets that pertain to each detailed 4-digit SIC industry. The reviews were, first, tests to see whether the economic concepts could be implemented in a pragmatic way, using mainly the type of information about industries that has been used in the past to make decisions about the U.S. and Canadian SIC systems. These reviews use the available information to assess economic concepts.

Secondly, the two reviews provide a preliminary assessment of the concepts embodied in the U.S. and Canadian systems by past decisions. Their results are subject to revision on the basis of industry expertise.

Some present 4-digit SIC industries are already constructed along production-oriented lines, or could be, with relatively small adjustments to definitions. In the United States, the study suggests that a little under a fifth (19 percent) of manufacturing shipments come from 4-digit industries that are fully defined on the production-oriented concept, and another two-fifths (actually 45 percent) originate from industries that

could be made consistent with the concept by combining and/or subdividing existing industries (the details for these estimates are contained in [6]).

On the other hand, one could emphasize the other side of the picture. Fully two-fifths of manufacturing industries have no discernible production-oriented basis and nearly as large a proportion of manufacturing shipments arise in these industries; to this one could add the two-fifths of the manufacturing industries that require some adjustments to be fully consistent with the production-oriented concept (as noted in the previous paragraph). From those numbers, it is evident that the U.S. system as presently developed does not conform to the production-oriented concept.

The situation is about the same for the 150 services industries that were reviewed in the U.S. study. Actually, a slightly higher proportion of services industries has been defined to be consistent with the production-oriented concept, but much additional refinement of service industry definitions will be required to produce adequate industrial data.

A little under a quarter (23 percent) of U.S. manufacturing shipments come from SIC industries that have been defined on a market-oriented basis. Nearly half of those (10 percent of manufacturing shipments) are industries that meet the conditions for both production-oriented and market-oriented conceptual systems: These were designed "Ideal" industries in the review, because statistics for them are appropriate for both of the major classes of economic analysis.

Another 35 percent of shipments arise in industries that have some market-oriented basis in their definitions. Many of those are cases where production-oriented and market-oriented reasoning has been combined into a compromise industry definition that fully satisfies neither.

In the traditional view, the classification problem is to find ideal industries--those that are satisfactory for both production and market analysis--on the implicit assumption that deviations from ideal in practice can be handled as "special cases," for which case-by-case compromises can be effected. That ideal industries have been found in the United States in only 10 percent of the cases is a measure of how far the traditional view of the classification problem is from the empirical reality of actual industry structure.

B. Heterogeneity index

The ECPC has developed a new statistical approach that will assist in determining production-oriented economic groupings. This method is explained in ECPC Report No. 2, "The Heterogeneity

Index: A Quantitative Tool to Support Industrial Classification" [7], which applies the new method to 4-digit manufacturing industries in the United States.

The heterogeneity index is based on the following regularity: When producing units have the same production function and face the same input prices, each producing unit will exhibit the same proportionate expenditure on each productive input (shares of inputs in total cost) as will every other producing unit. When producing units have different production functions, their input expenditures will differ. The heterogeneity index measures the dispersion in relative expenditures on inputs among the establishments in an industry, or in a proposed industry. When the establishments have the same production functions, they will have the same input shares in total costs, and the heterogeneity index will be zero. The value of the index rises as establishment heterogeneity within the industry increases; that is, the index takes on a larger value as establishments with dissimilar production processes are combined into a single category.

The heterogeneity index can be used, in conjunction with other information, to judge how closely existing industries correspond to a production-oriented grouping. It can also be used to evaluate proposals to form new production-oriented industries, or to break apart or combine existing ones.

ECPC Report No. 2 also compares the results from the new heterogeneity index with the judgments that were incorporated into the matrix of ECPC Report No. 1. Note that the matrix judgments were formed before the heterogeneity index was computed, so that the matrix and the index could be used as independent evaluations. The degree of correspondence between these two completely independent evaluations, though not perfect, is both intriguing and promising (see ECPC Report No. 2).

The heterogeneity index is an important new tool that is available for implementing a production-oriented economic concept in a classification system.

C. Services classifications

The three North American countries have agreed to give special attention to classifications for services industries, as well as for high-tech and new and emerging industries. The classification of services poses special difficulties and because of this the ECPC has released a paper (ECPC Issues Paper No. 6, "Services Classifications" [4]) that discusses the application of a production-oriented economic concept to services industries.

The ECPC has been especially challenged by those who have said that our approach may be practical for goods but will not work

for services. We believe the application of production-oriented reasoning to services industries is practical, and ECPC Issues Paper No. 6 discusses practical interpretations of the economic concept. Moreover, the matrix exercise (ECPC Report No. 1) also applied the production-oriented concept to services industries in a pragmatic way, and we believe that this exercise shows that the production-oriented concept can be applied to services.

The task of classifying services industries will, however, be especially difficult. Additional special reports on the classification of services will be released as the work proceeds.

V. Applications of the ECPC's Research Approaches to the Classification Systems of Other Countries

We believe it would be especially rewarding to know the economic concepts that have been incorporated into industry definitions in classification systems outside the United States and Canada. It would also be valuable to test the heterogeneity index on the industry classifications of other countries. Exchanging the results of similar studies carried out on classification systems in use in different countries would provide a good way to determine where--that is, in which classification systems--the best ideas for industry groupings are to be found.

In the past, comparisons of different classification systems have more or less given the result: We do ours this way and we think ours is best, and you do yours that way and you think yours is best. However, we can now do better: Carrying out analysis of classification systems along the lines of ECPC Reports Nos. 1 and 2 and the Statistics Canada study [10] potentially provides a much more productive exchange of information than has been possible in the past. Rather than "splitting the difference" between mutually exclusive classification outcomes, performing some economic analysis on classification systems, of the type incorporated into the Statistics Canada and the two ECPC reports described in this paper, would produce new and valuable information for improving industry classifications.

Moreover, explicit conceptual analyses of classification systems would offer the potential for melding the international desire for comparability in industrial statistics with the goal of improving the available industrial statistics for the needs of users. Rather than setting the two goals against each other, or elevating the one over the other, as has sometimes inadvertently been true in the past, we need to gain wider understanding and support for a new approach: Constructing internationally comparable industrial statistics--where internationally comparable economies exist--that conform to a consistent economic concept provides the worldwide best course for the future of industrial statistics.

REFERENCES

- [1] Bureau of the Census, *Proceedings, International Conference on Classification of Economic Activities*, Williamsburg, Virginia: U.S. Department of Commerce, November 6-8, 1991.
- [2] Coats, R. H., "The Classification Problem in Statistics," *International Labour Review*, April 1925, pp. 511 and 520.
- [3] Economic Classification Policy Committee, Issues Paper No. 1, "Conceptual Issues," *Federal Register*, March 31, 1993, pp. 16991-17000; Economic Classification Policy Committee, Bureau of Economic Analysis (BE-42), U.S. Department of Commerce, Washington, D.C. 20230.
- [4] _____, Issues Paper No. 6, "Services Classifications," March 1994. Available from Economic Classification Policy Committee, Bureau of Economic Analysis (BE-42), U.S. Department of Commerce, Washington, D.C. 20230.
- [5] _____, *Federal Register* (forthcoming). Available from Economic Classification Policy Committee, Bureau of Economic Analysis (BE-42), U.S. Department of Commerce, Washington, D.C. 20230.
- [6] _____, Report No. 1, "Economic Concepts Incorporated in the Standard Industrial Classification Industries of the United States," May 1994. Available from Economic Classification Policy Committee, Bureau of Economic Analysis (BE-42), U.S. Department of Commerce, Washington, D.C. 20230.
- [7] _____, Report No. 2, "The Heterogeneity Index: A Quantitative Tool to Support Industrial Classification," May 1994. Available from Economic Classification Policy Committee, Bureau of Economic Analysis (BE-42), U.S. Department of Commerce, Washington, D.C. 20230.
- [8] U.S. Department of Commerce, Bureau of Economic Analysis, *Survey of Current Business* 70 (February 1990): 2.
- [9] U.S. Executive Office of the President, Office of Management and Budget, *Standard Industrial Classification Manual, 1987*, 705 pages. (For sale by: National Technical Information Service, 5285 Port Royal Road, Springfield, Virginia 22161, Order no. PB 87-100012.)
- [10] Young, Kenneth, "The Conceptual Basis of the Standard Industrial Classification," Standards Division, Statistics Canada, February 1994.

REVISING THE UNITED STATES STANDARD OCCUPATIONAL CLASSIFICATION (SOC) SYSTEM

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

I. Background

Historically, various United States Federal agencies, primarily the Department of Labor and the Bureau of the Census, have developed their own separate occupational classification systems, designed to meet their own specific statistical and programmatic needs. The lack of comparability between these various sources of occupational information and data led to multi-agency interest in and action to develop a Standard Occupational Classification (SOC) system, beginning in 1966. The SOC, first published in 1977 and revised once in 1980, was intended to provide a mechanism for cross-referencing occupation-related data collected by various economic and social statistics programs in order to maximize the analytical utility of these data.

The major underlying principle of classification in the SOC is work performed, not skills, training, education, licenses or other credentials. More specific occupations are grouped into the most detailed SOC categories based on their similarity in terms of work tasks and activities. Other classification principles include the following: SOC groupings are independent of the work setting, unless it alters the nature of the job; supervisors are identified separately from workers; large or small size is not a determinative factor for separate identification; and comparability to the international standard classification of occupations (ISCO).

The SOC was intended to be comprehensive in coverage, including all occupations for which work is performed for pay or profit, including unpaid farm work. The 1980 SOC was comprised of 664 distinct occupations at the most detailed level. It was not intended to meet all specialized analytical or organizational management purposes, but to serve as a general tool for reconciling various sources of occupational data.

In 1983, the major sources of U.S. occupational employment data -- the establishment-based Occupational Employment Statistics (OES) survey and the Current Population Survey (and Decennial Census) of households -- became more comparable when each adopted a new classification structure based on the SOC.

II. The Need for a New SOC

The SOC, unfortunately, never was implemented fully across all Federal occupation-related data collection efforts. Various

federal agencies continue to use their own distinct occupational classification structures. For example, the Department of Labor's Employment and Training Administration uses the Dictionary of Occupational Titles (DOT); the Department of Education uses its Classification of Instructional Programs (CIP); and the Office of Personnel Management has its own occupational classification structure. As a result, reconciling different occupational data sources continues to be difficult at best. In addition, the 1980 version of the SOC is outdated, as new occupations -- particularly in technical and health-related fields -- have emerged since that time (and are incorporated into some of the current occupational classification structures).

There are other reasons that attention recently has focused on occupational information. Concern with the quality of the U.S. workforce, skill formation issues, and changes in occupational structures due to new technology and shifts to "high-performance" work organizations, all highlight the importance of accurate, timely, and comparable occupational information to support program planning, career guidance, and training development. As such many users and producers of occupational data feel that it is time to re-examine the SOC and to develop a classification structure that meets the occupational information needs of the twenty-first century.

III. Actions to Inform the SOC Revision

In November 1991, the Office of Management and Budget (OMB) designated the Department of Labor as the lead agency to coordinate the development of a new U. S. Standard Occupational Classification (SOC) system by 1997, in time for implementation in the 2000 Census. Since that time, the Bureau of Labor Statistics' Office of Employment and Unemployment Statistics and the Dictionary of Occupational Titles (DOT) staff of the Employment and Training Administration (ETA) have been working together to organize activities aimed at developing information and alternative approaches related to classification principles for the new SOC. These activities have included commissioning contract papers on major occupational classification issues.

An International Occupational Classification Conference was held in June 1993, sponsored by the Bureau of Labor Statistics. The Conference provided a forum for the discussion of new ideas and alternative approaches to occupational classification issues. It included many individuals and agencies directly involved with the occupational classification user community, as well as international occupational experts from numerous countries. The papers, discussions, and ideas generated at the Conference are serving to inform revision activities for the SOC.

Some of the major issue areas addressed at the Conference are described below.

1. *New challenges and alternative approaches to occupational classification:* Currently, all federal occupational classification systems are based on work performed or job titles. As the pace of occupational change has increased, many people are becoming more concerned with issues of skills transferability between jobs or occupations in order to facilitate transitions in an increasingly volatile economic environment. An important issue raised during the conference is whether a new U.S. SOC should be based primarily on skill type and skill level, rather than work performed.

2. *The feasibility and desirability of creating a unified occupational classification structure for government statistical and programmatic purposes:* Although some Federal agencies may prefer to maintain their separate classification structures, others feel that net value could be provided to users of occupational data by developing a more unified Federal classification structure. At a minimum, there seems to be consensus that a more unified Department of Labor occupational classification structure is desirable, and movement in this direction has been occurring, even prior to the Conference. In its final report, the Secretary of Labor's Advisory Panel on the Dictionary of Occupational Titles recommended that a revised Dictionary conform to the classification structure of a revised SOC system and, in the interim, conform to the Bureau of Labor Statistics' Occupational Employment Statistics (OES) system.

3. *How a revised SOC could meet the needs of users of occupational information who are dissatisfied with the current classification systems:* Due to the current system of multiple occupational classification structures, users must obtain important related information -- such as demographic characteristics, industry and geographical distribution, worker attributes and skill requirements, and wages -- from different sources with different underlying classification structures. As a result, the information obtained from one source is not compatible with information derived from another source, leading to frustration on the part of many users. Another source of dissatisfaction lies with the perceived currency and accuracy of current occupational classification structures. Some structures, including the SOC, have not been updated for more than a decade, and therefore, many new occupations that have emerged as a result of new technology and changed forms of work organization are not included in current classification structures.

4. *International perspectives on occupational classification and lessons for the U.S. SOC revision:* A full day of the conference was devoted to international occupational classification issues. The international experience is important for two reasons: One relates to the international comparability of data, and the other relates to lessons that can be learned from the experience of other countries. A decision to move towards a common international classification system, such as the International Standard Classification of Occupations (ISCO-88), would

inevitably result in a loss of nation-specific occupational detail that many users of national data regard to be critical. In addition, there are questions about the degree to which ISCO-88 is structured on clear, consistent, and appropriate principles. The second reason to examine the international experience is to try to draw lessons from other countries, many of which have recently made substantial revisions to their national occupational classification systems. Issues explored included new approaches to principles of occupational classification (e.g., skill type and skill level); the level of effort and resources required and methodologies used to develop new systems; and the feasibility and desirability of developing a unified national classification system to replace existing disparate ones.

IV. SOC Revision Process

Following the Conference, the Office of Management and Budget established an SOC Revision Policy Committee, chaired by the Bureau of Labor Statistics, with representatives from the Bureau of the Census, the Employment and Training Administration, the Office of Personnel Management, the Defense Manpower Data Center, and, *ex officio*, the Office of Management and Budget (OMB). A Charter for the Committee recently has been approved.

The Policy Committee is charged with an examination of the Federal Government's various occupational classification systems for statistical and administrative uses, and with providing recommendations to OMB on the structure and implementation of a new SOC. The charge to the Committee includes: (1) identifying the major statistical uses of occupational classifications; (2) identifying and developing new concepts, structures, and methodologies to determine what constitutes an occupation; (3) developing a standard classification system based on these concepts; (4) planning the implementation of the new classification system; and (5) ensuring that there is ample opportunity for widespread public participation in the revision process.

The principal use of a revised SOC would be statistical, but it also would serve as a framework for administrative purposes and other occupational classifications. The Policy Committee will evaluate the utility of alternative classification structures in consideration of the following: (1) Ensuring compatibility between the descriptive material of the new Dictionary of Occupational Titles (DOT) and the revised SOC; (2) current public interest in a skills-based classification system; (3) users' needs for historical comparability of data; (4) the expertise of other countries in revising national classification systems; (5) desirability, but not necessity, of compatibility with international occupational classification systems; and (6) the need for all Federal Government occupational classification systems to be part of the SOC framework.

The Policy Committee will adopt processes that ensure ample opportunity for public participation. These processes will involve all stakeholders, including the range of occupational data users, both government and private, as well as data collectors and data providers. The Policy Committee will consider forming a Consultation Group, composed of Federal agencies not represented on the Policy Committee and interested public and private parties (e.g., States, associations, private individuals). Such a group would meet on a flow basis, as necessary, to provide input to the work of the Policy Committee. Notice of the Policy Committee's work will be widespread and will be published in the *Federal Register*, and all interested parties will be given the opportunity to be included on a mailing list.

The conceptual framework for the new SOC is to be completed prior to July 1995 to allow for testing related to the 2000 Census, as well as for the administration of the 1996 DOT National Content Test. The completed occupational classification structure should be available by July 1997 to coincide with development of the 2000 Census.

COMMENTS ON THE REVISIONS OF THE STANDARD INDUSTRIAL AND OCCUPATIONAL CLASSIFICATIONS

Joel Popkin
Joel Popkin and Company

The beginnings of the present effort to revise the Standard Industrial Classification (SIC) originated in the Census Advisory Committee structure in the mid 1980s. At that time, I represented the American Economic Association. At those committee meetings, I recall getting the assignment to comment on papers about how the SIC ought to be revised, and asking myself why I had not drawn a more interesting assignment. I clearly did not recognize that a revolution in economic classification was afoot.

Charles Waite, Associate Director of the Census Bureau, was handing out those assignments, and the papers I commented on were written by his staffers, Pamela Powell-Hill and James Monahan. That was 10 years ago, and I think marks the birth of this much needed and very important current effort to conduct a "clean-slate" revision of the SIC. Following those developments, Charles Waite planned and convened in Williamsburg what turned out to be a seminal international conference on economic classification. About the time the plans for the conference were being initiated, Jack Triplett, Chief Economist of the Bureau of Economic Analysis, presented at the 1990 Census research conference a very important paper illuminating the conceptual issues relevant to the classification of economic activity. Hermann Habermann, then Chief Statistician of the U.S. government, lent support to a continuation and formalization of those efforts.

The project to revise the SIC now has a full head of steam with a target for implementation in the 1997 economic censuses. OMB appointed the BEA lead agency, and Jack Triplett is chairman of the government-wide Economic Classification Policy Committee (ECPC). There are three elements of U.S. leadership in this significant undertaking. The first was the Williamsburg conference itself. The second was the successful negotiation among the countries of NAFTA of an agreement to develop a common, North American industrial classification system (NAICS). The third key element in the pervasiveness of this effort, and its enhanced chances of success, was that the North American plan and the Williamsburg conference were instrumental in prompting Eurostat, the statistical agency of the European Community, to reconsider some decisions it had made about industrial classification and to explore moving in the direction of fundamental rethinking of classification systems that the United States has promulgated.

There are three fundamental kinds of decisions that have to be made in designing an industrial classification system. The first is the selection of the unit of observation. The second is the concept by which individual observations should be grouped. And the third is the hierarchy along which groups should be aggregated.

Decisions about two of those three elements have already been made. The first is that the classification system will retain the establishment as the unit of observation except in cases where its use is not appropriate or feasible. The second is that the underlying concept for classification will embody a production-side approach. Establishments will be grouped together that share the same kind of production techniques and processes. The third issue, that of hierarchy, is still being studied.

Part of my role here today is that of discussant of the SIC revision process that is now underway. Some of you will be familiar with my views and recommendations if you have seen the paper I wrote for the Williamsburg conference.

With respect to the unit of observation, my recommendation was to change from an establishment based system to one which I characterized as focusing on divisions, departments or subsidiaries (DDS) within companies as units of observation. I made that recommendation for three reasons. The first is that the establishment is not as prevalent an economic unit of observation as it once was. That is at least partly due to the advances in telecommunications which permit output to be produced with more inputs obtained from different establishments within the company. That leads me to the second reason I recommended a larger unit of observation such as the DDS. It is that at a higher level of aggregation, the matching of inputs and outputs and the full accounting of all inputs may be more feasible and data collection simplified. The problem posed in using the establishment is that not all inputs can be accounted for, especially some purchased services and inputs of information, technology, and management skills from central offices and other establishments within the company. I felt that by moving the unit of observation to a higher level of aggregation within the company, those inputs could be captured and the activity of separate business units (SBO) or DDSs could be relatively well accounted for and measured. That approach is not new. It is used currently in the Census M3 report on "Inventories, Shipments, New and Unfilled Orders" in which data are collected directly from divisions of companies; and it is also being utilized in the Annual Capital Expenditures Survey (ACES) that the Census Bureau has developed. The third reason I made this recommendation was that it seemed as though it might ease reporter burden to the extent that establishment records are increasingly being consolidated at company or division levels. Nonetheless, the ECPC recommendation to use the establishment, but with a recognition that there may be exceptions, goes some distance to alleviating my concern about the use of the establishment.

With respect to classification concept there were two candidates. The production-oriented approach or the market-oriented approach. Each approach serves many legitimate uses, and both can be justified. I thought it would be inappropriate to recommend multiple classification systems simply because the resources are limited. So I thought it was necessary to recommend one approach. For me, it was the demand approach. The ECPC has adopted the production approach, but also has indicated that it is undertaking work to develop a structure in which outputs can be

classified by market grouping for both goods and services. Such market groupings (commodity product classes) already exist in the government. For example, in the price statistics program, the PPIs classified by stage-of-process indexes reflect a market-oriented measure, while the PPIRs represent the kind of price index that would be used to deflate shipments or outputs and measure productivity. I also thought that the market-oriented system would fit better into the harmonized system being used internationally to collect trade statistics.

As I mentioned, the hierarchy issue is still undecided. One recommendation I have made, described more fully in an article in the November 1993 issue of the Survey of Current Business, is to break the large service sector, which as currently defined accounts for two-thirds of the economy, into two sectors. One part would be called "distribution networks" covering retail, wholesale, transportation, communication and other network suppliers. The other grouping would consist of the traditional kinds of services which tend to be labor intensive--such as personal and business services.

As if undertaking the revision of one classification was not enough, the researchers of the federal statistical system have assumed yet another undertaking--a clean-slate look at the way we classify occupations. This effort was lead by the Bureau of Labor Statistics which in June 1993 also convened an international conference. I gave a paper at that one, too. It stressed the need to define the unit of observation--the job as I see it; then to develop an underlying concept for the grouping of jobs; and finally, a hierarchical framework within which those groupings can be aggregated. Among my conclusions in that paper was that occupational classification would serve more purposes if it could be thought of in a three-dimensional context. Jobs should first be aggregated by both type and skill level. The third dimension, though not as well defined, could be along the line of whether the job involves symbolic logical work, production process work, or in-place personal service, a classification scheme developed by Robert Reich in his book, The Work of Nations. Perhaps information, goods, and services would be another way to view such a classification at higher levels of aggregation. Perhaps, this third dimension would capture, a classification index, which in concert with the other two dimensions, would approximate how employers view or define the labor markets in which they buy factors of production. In any event, if we move in that direction, our occupational classification would resemble a three-dimensional matrix, a Rubik's cube. That would facilitate not only the analysis of markets for certain kinds of occupations, but also provide a reading on the skill level required for those occupations and the kinds of training that individuals might need to reach that skill level.

In closing, I think these two efforts to completely revise the SIC and SOC are major statistical developments with considerable impacts. I am most pleased to see U.S. government statisticians take the lead in achieving progress in these fundamental areas.

Comments on Economic Classification Revisions

Joe Matthey ¹

May 24, 1994

¹The author is on the staff of the Board of Governors of the Federal Reserve System and is a research associate at the Center for Economic Studies (CES), U.S. Bureau of the Census. These remarks are for presentation at the seminar on classification sponsored by the Council of Professional Associations on Federal Statistics, May 25, 1994, in Bethesda, Maryland. The comments reflect the author's own views, not the official views of the Federal Reserve System or Census Bureau.

My remarks will be devoted to the industrial classification revision plans discussed by Carole Ambler in her presentation of Jack Triplett's paper. My background in this area stems both from my experience using the existing classification system—in, for example, analyzing productivity developments by industry—and from my work with plant-level data as a researcher at the Census Bureau's Center for Economic Studies.

Triplett's paper addresses six issues. First, he argues that the United States has mounted this effort to revise the classification system because users demand it; for example, users find the existing classification system inadequate for productivity studies. Second, Triplett explains that this revision of the SIC, unlike those in the past, is not going to be riddled with compromises among competing uses. The third section explains why a production-oriented concept has been chosen, and the fourth section discusses what could be done to appease those most interested in the classification concept that ran a close second, the market-oriented concept that groups products according to their degree of substitutability. The fifth section argues that a conceptually-based classification system is practical, and the final section advocates that the research approaches of the Economic Classification Policy Committee be applied to the classification systems of other countries.

I would like to elaborate on several of these issues. First, from the perspective of productivity studies in manufacturing, I believe that the need for an improved classification system largely arises from the difficulties we have in implementing the existing system consistently over time and across surveys. For example, the four-digit SIC classification of an individual manufacturing establishment often differs depending on whether the code has been assigned on the basis of product detail collected by the Census Bureau or on the basis of information available to others who initially identify the birth of new establishments, such as the Social Security Administration, Bureau of Labor Statistics or the IRS. Moreover, even within the Census Bureau's SIC assignment system, the industry affiliation of multi-product plants can switch frequently over time. These classification difficulties (and deterioration in sampling frames for a broader range of reasons) cause published individual industry-level time-series to change too abruptly from year-to-year. Users often cope by modelling productivity at more aggregate levels, following the SIC hierarchy for the aggregation. But the current SIC hierarchy was not designed to preserve similarity of input structures upon aggregation, and the resulting aggregate analyses often do not make much sense. Thus,

for productivity analysis there is a demand for an improved SIC system in two respects; we will be much better off if the new system achieves greater continuity at detailed levels (over time and across data sources) and if the hierarchy for aggregation better preserves similarities of the production process upon aggregation.

Whether this consistency goal will be achieved ultimately is an empirical matter. The production-oriented unifying concepts of the revised classification system offer some promise of greater consistency, particularly if the process of how things are made in a given establishment tends to be more fixed than what an establishment makes. In other words, there is some hope that the new technologically based classification system will reduce the extent of SIC switching because it is easier for, say, a manufacturing plant to alter its product mix among goods that are not close substitutes than it is for that plant to change the basic manufacturing process. Ultimately, then, a fundamental task of classification is to find meaningful characteristics of establishments that are relatively fixed.

As an economist, any discussion of fixed factors of production automatically evokes images of the capital stock in place and also, to a certain extent, the human capital embodied in a firm's employees. The production orientation favored by the committee seems quite natural. My only advice is that when production processes are analyzed, particular attention should be paid to the fixity of the elements when deciding whether they are defining features of the industry.

Triplett discusses how the committee would be likely to proceed in determining the defining features of the industry in the section of his paper on whether a conceptually-based classification system is practical. He mentions three studies that demonstrate how classification decisions could be made. Two of these are "matrix papers" that offer subjective descriptions of the extent to which the existing classification systems in the United States and Canada fit the production orientation. A third paper presents a quantitative heterogeneity index for use as a diagnostic tool.

I have had the opportunity to read drafts of these papers. The overall impression that they leave is that a lot of work remains to be done, particularly on achieving a consensus on the defining features of industries. The Canadian paper puts it well in saying: "In the United States, the E.C.P.C. has analyzed part of the SIC. There is an official concordance between the two classifications, so the results could be compared for similarly defined in-

dustries. The initial comparisons showed numerous differences in the way the two countries had applied the concepts. (p. 14)" Given this illustration that in a subjective process of assigning characteristics to industries, experts will differ in their application of concepts, they conclude that it would be desirable to have objective measures.

My final remarks concern the one objective measure of heterogeneity that has received the most attention, the heterogeneity index originally proposed by Frank Gollop in 1986. The second report of the Classification Committee presents this heterogeneity index for selected manufacturing industries. The basic process of using the index for classification starts with a tentative grouping of establishments into various industries. Then, for each industry, a weighted average of the differences in input cost shares among establishments in the tentative industry is computed. The relative sizes of the establishments in terms of, say, shipments, can be used as weights.

Thus far, this index has been calculated using only ten types of inputs, each of which is a very aggregate concept: production workers, other labor, fuel, electricity, purchased services, agricultural materials, mineral inputs, nondurable materials, durable materials and capital. I must confess that when inputs are defined at such an aggregative level, I find the heterogeneity index relatively useless for classification. To see this, one can contrast the results of the heterogeneity index for the fluid milk industry with the subjective process illustrated in the U.S. matrix paper. The latter paper states that "...the physical properties of fluid milk dictate many of the processing methods and the types of machinery and equipment that must be used to handle it (p. 11)." I interpret this as meaning that if a plant has the types of machinery and equipment specially designed for handling fluid milk, than it must be a milk processing plant. In contrast, the quantitative heterogeneity index just looks at the overall cost of capital among plants within the industry, without regard to the type of capital equipment. Similarly, the heterogeneity index as computed just looks at the overall cost of agricultural materials, whether or not these materials have anything to do with the defining features of milk production.

In the case of capital equipment and structures the use of the aggregative data can be defended on the grounds that detailed information is not available. However, detailed information on materials use is available from the Census of Manufactures. In some of my own work with the plant-level microdata, I have gone to the opposite extreme, singling out specific detailed

materials that comprise a large fraction of total materials costs in the industry in question. For example, I analyzed the degree of heterogeneity in the use by fluid milk processors of whole milk from dairy farms.

My own statistical work demonstrates some of the difficulties one encounters when attempting to develop a quantitative index of heterogeneity, and whether or not the Gollop index can be successfully applied depends on how these issues are resolved. For example, not all plants report data on specific materials use. Small plants, in particular, omit information on detailed materials use because the Census forms instruct them to do so if a minimum value threshold is not surpassed. Moreover, in any given industry, the inquiries on specific materials are restricted to only a few pre-selected materials. Which Census form a plant receives depends on the tentative classification of the plant. So, anyone trying to develop a quantitative index of heterogeneity for re-classifying plants faces the problem that the data needed to make such a reclassification might not be collected, exactly because the initial classification was inappropriate.

In summary, the revised classification system has the potential for helping users of the data quite a bit, particularly those interested in production function relationships. However, it seems like a lot of work remains to be done to develop the consistency needed to achieve this benefit.

Session 2

DISCLOSURE LIMITATION METHODOLOGY

RESTRICTED DATA VERSUS RESTRICTED ACCESS:
A PERSPECTIVE FROM
PRIVATE LIVES AND PUBLIC POLICIES

George T. Duncan
H. John Heinz III School of Public Policy and Management
Carnegie Mellon University
Pittsburgh, PA 15213
Phone/FAX: (412) 268-2172/7036
email: George.Duncan@cmu.edu

1994 July 8

A paper presented to the Council of Professional Associations on Federal Statistics Seminar on New Directions in Statistical Methodology, Bethesda, MD, 1994 May 25-26. This paper draws directly on Duncan, G., Jabine, T., and de Wolf, V. (eds.) Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, the final report of the Panel on Confidentiality and Data Access of the National Research Council and the Social Science Research Council. Thanks are extended to the panel members for their many contributions to the report. Special thanks go to Thomas Jabine and Virginia de Wolf for both their contributions to the report and for thoughts on this paper. It is dedicated to the memory of Roger Herriot, who in his work in the federal statistical system demonstrated so clearly the value of innovative thinking.

RESTRICTED DATA VERSUS RESTRICTED ACCESS: A PERSPECTIVE FROM
"PRIVATE LIVES AND PUBLIC POLICIES"

George T. Duncan¹
Carnegie Mellon University

1. Stewardship of Statistical Agencies.

A statistical agency is more an art museum than a confessional booth. Certainly the three institutions are similar in eliciting valuables under pledges of protective stewardship—indeed both the survey respondent and the penitent entrust their personal information. But more consequentially, the statistical agency shares only with the art museum a commitment to responsible dissemination to the legitimately curious. Alike, the statistical agency and the art museum must address the tension between protection and access.

Long before statistical agencies had ever sponsored a survey to obtain personal facts, the cloak of confidentiality had been extended in a religious setting. In 1215, the Lateran IV Council decreed that "all the faithful, of both sexes, when they have reached the age of discretion, are to confess all their sins at least once a year to their own priest." (Bok 1983: 78) Traditionally, the received confession is treated as protected personal information, with the priest serving as an instrument of God. On the statistical front, it was not until 1890 that U.S. census legislation required census workers to swear under oath not to disclose census data except to their superiors. Likewise, art museums view protection of their treasured works as an essential function. Motivating the extension of protection by all three is a pragmatic footing: without assurances of security, each would be severely hampered in obtaining the largely voluntary contributions they require.

How does each institution protect its data? The priest silent to the curious is honorable. Contrarily, the art museum hidden to the inquisitive is ineffectual. Likewise, the statistical agency in secreting its data fails its mission.

¹This paper draws directly on Duncan, G., Jabine, T., and de Wolf, V. (eds.) Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics, the final report of the Panel on Confidentiality and Data Access of the National Research Council and the Social Science Research Council. Thanks are extended to the panel members for their many contributions to the report. Special thanks go to Thomas Jabine and Virginia de Wolf for both their contributions to the report and for thoughts on this paper. It is dedicated to the memory of Roger Herriot, who in his work in the federal statistical system demonstrated so clearly the value of innovative thinking.

Restricted Data versus Restricted Access

Whether for museums or statistical agencies, the dual role of protection and dissemination is challenging, but these two pillars cannot be compromised without risking institutional collapse. Original microdata as collected from statistical surveys can no more be provided to all who might want it than the new Andy Warhol museum in Pittsburgh could freely hand over one of his renderings of Marilyn Monroe.

Generically, two dissemination strategies are possible: provide the good in restricted form, i.e., as a transformation, to a quite general audience without preconditions on use, or provide access to the good itself, but only to a restricted audience under restricted conditions. For art museums, the first strategy calls for providing reproductions, while the second strategy calls for guarded galleries. For a statistical agency, the first strategy results in dissemination of restricted data. The second strategy results in restricted access. Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics (1993), the report of the National Research Council/Social Science Research Council Panel on Confidentiality and Data Access, explores these two strategies in its Chapter 6. The purpose of this paper is to provide some perspective on the ideas and recommendations of the report on these topics of restricted data and restricted access.

2. Restricted data

Restricted data is a confidentiality-motivated transformation of the original data; it results from the application of a statistical disclosure limitation technique. Before releasing a microdata file, for example, a statistical agency might go beyond removing explicit identifiers like name, address, and Social Security number. To limit disclosure risk, the agency could, for example, give people's ages in five-year intervals rather than by the exact date of birth.

Private Lives and Public Policies gives an overview of some key concepts and techniques of disclosure limitation:

- Disclosure risk, including identity, attribute, and inferential disclosure
- Statistical procedures for disclosure limitation, both for microdata and for tabular data
- Impact of improved computer and communications technology
- Recent research on disclosure limitation

A review and evaluation of statistical disclosure limitation techniques and their

Restricted Data versus Restricted Access

application is given in the Report on Statistical Disclosure Limitation Methodology (1994) and in Dalenius (1988) (also see, Fienberg 1993), so the treatment here will not be detailed.

Disclosure risk

As explored in Duncan and Lambert (1989), disclosure occurs when a data subject is identified from a released file (identity disclosure), sensitive information about a data subject is revealed through the released file (attribute disclosure), or released data makes it possible to infer the value of an attribute of a data subject more accurately than otherwise would have been possible (inferential disclosure).

Statistical procedures for disclosure limitation

Statistical disclosure limitation techniques involve transformations of data to limit the risk of disclosure. Use of such a technique is often called masking the data, because it is intended to hide characteristics of data subjects. Some statistical disclosure limitation techniques are designed for data accessed as tables (tabular data), some are designed for data accessed as records of individual data subjects (microdata), and some are designed for data accessed as computer databases. Common methods of masking tabular data are deleting table entries (cell suppression) and altering table entries (random error, or noise introduction). Common methods of masking microdata are deleting identifiers, dropping sensitive variables, releasing only a small fraction of the data records, and grouping data values into categories. Direct access of computer databases raises new statistical disclosure limitation issues which are only recently being addressed (see, e.g., Duncan and Mukherjee 1992; Keller-McNulty and Unger 1993).

In the case of a public-use microdata file, statistical disclosure limitation techniques can be classified into five broad categories (Duncan and Pearson, 1991):

1. Collecting or releasing only a sample of the data: For example, the Bureau of the Census first released a public-use microdata file with a 1-in-1000 sample from the 1960 Census of Population and Housing.
2. Including simulated data: This technique has not been implemented, but it is conceptually akin to including several identical limousines in a motorcade that is under threat of terrorist attack.

Restricted Data versus Restricted Access

3. "Blurring" of the data by grouping or adding error to the individual values: Presenting subjects' ages in 10-year intervals is an example of grouping. A microdata file prepared by the Census Bureau for the National Opinion Research Center from the 1980 census masked census tract characteristics (e.g., percentage of blacks, unemployment rate) by adding random noise (Kim 1990).
4. Excluding certain attributes: Information on a doctoral graduate field of specialization might be omitted.
5. Swapping of data by exchanging the values of certain variables between data subjects: The value of some sensitive variable could be exchanged for that in, say, an adjacent record.

For data released as tables, the blurring and swapping techniques described above have been used. Three other statistical disclosure limitation techniques are unique to tables (Cox 1980):

1. Requiring each marginal total of the table to have a minimum count of data subjects.^a
2. Using a "concentration" rule, also known as the (N, K)-rule, where N entities do not dominate K percent of a cell; for example, requiring that the reported aspects of two dominant businesses in a cell comprise no more than a certain percentage of a cell.
3. Using controlled rounding of table entries to perturb entries while maintaining various marginal totals.

Statistical disclosure limitation practices of federal statistical agencies

The practices of federal statistical agencies regarding statistical disclosure limitation is well-covered in Jabine (1993b), a paper commissioned by the Panel on Confidentiality and Data Access. Based on a detailed study of twelve statistical agencies, their basic finding is that, although most have standards, guidelines, or formal review mechanisms, there is great diversity in policies, procedures, and practices among them.

This finding provides the basis for the Panel's first recommendation in this area (all eight recommendations are given for convenience in the Appendix):

Recommendation 6.1. The Office of Management and Budget's Statistical Policy Office should continue to coordinate research work

Restricted Data versus Restricted Access

on statistical disclosure analysis and should disseminate the results of this work broadly among statistical agencies. Major statistical agencies should actively encourage and participate in scholarly statistical research in this area. Other agencies should keep abreast of current developments in the application of statistical disclosure limitation techniques.

Beginnings have been made in implementing this recommendation. In early 1992 the Statistical Policy Office convened an ad hoc interagency committee of ten persons to be chaired by Nancy Kirkendall of the U. S. Energy Information Administration. The mandate of the committee was to review and evaluate statistical disclosure limitation methods used by federal statistical agencies and to develop recommendations for their improvement. Subsequently, the ad hoc committee became the Subcommittee on Disclosure Limitation Methodology, operating under the auspices of the Federal Committee on Statistical Methodology. Its final product, the Report on Statistical Disclosure Limitation Methodology, notes, "the development and publication of this report is directly responsive to the CNSTAT Panel's Recommendation 6.1, which says, in part, that 'The Office of Management and Budget's Statistical Policy Office should continue to coordinate research work on statistical disclosure analysis and should disseminate the results of this work broadly among statistical agencies.'" In the report's Chapter VII, a research agenda is laid out for disclosure limitation methodology. A reasonable expectation is that further progress on dissemination will be made by the dissemination of the Subcommittee's report, the presentations at the Council of Professional Associations on Federal Statistics (COPAFS) Seminar on New Directions in Statistical Methodology, and publications in the OMB Statistical Policy Working Paper series.

The Panel was concerned with the impact of statistical disclosure limitation procedures on the quality of the data as it is disseminated to data users. Statistical disclosure methods can hide or distort relations among study variables and result in analyses that are incomplete or misleading. Because of this possibility, policy researchers have expressed serious reservations about the implementation of statistical disclosure limitation (e.g., Smith 1991). Further, data masked by some disclosure limitation methods can only be analyzed accurately by researchers who are highly sophisticated methodologically. Based on these findings, the panel made the following recommendation:

Recommendation 6.2. Statistical agencies should determine the impact on statistical analyses of the techniques they use to mask data. They should be sure that the masked data can be accurately analyzed by a range of typical researchers. If the data cannot be

Restricted Data versus Restricted Access

accurately analyzed using standard statistical software, the agency should make appropriate consulting and software available.

Unfortunately, this recommendation has yet to be addressed, or to appear on the research agenda of statistical agencies. The Report on Statistical Disclosure Limitation Methodology is moot on this topic.

Given the potential difficulties that certain statistical disclosure limitation techniques can cause for analysts, it is important that federal statistical agencies involve data users in selecting such procedures. As Greenberg (1991:375) notes, "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." These thoughts led to the Panel's third recommendation:

Recommendation 6.3. Each statistical agency should actively involve data users from outside the agency as statistical disclosure limitation techniques are developed and applied to data.

Steps toward implementation of this recommendation are being made through the inclusion of individuals outside the agency on microdata review panels. It remains to be seen whether the views of data users will be adequately represented.

Finally, over the past thirty years various agencies have released public-use microdata files successfully. Based on experience, such data dissemination has met a two-pronged test: (1) the microdata files have been useful to researchers and policy analysts and (2) confidentiality has been protected. Based on this finding, the panel made a final recommendation in this area:

Recommendation 6.4. Statistical agencies should continue widespread release, with minimal restrictions on use, of microdata sets with no less detail than currently provided.

Given an increased public concern over privacy and confidentiality issues,

Recommendation 6.4 presents a real challenge to statistical agencies. Far easier it would be to turn inward and protective. To do so, however, would be to abdicate the statistical agency's responsibility to provide the data a democratic society needs.

The panel noted that expansion of the number and richness of public-use microdata files to be disseminated would be better justified if all users were

Restricted Data versus Restricted Access

subject to sanctions for disclosure of information about individually identifiable data subjects. Reference was made to a recommendation, in another chapter, as follows:

Recommendation 5.3 There should be legal sanctions for all users, both external users and agency employees, who violate requirements to maintain the confidentiality of data.

3. Restricted access: Administrative procedures to protect confidentiality

Procedures for providing restricted access to data typically establish eligibility requirements for access and impose a variety of conditions governing the purposes for which the data can be used, which organizations and individuals can have access, the location of access, physical security measures, and the retention and disposition of initial and secondary data files.

Arrangements for providing restricted access to federal data for statistical purposes do exist. Jabine (1993a) provides 19 examples, including both interagency data sharing and arrangements with data users external to the federal government.

Interagency data sharing

There have been instances of agreements to permit interagency sharing of identifiable, or potentially identifiable, personal records for statistical purposes. Some of the instances involved transfers of administrative records; others involve transfers of data collected in statistical surveys. As identified in Private Lives and Public Policies, the mechanisms used to insure confidentiality in a selected set of instances included the following:

- Making data users in the receiving agency special sworn employees of the sharing agency
- Restricting further dissemination of data and follow up with respondents
- Periodic on-site inspections of the receiving agency's security measures by the sharing agency
- Regular review of the benefits of the sharing arrangement
- Written agreement that a specified data match would not be used for any other purpose and that the receiving agency would return the shared data

Restricted Data versus Restricted Access

file when the match was completed

- Minimizing the possibility of using linked data to identify an individual in a public-use file and then using the survey information in the identified individual's record for administrative purposes by data masking

In general, an obvious requirement for interagency data sharing is that the statutory requirements for confidentiality of all of the agencies involved must be observed. A second requirement is that the transfer of data among agencies must be consistent with statements made to data providers when the data were obtained from them.

Developing arrangements for interagency data sharing can be a complex and time-consuming process, especially if more than two agencies are involved or if novel applications of the data are planned. New initiatives are likely to pose new legal, ethical, administrative, and policy questions. The expected benefits in cost savings or better quality data must be substantial to justify the level of effort and perseverance needed to find acceptable answers. It helps if the proposed data-sharing arrangements offer benefits to all of the parties concerned.

The success of the instances examined in efficiently using data resources while protecting confidentiality support the panel's first recommendation regarding restricted access.

Recommendation 6.5. Federal statistical agencies should strive for a greater return on public investment in statistical programs through carefully controlled increases in interagency data sharing for statistical purposes and expanded availability of federal data sets to external users.

Full realization of this goal will require legislative changes, as discussed in Chapter 5 of Private Lives and Public Policies, but much can be accomplished within the framework of existing legislation.

External data users

The availability of high-speed computers and sophisticated analytic techniques and software have generated vastly increased appetites for federal statistical data. In many cases if the data are restricted sufficiently to ensure confidentiality, the released data will not satisfy the needs of users. Appropriate to such cases, several modes of restricted access for external data users have

Restricted Data versus Restricted Access

been developed by statistical agencies. Some of the important features of these access modes are eligibility criteria, location of access, cost and convenience for agencies and users, and methods of protecting confidentiality. Particular modes of restricted access include the following:

- Use of a fellows program with access at the agency's central facility, for a limited term, and only for projects that the host agency deems to be of interest
- Remote access to computer databases with automated screening of batch process programs
- Use of encrypted CD-ROM products which have statistical software that is restricted so as to prevent the user from obtaining unencrypted individual records or statistics that would tend to disclose individual information.
- Release of microdata under licensing agreements that provide for special sworn employee status, authorize unscheduled site visits to the data user, provide for prepublication review by the disseminating agency, and require return or destruction of the data when the research is completed.
- Ease on-site access of data users by providing access at agency regional centers.

Given this history and the value to society of broad dissemination of federal statistical data, the panel made the following two recommendations:

Recommendation 6.6. Statistical agencies, in their efforts to expand access for external data users, should follow a policy of responsible innovation. Whenever feasible, they should experiment with some of the newer restricted access techniques, with appropriate confidentiality safeguards and periodic reviews of the costs and benefits of each procedure.

Recommendation 6.7. In those instances in which controlled access at agency sites remains the only feasible alternative, statistical agencies should do all they can to make access conditions more affordable and acceptable to users, for example, by providing access at dispersed agency locations and providing adequate user support and access to computing facilities at reasonable cost.

Restricted Data versus Restricted Access

Finally the panel supported archiving of important statistical data:

Recommendation 6.8. Significant statistical data files, in their unrestricted form, should be deposited at the National Archives and eventually made available for historical research uses.

This recommendation is intended to cover statistical databases from censuses and surveys and those, like the Statistics of Income and the Continuous Work History Sample databases, that are derived from administrative records. The panel was purposely not specific as to the content of such archived databases and the length of time for which confidentiality restrictions should continue to apply. Some databases, like the economic and population censuses, might include explicit identification of data providers. Others, especially those based on samples, might not include names and addresses, but would not be subject to statistical disclosure limitation procedures of the kind that are applied to public-use microdata sets for contemporary use.

4. Conclusions

There is an inverse relationship between restrictions on data and restrictions on access: as data restrictions increase, fewer restrictions on access are needed and vice versa. A given level of confidentiality can be achieved with various combinations of restricted data and restricted access. Just as an art museum may sell reproductions, provide carefully monitored access to galleries, and allow qualified art historians considerable latitude in examination of a work, a statistical agency must choose an appropriate mix of data products to disseminate that will serve the needs of their various data users. A strong beginning has been made by the federal statistical system in developing a research and implementation agenda for restricted data. This is evident from the important contribution of the Report on Statistical Disclosure Limitation. I ponder the contribution that might be made through a comparable effort in developing a research and implementation agenda for restricted access. No less, I ponder the restricted data and restricted access procedures that will be required to ensure data access with confidentiality in the computer databases of the Global Information Infrastructure.

Restricted Data versus Restricted Access

REFERENCES

- Bok, S. (1983) Secrets: On the Ethics of Concealment and Revelation New York: Random House.
- Dalenius, T. (1988) Controlling Invasion of Privacy in Surveys Department of Development and Research, Statistics Sweden.
- Duncan, G. T., Jabine, T., and de Wolf, V. (1993) Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics Washington, DC: National Academy Press.
- Duncan, G. T. and Lambert, D. (1986) The risk of disclosure for microdata. Journal of Business and Economic Statistics 7(2):207-217.
- Duncan, G. T. and Mukherjee, S. (1992) Confidentiality protection in statistical databases: a disclosure limitation approach. Proceedings of the International Seminar on Statistical Confidentiality, International Statistical Institute.
- Fienberg, S. (1993) Conflicts between the needs for access to statistical information and demands for confidentiality. Technical Report #577, Department of Statistics, Carnegie Mellon University.
- Jabine, T. (1993a) Procedures for restricted data access. Journal of Official Statistics 9(2):537-589.
- Jabine, T. (1993b) Statistical disclosure limitation practices of United States statistical agencies. Journal of Official Statistics 9(2):427-454.
- Keller-McNulty, S. and Unger, E. (1993) Database systems: inferential security. Journal of Official Statistics, 9(2)475-499.
- Kim, J. (1990) Masking Microdata for National Opinion Research Center. Final Project Report. Bureau of the Census, Washington, D.C.
- Report on Statistical Disclosure Limitation Methodology (1994) Statistical Policy Office, Office of Management and Budget, Washington, DC.
- Smith, J. P. (1991) Data confidentiality: a researcher's perspective. American Statistical Association 1991 Proceedings of the Social Statistics Section. Alexandria, VA: American Statistical Association.

Restricted Data versus Restricted Access

APPENDIX. Recommendations

Recommendation 6.1. The Office of Management and Budget's Statistical Policy Office should continue to coordinate research work on statistical disclosure analysis and should disseminate the results of this work broadly among statistical agencies. Major statistical agencies should actively encourage and participate in scholarly statistical research in this area. Other agencies should keep abreast of current developments in the application of statistical disclosure limitation techniques.

Recommendation 6.2. Statistical agencies should determine the impact on statistical analyses of the techniques they use to mask data. They should be sure that the masked data can be accurately analyzed by a range of typical researchers. If the data cannot be accurately analyzed using standard statistical software, the agency should make appropriate consulting and software available.

Recommendation 6.3. Each statistical agency should actively involve data users from outside the agency as statistical disclosure limitation techniques are developed and applied to data.

Recommendation 6.4. Statistical agencies should continue widespread release, with minimal restrictions on use, of microdata sets with no less detail than currently provided.

Recommendation 6.5. Federal statistical agencies should strive for a greater return on public investment in statistical programs through carefully controlled increases in interagency data sharing for statistical purposes and expanded availability of federal data sets to external users.

Recommendation 6.6. Statistical agencies, in their efforts to expand access for external data users, should follow a policy of responsible innovation. Whenever feasible, they should experiment with some of the newer restricted access techniques, with appropriate confidentiality safeguards and periodic reviews of the costs and benefits of each procedure.

Recommendation 6.7. In those instances in which controlled access at agency sites remains the only feasible alternative, statistical agencies should do all they can to make access conditions more affordable and acceptable to users, for example, by providing access at dispersed agency locations and providing adequate user support and access to computing facilities at reasonable cost.

Restricted Data versus Restricted Access

Recommendation 6.8. Significant statistical data files, in their unrestricted form, should be deposited at the National Archives and eventually made available for historical research uses.

Statistical Disclosure Limitation Methodology

by

Nancy J. Kirkendall, Energy Information Administration

Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology was released in May 1994. This working paper reflects the efforts of the Subcommittee on Disclosure Limitation Methodology of the Federal Committee on Statistical Methodology. I was the chair of the Subcommittee. The other members are William Arends, National Agricultural Statistics Service; Lawrence Cox, Environmental Protection Agency; Virginia de Wolf, Bureau of Labor Statistics; Arnold Gilbert, Bureau of Economic Analysis; Thomas Jabine, Committee on National Statistics; Mel Kollander, Environmental Protection Agency; Donald Marks, Department of Defense; Barry Nussbaum, Environmental Protection Agency; and Laura Zayatz, Bureau of the Census.

Working Paper 22 presents a basic introduction to statistical disclosure limitation, describes the methods used by 12 Federal Statistical Agencies, provides more detail on techniques used to protect tables and microdata, and discusses needed research. It also presents the Subcommittee's recommendations. The previous *Statistical Policy Working Paper* on the subject of disclosure limitation was *Statistical Policy Working Paper 2*, which was published in 1978. While *Working Paper 22* is an update of *Working Paper 2* in some sense, one of our primary purposes was to summarize and describe the current techniques which are used to protect data, and to make recommendations concerning what the subcommittee felt should be done. It is primarily intended to serve as a practitioner's handbook.

The purpose of this paper is to summarize the information and the recommendations made in *Working Paper 22*.

Disclosure Limitation

"Federal agencies and their contractors who release statistical tables or microdata files are often required by law or established policies to protect the confidentiality of individual information. This confidentiality requirement applies to releases of data to the general public; it can also apply to releases to other agencies or even to other units within the same agency. The required protection is achieved by the application of statistical disclosure limitation procedures whose purpose is to ensure that the risk of disclosing confidential information about identifiable persons, businesses, or other units will be very small."¹

The historical method of providing data to the public is via tables. Beginning in 1962 with the advent of the computer age, agencies also started releasing microdata files. In a microdata file, each record contains a set of variables that pertain to a single respondent. The variables relate to that respondent's reported values. However, there are no identifiers on the file, and the data may be disguised in some way to make sure they do not reveal the respondent's identity.

¹*Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, p. 1.

For our purposes there are two types of disclosure. **Identity disclosure**, occurs when a specific respondent can be identified from the data. Identity disclosure is particularly important to microdata files, and the solution is to limit or modify the identifying information on the file. **Attribute disclosure** occurs when confidential information about the respondent is revealed. This type of disclosure is particularly important to tables (where it is assumed that one might know if a person is represented in the table), and the solution is to make sure a sufficient number of respondents contribute to each cell in the table.

A distinction is also made between tables of **frequency data** and tables of **magnitude data**. A simple example illustrates the difference. Assume that a survey provides data on a person's profession, his salary, and the county in which he lives. Let us assume that in Franklin county, we had the following three respondents who reported that they were doctors.

**Example Cell in Profession x County table
{Doctors, Franklin county}.**

Number	Count	Salary
1	1	\$600,000
2	1	\$ 75,000
3	1	\$ 75,000
Total	3	\$750,000

With this example, if we publish the total for counts (3), we say we have count data. If we publish the percent of people surveyed who were doctors, we say we have frequency data. With frequency or count data every respondent contributes exactly the same amount to the cell, and methods of identifying sensitive cells depend only on the number of respondents contributing to a cell.

On the other hand, the salaries are called magnitude data. Here the respondent's contribution to the cell total depends on his reported value. Let us assume that the two doctors who are less well paid are local general practitioners, and the third is a heart surgeon who works in the city, but lives in Franklin County. Publishing the total salary would allow each of the local doctors to make a very good estimate for the salary of the heart surgeon. If they can estimate his salary "too closely", we would say that we have attribute disclosure. Thus, for tables of magnitude data, the method of determining sensitive cells depends on the values reported by each respondent.

In the next few sections of this paper, we will illustrate the methods used to protect data and present the Subcommittee's recommendations. Section 1 concerns tables of frequency or count data; Section 2 tables of magnitude data; and Section 3 microdata. Section 4 is a summary.

1.0 Tables of Frequency (Count) Data

A cell in a table of frequencies or counts is sensitive if there are too few respondents. The methods used to protect such cells include:

1. Collapse categories (combine rows or columns).
2. Suppression.
3. Controlled (random) rounding.
4. Confidentiality edit.

Both collapsing categories and suppression are widely used by Federal agencies, and have been for years. Random rounding and controlled rounding have not actually been used by Federal agencies. The confidentiality edit is a new method which was used to protect tables from the 1990 decennial Census.

Assume that cells are defined to be sensitive if they have three or fewer respondents. The following table is an example we will use to illustrate different ways of protecting the sensitive cells. The cells which are sensitive are printed in bold with an asterisk.

Table 1 -- Example -- with Disclosure

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	1*	3*	1*	20
B	20	10	10	15	55
C	3*	10	10	2*	25
D	12	14	7	2*	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

1.1 Combine categories

As noted above, one way of protecting the sensitive cells is to combine rows and/or columns. In the following table, the education levels are combined into two categories. Clearly, the result is that there are no sensitive cells. However, a lot of information is lost.

**Table 2 -- Example Without Disclosure
Protection Provided by Combining Rows or Columns**

Household Head Education Level			
County	Low/Medium	High/Very High	Total
A	16	4	20
B	30	25	55
C	13	12	25
D	26	9	35
Total	85	50	135

1.2 Suppression

The second method of providing protection is to simply withhold from publication the sensitive cells and a combination of other cells in each row and column so that it is not possible to derive the value of the sensitive cells by subtraction using the published marginal totals. Clearly, we need at least two suppressed cells in every row and column, but is that enough? The answer is no, and here is the counter example.

**Table 3 -- Example With Disclosure
Protection Not Provided By Suppression**

Household Head Education Level					
County	Low	Med	High	Very High	Total
A	15	S ₁	S ₂	S ₃	20
B	20	S ₄	S ₅	15	55
C	S ₆	10	10	S ₇	25
S	S ₈	14	7	S ₉	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

To show that this table still contains disclosures, consider the sum of row 1 and row 2 minus the sum of column 2 and column 3. This reduces to the following equation:

$$(15 + S_1 + S_2 + S_3) + (20 + S_4 + S_5 + 15) - (S_1 + S_4 + 10 + 7) - (S_2 + S_5 + 10 + 7) = 20 + 55 - 35 - 30$$

or

$$S_3 = 1$$

This illustrates that selection of cells for complementary suppression is not a trivial matter. Methods of linear programming are used to select the set of cells which are "optimal" in some sense and which protect the sensitive cells. The following table with suppressions does protect the sensitive cells.

Table 4 -- Example Without Disclosure Protection Provided by Suppression

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	S	S	S	20
B	20	10	10	15	55
C	S	S	10	S	25
D	S	14	S	S	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

This example leads to the first of our recommendations. When suppression is used to protect tabular data, whether frequency or magnitude data, the table with suppressions should be *audited*. Auditing involves applying a linear programming algorithm to calculate the largest value a suppressed cell can take and the smallest value it can take. If the largest value and the smallest value are equal, the cell total is revealed exactly. If they are "too close" then the cell value can be estimated "too closely".

1.3 Random Rounding or Controlled Rounding

With random or controlled rounding, each cell count is rounded using some base value. In the following example, the base value is 5. In this case each cell count can be written as $X = 5q + r$. For random rounding each cell is rounded at random. This cell would be rounded up with probability $r/5$, and down with probability $1-r/5$. The problem with this procedure is that tables do not add, as illustrated in the Table 5.

**Table 5 -- Example Without Disclosure
Protection Provided by Random Rounding**

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	0	0	0	20
B	20	10	10	15	55
C	5	10	10	0	25
D	15	15	10	0	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

Random rounding has been used by Statistics Canada and was used by the New Zealand Department of Statistics before they moved to controlled rounding. The New Zealand Department of Statistics moved to controlled rounding primarily because users complained that the randomly rounded tables did not add (George and Penny, 1987.)

Controlled rounding is like random rounding except that a linear programming method is used to impose the constraint that the table must add. Controlled rounding was a topic of research during the 1980's, and for two dimensional tables and most three dimensional tables current methods work very well. It was proposed for use with the 1990 decennial census (Greenberg, 1986), but has not yet been used by any Federal statistical agency. An example of our table protected with controlled rounding is presented below.

**Table 6 -- Example Without Disclosure
Protection Provided by Random Rounding**

Household Head Education Level

County	Low	Med	High	Very High	Total
A	15	0	5	0	20
B	20	10	10	15	55
C	5	10	10	0	25
D	10	15	5	5	35
Total	50	35	30	20	135

Source: Numbers taken from Cox, Johnson, McDonald, Nelson and Vazquez (1985). Titles, row and column heading are fictitious.

1.4 Confidentiality Edit

All of the above methods are applied to a specific table. If the table is changed in some way, or another table containing data from the same data file is constructed, another detailed analysis must follow to assure that consistent protection is applied.

The confidentiality edit is a new method which was developed at the U. S. Census Bureau and used to protect tables from the 1990 Census (Griffin, Navarro, and Flores-Baez, 1989). With this method the original microdata file is manipulated (much as it would be if it were going to be released for public use). After manipulation the microdata file can be used directly to make tables. Other tables made from the same manipulated microdata file will also be protected, and the protection will be consistent. The approach described below was used for the regular decennial Census data file (the 100 percent data file), it uses a microdata protection technique called "data swapping" or "switching" (Dalenius and Reiss, 1982).

To apply the confidentiality edit the following steps are applied.

1. Take a random sample of records from the microdata file;
2. Find a match with them in some other county, based on a set of key variables;
3. Swap all other variables on the matched records;
4. Make tables

After the confidentiality edit, our table might appear as below.

**Table 7 -- Example Without Disclosure
Protection Provided by Confidentiality Edit**

Household Head Education Level

County	Low	Med	High	Very High	Total
A	13	2	4	2	20
B	18	12	8	17	55
C	5	9	11	0	25
D	14	12	8	1	35
Total	50	35	30	20	135

The only disadvantage I have seen quoted is that the table does not look as if disclosure limitation has been applied.

1.5 Recommendation

While each of these methods has advantages and disadvantages, the Subcommittee was unable to determine which of these methods were preferable in terms of the level of protection applied, and the usefulness of the result. Our recommendation is that further research should be done to address this question, and the result widely disseminated.

2.0 Tables of Magnitude Data

For tables of magnitude data only two methods can be used to protect sensitive cells. They are combining categories, and suppression. Each has the same strengths and weaknesses as discussed above, and if suppression is used the table should be audited. For tables of magnitude data, the new question is how to identify sensitive cells?

We indicated above that the respondents' reported values are used. In fact, cells are identified as sensitive if a simple linear combination of respondent level data is positive. The linear equation is called a linear sensitivity rule and the coefficients depend on the specific rule used and the parameters chosen. There are three rules which are commonly used:

(n,k) rule -- a cell is sensitive if n respondents contribute k% or more to the cell total;

p-percent -- a cell is sensitive if the published total can be used to estimate any respondent's data more accurately than p-percent;

pq -- like the p-percent rule, but acknowledges that before data are published, common knowledge allows estimation of any respondents' data to within q percent ($q > p$).

Recommendations

The Subcommittee's recommendations for tables of magnitude data are:

1. Only subadditive linear sensitivity measures should be used to identify sensitive cells. Subadditivity is a mathematical property that assures that if two or more cells are not sensitive, then their sum (union) is not sensitive either. Fortunately, all three commonly used linear sensitivity rules are subadditive.
2. The committee prefers the p-percent or pq rules as providing more consistent protection.
3. Suppression or collapsing categories are the only accepted methods of protecting sensitive cells.
4. The parameter values used in practice should not be revealed.
5. Tables containing suppressions should be audited.

For tables of magnitude data research is needed into identifying summary statistics to publish as a replacement for a sensitive cell total. If it could be shown that the summary statistics do not reveal individual data, they could be used instead of suppression and provide users with more information.

3.0 Microdata

For tables, we have associated "disclosure" with the publication of "sensitive cells", and have justified a simple way to identify which cells are sensitive. Once that is done, several approaches have been used to protect the sensitive cells. Unfortunately, for microdata files there is no standard agreed to definition of what constitutes "disclosure", other than uniquely identifying an individual in a data file.

The following four common ways to protect microdata files are used by virtually every agency which releases microdata files.

1. Use only a sample of the population. (A sample protects an individual's data, because it is not generally known whether or not a particular individual is included in the file.)
2. Remove obvious identifiers (eg. name, address, social security number).
3. Limit geographic detail (detailed data about an individual from too small a geographic region increases the risk of identification.)
4. Top code, bottom code and/or recode continuous high visibility variables. (Recoding continuous variables essentially makes them discrete. The larger values are shown only as greater than some number, the smaller values are shown as less than some number, and the intermediate values are assigned to a range.)

Salary is an example of a high visibility continuous variable. It may take many different values, and either very large ones, very small ones, or very precise recording of the value may reveal a respondent's identity. (Like our highly paid heart surgeon.) Other ways of protecting microdata are also applied to high visibility continuous variables. They include:

5. Masking (add or multiply by random numbers);
6. Swapping or rank swapping (find two records which match on a selected set of variables and exchange (swap) the remaining variables);
7. Blank and impute for randomly selected records. (randomly select a set of records, eliminate specific reported variables and replace them by imputed values);
8. Blurring -- aggregate values across small groups of respondents. (find a group of respondents, average some of their variables, and replace the reported values by the average.)

Recommendations

The subcommittee could only make one fairly obvious recommendation for protecting microdata files.

Remove direct identifiers and limit other identifying information.

Research is needed into defining disclosure or an unacceptable likelihood of disclosure for microdata files. Another area of needed research is into the impact of disclosure limitation techniques on the usefulness of the resultant data file. The subcommittee believes that research into these topics was of the highest priority.

4.0 General Recommendations and Summary

In addition to the specific recommendations above, the subcommittee had the following general recommendations. Agencies should

1. Seek advice from respondents and data users. Respondents should be asked about variables they consider sensitive and those they do not consider sensitive. It would be better if agencies applied disclosure limitation methods only to variables considered sensitive by respondents. Data users should be offered the opportunity to comment on disclosure limitation methods. Agencies should use this information in selecting the disclosure limitation methods to use.
2. Centralize review of disclosure limited products within an agency. A centralized review of disclosure limited products assures consistency in the application of disclosure limitation within an agency. In addition, a centralized review provides greater assurance that the data are adequately protected.
3. Share software and methodology. Agencies need to help each other to assure consistency in practice, and to make more advanced methodology and software widely available.
4. Agencies which release the same or similar data sets should cooperate in the application of disclosure limitation to those data sets. If there is no coordination, it is more likely, for example, that cells selected for complementary suppression by one agency, might not be suppressed by the other agency. This would lead to disclosure.

This paper has provided an elementary description of statistical disclosure limitation methodology and the principle recommendations of the Subcommittee on Statistical Disclosure Limitation Methodology. *Working Paper 22* provides considerably more detail on statistical disclosure limitation methodology, agency practices and needed research. It also provides an extensive annotated bibliography. The Subcommittee hopes that you find the information useful.

References

- Cox, L. H., Johnson, B. McDonald, S. Nelson, D. and Vazquez, V. (1985) "Confidentiality Issues at the Census Bureau," Proceedings of the Bureau of the Census First Annual Research Conference, Bureau of the Census, Washington D.C. pp 199-218.
- Dalenius, T. and Reiss, S. P. (1982). "Data Swapping: A Technique for Disclosure Control". Journal of Statistical Planning and Inference, Vol 6, pp. 73-85.
- George, J. A. and Penny, R. N. (1987), "Initial Experience in Implementing Controlled Rounding for Confidentiality Control, Proceeding of the Bureau of Census Third Annual Research Conference, Bureau of the Census, Washington DC., pp 253-262.
- Greenberg, B. (1986), "Designing a Disclosure Avoidance Methodology for the 1990 Decennial Census," presented at the 1990 Census Data products Fall Conference, Arlington, VA.
- Griffin, R. A., Navarro, A. and Flores-Baez, L. (1989), "Disclosure Avoidance for the 1990 Census, Proceedings of the Section on Survey Research Methods, American Statistical Association, Alexandria, VA, pp. 516-521.

Discussion of Presentations on Statistical Disclosure Methodology¹

Stephen E. Fienberg²

1 Prologue

This past weekend my wife and I were attending a Bat Mitzvah and the daughter of our friends read her portion of the Torah from the *Book of Numbers* dealing with the census of the Israelites in the desert. As I listened to her, I read this passage from the bible again with special care with the hope of some divine inspiration for my discussion of the two papers presented today. Let me share with you what I learned about disclosure limitation.

First, the census seemed to be much easier to take than we have found to be the case in modern times in the United States. There is no mention of an undercount, differential or otherwise, although women and children were intentionally omitted from the count. It turns out that there were 603,550 Israelites aged 20 and above, and the bible gives various breakdowns of these totals, without any reference to or apparent concern for confidentiality.

¹Presented at "Seminar on New Directions in Statistical Methodology," sponsored by the Council of Professional Associations on Federal Statistics, Bethesda, MD, May 25-26, 1994

²Stephen E. Fienberg is Maurice Falk Professor of Statistics and Social Science at Carnegie Mellon University, Pittsburgh, PA, 15213. The preparation of this discussion was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada to York University.

Second, the 12 tribes were organized around the tabernacle and in that sense we could think of the tribes as corresponding to geographic areas. Part of the reported data goes down to subgroups whose order of magnitude is a few thousand. Clearly, this would not meet the Census Bureau requirement for the release of identification of geographic codes for microdata sets where the threshold is now 100,000.

Third, while the bible contains no cross-tabulations as we know them today, it does include considerable information that could be displayed in cross-classified form. But even the smallest numbers reported, e.g., the 273 for the number of first born of the Levites, would not seem to provide an example requiring cell suppression.

Fourth, the bible actually releases the names of several individuals who participated in the census, especially the names of a number of the tribal leaders and their sons. This suggests that the Israelites didn't have any hang-up about the issue of uniques in the population for the release of census data. The idea seems to be that there is the need to distinguish whether or not the release is in fact harmful. After all, everyone knew that Moses, Aaron, and a number of others were included in the census and what their demographic classifications were. Therefore, identifying them by name did not compromise them in any way.

Finally, there were few or no subsequent releases from the biblical census

so we don't have much evidence about how the Israelites would have treated concerns about confidentiality. We do know, however, that there was nothing corresponding to Title 13 in the Torah or in the commentaries such as the Talmud.

Having noted all of this in the way of prefatory remarks, let me now turn to the two papers in this session.

2 Duncan on *Private Lives and Public Policies*

George Duncan has summarized the major recommendations from Chapter 6 of *Private Lives and Public Policies*, a report issued by the NRC-SSRC panel he chaired as they relate to statistical procedures to protect confidentiality. His paper begins with a discussion of a 1215 Latern IV Council decree on confidentiality and quickly shifts to statistical agencies' dual role of protector and disseminator of data. He then takes up the panel's themes of restricted data (via some transformation) versus restricted access. To do this, he needs to define disclosure and, in keeping with the literature, discusses this at three levels: individual disclosure, attribute disclosure, and inferential disclosure, and he lists some standard techniques for providing restricted data to achieve disclosure avoidance. This material is a brief introduction to that which is covered in much greater detail in chapter II of the draft Federal Committee on Statistical Methodology Working Paper 22, *Statistical Disclosure Limitation Methodology*, described by Nancy Kirkendall. In my remarks I will focus

on the panel's recommendations regarding restricted data and those aspects of the topic dealt with in Working Paper 22.

Because of the great diversity in policies and practices of the statistical agencies (documented in the panel report and in chapter III of Working Paper 22) the panel recommended that OMB should continue to coordinate research work on disclosure limitation and disseminate the results widely. The existence of the Subcommittee on Disclosure Limitation Methodology of the Federal Committee on Statistical Methodology and its recently released working paper represent OMB's and the agencies' positive response. The panel's second recommendation relates to agency assessments of the impact of their own data disclosure limitation techniques and Working Paper 22 remains silent on the matter, a point to which I will return in a few moments.

A few years ago I argued that the statistical agencies in the U. S. clearly were using techniques that were too conservative, i.e., that they erred too much on the side of restricting data in order to ensure that guarantees of confidentiality are not compromised as opposed to increasing the extent and utility of released data. I was immediately challenged and I offered as evidence to support my proposition the total absence of anecdotes where, despite agency actions, confidentiality had been breached. Agencies must remember that they are only public protectors and not owners of the data and they need to involve users in the choice of disclosure avoidance procedures. This is the third of the panel's recommendations and this, according to Duncan, is in the process of implementation by a number of agencies. The panel's final

recommendation encouraged the continued widespread release of microdata sets.

I have watched the NRC/SSRC panel from conception through the completion of its report. While the four recommendations I have singled out here from Chapter 6 of the report sound much like apple pie and motherhood, they and the other recommendations of the panel are clearly designed to move the practice of statistical data disclosure forward and encourage the development of a statistical basis for confidentiality practices. I heartily recommend the report and its companion volume of technical commissioned papers which appeared as a special issue of the *Journal of Official Statistics* in the fall of 1993.

3 Kirkendall on *Statistical Disclosure Limitation Methodology*

Nancy Kirkendall has described some of the ideas and materials from Working Paper 22 of the Federal Committee on Statistical Methodology Subcommittee on Disclosure Limitation Methodology, an activity which she chaired. This working paper needs to be considered against the backdrop of an earlier Federal Committee on Statistical Methodology working paper on the topic issued in 1978. What we have here is a major update with considerable detail and an extensive annotated bibliography. Depending on how we approach the topic, we find both good news and bad.

First, the good news. Much has happened in the intervening 16 years. The earlier working paper was technically innovative and it served as a catalyst to the development of new disclosure limitation methodology, especially in such agencies as the Bureau of the Census, but also by those in universities such as George Duncan and my former colleague Diane Lambert, and by Tore Dalenius, my fellow discussant today. The new working paper documents many of these advances and the extent of the research developed is impressive. So too are the advances in the uses of disclosure limitation methodology by federal statistical agencies. The current agency practices, as described in chapter III of Working Paper 22, are far more advanced thanks both to the methodological developments and to attendant advances in computation. In these senses, the new working paper represents a major progress report on the health of the federal statistical system.

Next, the bad news. I found the new working paper disappointing, largely because it represents an intellectual backsliding from the innovative stance staked out by its predecessor and because of its failure to adopt what I would argue is a badly needed statistical foundation for the very methods whose cause it advances. Let me explain.

Chapter II of the report captures some the current discussions in the literature about the the definitions of disclosure, but it fails to build on Dalenius' statistical definition of disclosure that formed the foundation for the structure of the 1978 paper. As a consequence, we have descriptions of methodology

such as cell suppression which, while seemingly advanced, represent mathematics but not statistics. The techniques have been honed so that they can be implemented for large collections of cross-classifications utilizing linear programming and other techniques but we are never told, either by those who developed the approach or by the Subcommittee preparing this working paper, what statistical criteria the methods attempt to optimize and the extent to which they succeed. Thus we are told, for example, about the need to keep the values of n and p in the cell suppression rules confidential, but there is no recognition that statistical learning by those outside the agency might easily make such a statement essentially moot. Similarly, in the discussion of three-way and multiway cross-classifications, there is no recognition of relevant statistical methodology that might inform the very methods under discussion such as the probabilistic theory for Fréchet bounds on cell values (e.g., see Kwerel, 1983). When we get to the discussion of research issues relating to cell suppression, we find more of the same: advances in optimization of network flow methods, more elaborate computer programs, faster software. Where is the statistics in statistical disclosure limitation methodology? Where is the recognition that the data collected by statistical agencies is not error free? I contend that this very measurement error ultimately drives the statistical properties of attempts to compromise otherwise confidential data and disclosure limitation methodology to counter such attempts.

Nancy Kirkendall presented an example of an application of cell suppression which produces through complementary suppressions the following table (in which S stands for a suppressed cell and x a released cell):

S_1	x	x	S_2	S_3
S_4	x	x	S_5	x
x	S_6	x	x	S_7
x	S_8	x	x	S_9

She uses this to illustrate the need for auditing tables prior to release since the cell with entry S_3 can be determined via the other cells. It is interesting to note that all of this is related to the theory of existence of maximum likelihood estimates under quasi-independence for two-way tables. (e.g. see Chapter 5 of Bishop, Fienberg, and Holland, 1975). That those developing methods in this area seem unaware of such links to the statistical literature serves to reinforce my point on the need to make statistical disclosure methods more statistical.

I have a similar reaction to the briefer materials described in the Working Paper on data swapping, especially as it was implemented in the 1990 decennial census. This method grew out of a novel notion suggested by Tore Dalenius, but there appears to be little recognition by those who implemented the approach regarding the effect that the method has had on the utility of the resulting data, for example, as it is to be used for enforcement of the

Voting Rights Act. I understand that considerable effort went into some of these considerations in advance, but we have little documentation and no post-censal evaluations.

The Working Paper also places what I believe to be a misguided emphasis on "population uniques." As various authors have noted, uniqueness in the population is a necessary but not sufficient condition for identity disclosure, and there is no reason to believe that identity disclosure necessarily compromises confidentiality guarantees. My example of the identity release in the biblical census I believe makes this point well. The Working Paper relegates the more interesting and more important statistical problems of inferential disclosure and measuring disclosure risk to the research agenda.

Finally, the report tries to make a clear demarcation between methods for microdata and methods for tabulations. What it fails to recognize is that many examples of tabulations are in fact restricted microdata. For example, tables of counts are microdata in which either the original variables are categorical or are continuous but have been disguised through the use of conversion through categories, and where the data have been truncated by the dropping of variables. Surely there should be some linkage between the methods for microdata and for tabulations. This is less a criticism of Working Paper 22 than it is of the state of the art of research on disclosure limitation. (See the related remarks in Fienberg, 1994).

There are interesting statistical ideas and proposals for a unified theory

of disclosure control in the research literature, such as those captured by the papers by Fuller, Lambert, Little, and Rubin in the recent special issue of JOS on privacy and confidentiality, but these are not given appropriate coverage in the Working Paper nor are they reflected in agency thinking. Perhaps this simply reflects the lag between research and practical implementation. Despite such shortcomings, Working Paper 22 is an excellent summary both of current methods and practices in the agencies. The Subcommittee should be applauded for its efforts.

4 Restricted Access or Expanded Access?

George Duncan's second major topic was the NRC/SSRC panel's recommendations on administrative procedures to protect confidentiality. The panel has emphasized the role of interagency data sharing as well as technological aids that facilitate such access. While the need for such restricted access clearly will continue, I believe that the future will be one of expanded rather than restricted access. Working Paper 22 is especially helpful in this regard. Chapter V on "Methods for Public-use Microdata Files" provides a concise primer on the developments in this area.

I've mentioned the role of technology in restricted access, but technology is even more important when we come to expand access. A number of federal statistical agencies are playing leadership roles in this regard. Nancy Kirkendall referred to the innovative approach being explored by the National

Center for Educational Statistics, but there are many other examples. For example, micro-data from the 1990 decennial census are currently available over the Internet via the Consortium for International Earth Science Information Network (CIESIN) in Michigan. Further, the Bureau of the Census has created SIPP-On-Call, a new interactive approach to allow access to files from the Survey of Income and Program Participation over the Internet. Special user-friendly access is available via Gopher or NSF's Mosaic. Even *Wired* magazine, in its June 1994 issue, describes such access to its readers and points out that one also has on-line access to the Privacy Act and Title 13 as hypertext documents!

The new world of immediate user and intruder access over the "information highway" will place greater demands on released microdata and it will test, in new ways, the appropriateness of disclosure limitation methods both for the preservation of confidentiality and for the increased utility of the released data. This, I predict, will be a major topic for the next Federal Committee Subcommittee effort in this area and I expect that new statistical approaches to disclosure limitation will accompany these emerging changes.

5 Summary

There is much meat for statistical thought in *Private Lives and Public Policies*, the report of the NRC/SSRC panel, and in both the original Working Paper No. 2 and the recently released Working Paper No. 22, *Statistical Disclosure Limitation Methodology*, produced under the sponsorship of the

Federal Committee on Statistical Methodology. I fully expect that the next COPAFS-sponsored seminar on new statistical methodology, will highlight new advances in this area that build on the substantial contributions to date, that will also better link to statistical ideas, and that will report on the enhanced utility of released data resulting from these new developments.

6 References

- Bishop, Y. M. M., Fienberg, S. E., and Holland, P. N. (1975). *Discrete Multivariate Analysis: Theory and Practice* MIT Press: Cambridge, MA.
- Fienberg, S. E. (1994). Conflicts between the needs for access to statistical information and demands for confidentiality. *Journal of Official Statistics*, 10, in press.
- Kwerel, S. M. (1983). Fréchet bounds. In *Encyclopedia of Statistical Sciences*. Vol. 3, (S. Kotz and N. L. Johnson, eds.), Wiley: New York, 202-209.

Tore Dalenius
DISCUSSION

Introduction

Before around 1970, the main direction of the methodological development was on the development of survey designs enhancing the efficiency, i.e. increasing the amount of information provided by a survey by other means than increasing the size of the survey.

Around 1970, a decisive change may be observed. The attention of the survey statisticians was now gradually directed towards how to recognize and hopefully address the problem of invasion of privacy. To address that problem, it proved necessary to apply methods which in fact served to reduce the amount of information made available. The subject of this meeting – Disclosure Limitation Methodology – reflects the above-mentioned change of methodological direction.

Dr. Duncan's presentation is based on ch. 6 of the book "Private Lives and Public Policies". This chapter considers two main options for protection of the confidentiality of released data: providing 'restricted data' and providing 'restricted access'. Dr. Kirkendall's presentation is based on the report Statistical Disclosure Limitation Methodology, prepared by a subcommittee of the Federal Committee on Statistical Methodology. This report is limited to disclosure limitation by means of 'restricted data'. Obviously, both documents are *final* products, a fact of relevance for the shaping of my discussion.

In what follows, I will first discuss selected aspects of restricted data and restricted access, respectively, to be followed by brief accounts of some additional aspects.

SELECTED ASPECTS OF RESTRICTED DATA

1. Two Classes of Data

Dr. Duncan and Dr. Kirkendall discuss in some detail two classes of data, viz. tabular data (frequency data and magnitude data), and microdata.

2. Frequency Data

The data to be restricted are represented by a table $T(N)$ with $R \times C$ cells. The restriction is achieved by a two-step procedure:

- i. the sensitive cells of $T(N)$, if any, are identified by subjecting the table to a threshold rule: cells with a small number of data subjects (such as $n = 3$) are considered sensitive;
- ii. next, some cells are combined, suppressed or rounded.

3. Magnitude Data

Typically these data are non-demographic, such as income or sales, accounted for by a table $T(X)$ with $R \times C$ cells. The variable X has in most cases a skew distribution: a small number of data subjects may account for a large proportion of the cell values. These cells may accordingly be sensitive, i.e. make it possible to link the cells with the data subjects accounted for, that is, to identify the data subjects. Hence, some kind of a restriction has to be applied to these cells.

The restriction of the data is achieved by:

- i. first identifying cells to which a small number of data subjects contribute a large percentage of the cell value - this may be done by using the p percent rule, the pq rule, or the (n, k) rule, also called the 'dominance rule';
- ii. next, these cells are subjected to restrictions, such as top-coding.

4. Microdata

Most releases of microdata are made up by a set of records with data about *individuals*. Only in exceptional cases do the data refer to business establishments.

Before the records can be released, formal identifiers must be removed ('deidentified'). But it may still be possible to link a record with a data subject: unique combinations of data concerning some attributes may serve as 'quasi-identifiers'. Hence additional restrictions are necessary, such as:

- i. sampling;
- ii. excluding data for one or more variables;
- iii. representing the data by broad classes; age may for example be represented by an interval (age class);
- iv. releasing data only for large populations; and
- v. confidentiality edit of the data,

to give but five examples.

SELECTED ASPECTS OF RESTRICTED ACCESS

5. A Wide Class of Procedures

Dr. Duncan includes in this class several disclosure limitation approaches. Common to them is that the statistical agency establishes eligibility requirements for the data users who are to be included in the group of users given access. I will briefly consider four procedures.

6. Interagency Data Sharing

This term is used to denote two cases:

- i. transfer of *administrative* data from a government agency to the statistical agency; and
- ii. transfer of *statistical* data from a government agency to the statistical agency.

7. Swearing In of Users

Formally, this kind of restricted access means that potential users are given status of employees of the statistical agency concerned, either at the main office, or at some local office near the place where the potential users live.

Clearly, the statistical office will have an opportunity of critically assessing the users' research projects and also the merits of the users.

8. Site Inspection

Assume that there is a government agency with authority to inspect how a statistical agency performs with respect to protection of the confidentiality of the data to be released. Then this "control agency" may implement a scheme for inspection of the performance of the statistical agency.

The scheme may call for inspection every k th month. A better scheme would, however, call for inspection at dates chosen at random. This would make it impossible for the statistical agency to perform well during an inspection but not between inspections.

SELECTED MISSING TOPICS

9. The Coverage of the Two Presentations

It goes without saying that it is possible to identify topics which have not been presented, or possibly only touched upon. I will provide three such examples.

10. Example No. 1 – Schemes for Rounding

Rounding the counts in a table may be carried out in several ways. The main ways are related to:

- i. the choice of a base different from the standard $b = 5$;
- ii. the simultaneous use of more than one base, especially if the table is large (many rows and columns);
- iii. rounding all cells in the table rather than a subset of cells; this type of scheme has in fact been proposed for use in the British population census; and
- iv. the use of deterministic rather than random rounding.

11. Example No. 2 – The Multi-Table Problem

Let T_1 be a table with no disclosure. And let T_2 be another similar table. Release of both T_1 and T_2 is not necessarily safe. Access to both tables may make it possible to derive a combined table T_3 which is disclosing.

12. Example No. 3 – Release by a Database

The statistical agencies should develop schemes for releasing statistics by means of a database. There is no reason to 'wait and see' what comes out with respect to a data superhighway.

TOPICS FOR RESEARCH AND DEVELOPMENT

13. Terminology

There is as yet no generally agreed upon terminology in the area under consideration here. It suffices to mention the following facts:

- i. privacy is defined in a great many different ways;

- ii. confidentiality is sometimes viewed as 'anonymity'; and
- iii. what in the two presentations is called 'disclosure limitation' was called 'disclosure avoidance' in the 1978 report; an alternative term is 'disclosure control', which I prefer.

It is indeed high time to develop a standard terminology.

14. A Catalogue of Potential Research Topics

In the report from the subcommittee there are some suggestions about research topics. But additional topics are needed. I will suggest one topic, viz. design of microdata about business establishments.

15. Inventory and Analysis of Sensitive Topics

In the last two decades, the non-response rate in surveys has shown a tendency to grow, possibly reflecting an increasing unwillingness to answer questions about sensitive topics.

In my view, the survey statisticians should process surveys already carried out and generate an *inventory* of sensitive topics which may explain the development. Such an inventory would be useful in the design of future surveys, by drawing the statisticians' attention to the need for special measures (such as special measurement methods) to improve the rate of cooperation.

The inventory should be *analyzed* to identify groups of data subjects with very large non-response rates. Such groups may then be singled out for special action.

CONCLUDING REMARKS

By way of presenting a summary of my views about the two presentations, I want to say that I have found them very informative and helpful. Dr. Duncan and Dr. Kirkendall are to be congratulated to the contributions they and their cooperators have made.

Session 3
CUSTOMER SURVEYS

Quality Management for Customer Satisfaction Surveys

Richard M. Devens, Jr.
Executive Editor
Monthly Labor Review
U.S. Bureau of Labor Statistics

Introduction.

This is the story of a customer satisfaction survey done for the Employment and Unemployment Statistics Quality Council at the Bureau of Labor Statistics. This is was the program's first customer satisfaction survey, and we are still learning from it. What I hope to pass on in this paper are the lessons learned about serving the "other customers," the executives that sponsored the survey and the front-line staff at the survey's focus. In other words, how "fit for use" was the National Survey of Users of Employment and Unemployment Statistics? Before plunging into that, I will take a few minutes to explore the importance of customer surveys and to outline the technical process of designing and conducting this one.

The reason for conducting any customer satisfaction survey is the position customers hold in the guiding principles of total quality management (TQM):

- customer focus,
- employee involvement,
- continuous improvement.

Customer focus, in my mind, is both the most important of these principles and the most difficult to persuade many public-sector managers to accept.

Continuous improvement is normally accepted straight off, usually with the assertion that the organization is already practicing it. The TQM purist might quibble that managers often mean their organization is always on the lookout for the big breakthrough, rather than practicing Deming's Fifth Point. [Improve constantly and forever every process for planning, production, and service.] My own observation has been that managers really do want to improve their operations, one way or another.

Employee involvement is a bit harder to sell. Many executives are used to and, quite frankly, happy with a command-and-control structure. In the case of the statistical agencies, such organizations were tremendously successful at organizing the armies of data collectors, mail room clerks, document controllers, coders, key entry workers, data reviewers, statistical assistants, statistical typists, junior economists, computer operators, computer programmers, research assistants, supervisory statisticians, senior economists, printers, and Assistant Commissioners that it took, and still takes to a fair extent, to produce a few tables of accurate, timely, relevant numbers.

Whether these hierarchies will work as well when data collection becomes automated, databases are connected through electronic data interchange (EDI), and performance becomes more dependent on the commitment of highly-skilled, self-confident, and very independent professionals is the issue. I believe such developments will lead organizations to embrace employee involvement models sooner rather than later--and most executives realize it, however grudgingly.

Customer focus, in contrast, is a very difficult concept for public-sector managers to accept at all, let alone embrace. The first reaction is, "We don't sell anything, so we don't have customers." Even after getting over this "filthy lucre" barrier, there is, especially in "craft" or "engineering" cultures such as those of the statistical bureaus, a deep skepticism about the fitness of the customer to make rational decisions or even to know what they want. These reactions are evident deep down into the structure of such agencies. Where the first-line will quickly accept the notion of getting involved in and taking greater responsibility for technical improvement, there is little enthusiasm for treating their work as a customer-satisfying process, not an estimates- or analysis-producing process.

The upshot of all this for the manager of a customer satisfaction survey is that there are two other--and perhaps more difficult --customers that must be considered in parallel with the external customer: the executive-level sponsors and the front-line staff. The rest of this paper overviews the National Survey of Users of Employment and Unemployment Statistics and its findings, the interaction of the project with its sponsors, the interaction of the project with the front-line staff, and the reactions of these "other" customers to the survey.

Outline of the National Survey of Users of Employment and Unemployment Statistics

The National Survey of Users of Employment and Unemployment Statistics is based on the premise that customer satisfaction is measured by the discrepancy between the client's needs and expectations and the client's perception of our performance. In the marketing literature, this is known as "disconfirmation" theory. The survey measures expectations and performance in five broad factors:

- **Data quality:** The accuracy, relevance, and timeliness of our statistics.
- **Tangibles:** The appearance and understandability of our materials.
- **Dependability:** Our demonstrated ability to perform promised services reliably, correctly, and promptly.
- **Assurance:** The knowledge of our employees and their ability to convey trust and confidence.
- **Empathy:** The caring, courteous, individualized attention we provide.

Each factor is represented by specific statements in the questionnaire. (See box.) The questionnaire also provides for an independent ranking of the importance of the factors and for general evaluations of satisfaction with our statistics and associated services.

Quality Factors and Their Proxies
(Question number in parentheses)

Data Quality

- (2) The demographic, geographic, and industrial coverage of the statistics is sufficient for my needs.
- (7) The data provided meet my standards of accuracy and reliability.
- (8) The data provided meet my standards of timeliness and currency.

Dependability

- (1) Staff are always available during their normal working hours.
- (4) My questions are answered promptly and dependably.
- (5) It is easy to get in touch with someone who can answer my questions.
- (14) The information I ask for is sent in the medium and format requested.

Tangibles

- (6) Materials provided make sense and can be understood without additional information.
- (14) The information I ask for is sent in the medium and format requested.

Assurance

- (9) Staff are knowledgeable and competent.
- (11) Staff can clearly explain conceptual and analytical issues without using overly technical language.
- (12) Staff can clearly explain the technical limitations of the data.

Empathy

- (3) Staff make me feel that I can call back for additional clarification or data.
- (4) My questions are answered promptly and dependably.
- (10) Staff are courteous.
- (13) Staff go out of the way to understand and fulfill my requests.
- (15) Apologies are rendered for inconveniences such as delay or misunderstanding of my needs.

Clients rated the statements on the quality of our statistics and services on 5-point scales for their expectations of quality and their perception of our performance. The expectation score is subtracted from the performance score to yield the "performance gap" for any specific statement. The performance gap for a factor is the mean gap for the set of statements that represent it.

In addition to the customer satisfaction scales, the survey asks how clients use our data, which programs they have utilized, and what channels of distribution were used to access data. We also provided space for comments.

Designing the Survey

We developed this user-friendly questionnaire using cognitive research methods including focus groups, think-aloud interviews, and a pilot test. Each of these methods identified errors and we were able to take corrective action before taking the final survey into the field.

In the field, Dillman's Total Design Method was followed closely, with the exception of experimental variations in the third and final follow-ups. Clients selected for the survey received several mailings:

- A notice arrived at the customer's address a few days before the primary questionnaire package.
- A thank-you/reminder letter followed the questionnaire by about a week.
- A second package went out two weeks after the "tickler."
- Final prompting, experimentally split between certified mail and telephone prompts, began 2 weeks after that.

This intensive data collection methodology yielded a usable response rate of 87.8 percent.

Two minor modifications to the Dillman method were necessary. First, the front cover of the questionnaire was not illustrated with graphics because of the limited space, and the stationery size was the ordinary 8 1/2 by 11. Second, the reminder/thank-you postcard was replaced by a reminder/thank-you letter because in-house constraints allowed letter production only.

The experimental exercise conducted in the third follow-up tested certain refinements to the Total Design Method for use in the establishment setting. Two weeks after the second follow-up, each of the remaining nonrespondents was randomly assigned to either certified mail follow-up or telephone prompting. In the control group, "holdouts" received the third follow-up packet by certified mail containing a replacement questionnaire, a business reply envelope, and a cover letter. The wording of the cover letter was different from the cover letters used in the preceding follow-ups; we softened and relaxed the wording but emphasized explanations of why this additional follow-up is important and is sent by certified-mail.

In the treatment group, nonrespondents were contacted by trained, experienced telephone prompters. We prepared a survey-specific training agenda, drawing on insights from nonresponse

conversion efforts in telephone follow-up surveys. The training included practice of scripted telephone procedures including appropriate reactions to specific reasons for refusal, discussion of persuasive techniques, and use of call record sheets. Approaches to locate the sample subject and find the best time to call back were also included in the training.

The survey's sample frame was constructed from two sources. First, client contact staff in the National program offices, the Regional Offices, and the Inquiries and Correspondence section logged contacts during September-November 1992. Program managers and senior executives provided separate lists of "regular" clients--persons maintaining on-going professional contact with our programs.

The lists were merged and duplicate entries removed. The resulting sample frame contained 3553 names which were stratified based on the program office that was the point of contact. Two additional strata were formed: one for all of the regular clients and another for all of the customers who were logged in by more than one program. The total sample of 999 clients was obtained by selecting samples of approximately equal size from all strata except one. Members of the stratum of regular users were included in the sample with certainty.

The response rate figures from the national survey of users of Employment and Unemployment Statistics are shown in Table 1. After the second replacement questionnaire mailout, the overall response rate had already reached 75%, which is the average overall response rate for TDM-based surveys. The third and final follow-up boosted the response rate by 13 percentage points to approximately 88%.

Table 1. Response and Conversion Rates

Mailing	Conversion Rate (%)	Overall Rate (%)	N*
Prenotice (Day 1)			999
1st mailout (Day 8)	28.65	28.65	998
Reminder (Day 15)	50.87	65.24	978
2nd mailout (Day 29)	28.53	75.77	970
3rd follow-up (Day 43)			
Certified mail	48.21		
Phone Prompting	44.44		
Overall	46.45	87.68	950
Close-out (Day 71)			

* The sample size declined as ineligible were uncovered through the data collection process.

Summary of Findings. Despite averaging 4.08 out of a possible 5 points on the performance scale of our survey, we did not fully meet our customers' expectations. (Expectations averaged 4.46 out of 5.)

Considering the major factors displayed on chart 1, our "performance gaps"--the average difference between our performance and our customers' expectations across the statements that represented the factors--were:

1.	Data quality	(-0.66)
2.	Dependability	(-0.52)
3.	Tangibles	(-0.34)
4.	Assurance	(-0.28)
5.	Empathy	(-0.17)

Using an expectations/performance grid --a "customer window" in the most recent jargon-- to analyze individual statements shows specific areas to concentrate our efforts on. (See chart 2.) In this graphic display, the intersection of the axes represents the grand means for customers' expectations (Y-axis) and their perception of our performance (X-axis). The points plotted for each statement are the ordered pair of Z-scores. According to this analytical tool, the important places to "Concentrate" corrective strategies are:

1. More timely (#8) and detailed (#2) data
2. Making it easier to find someone to answer your questions (#5)
3. Providing clearer materials (#6)

The survey report expressed these in terms of three strategic themes for improvement:

- **Get Faster:** Make statistical products available to the public more quickly.
- **Basic Service First Time/Full Service Every Time:** Have analysts able to answer broader ranges of inquiries, rather than transfer customers across program lines.
- **Clarity, Clarity, Clarity.** Make products and services easier to understand from the customer's point of view.

Chart 1

Expectation/Performance Gap Analysis

Overall

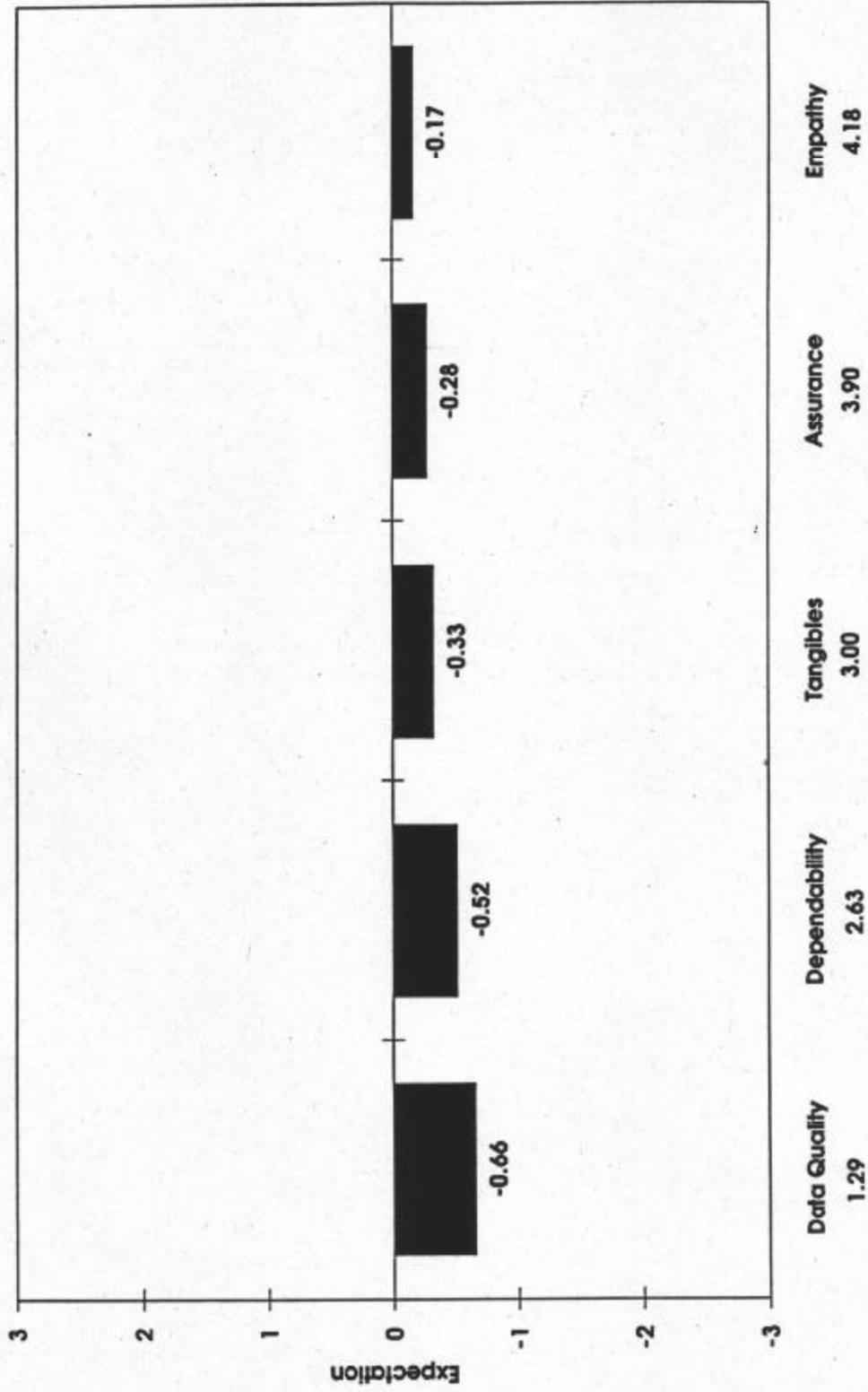
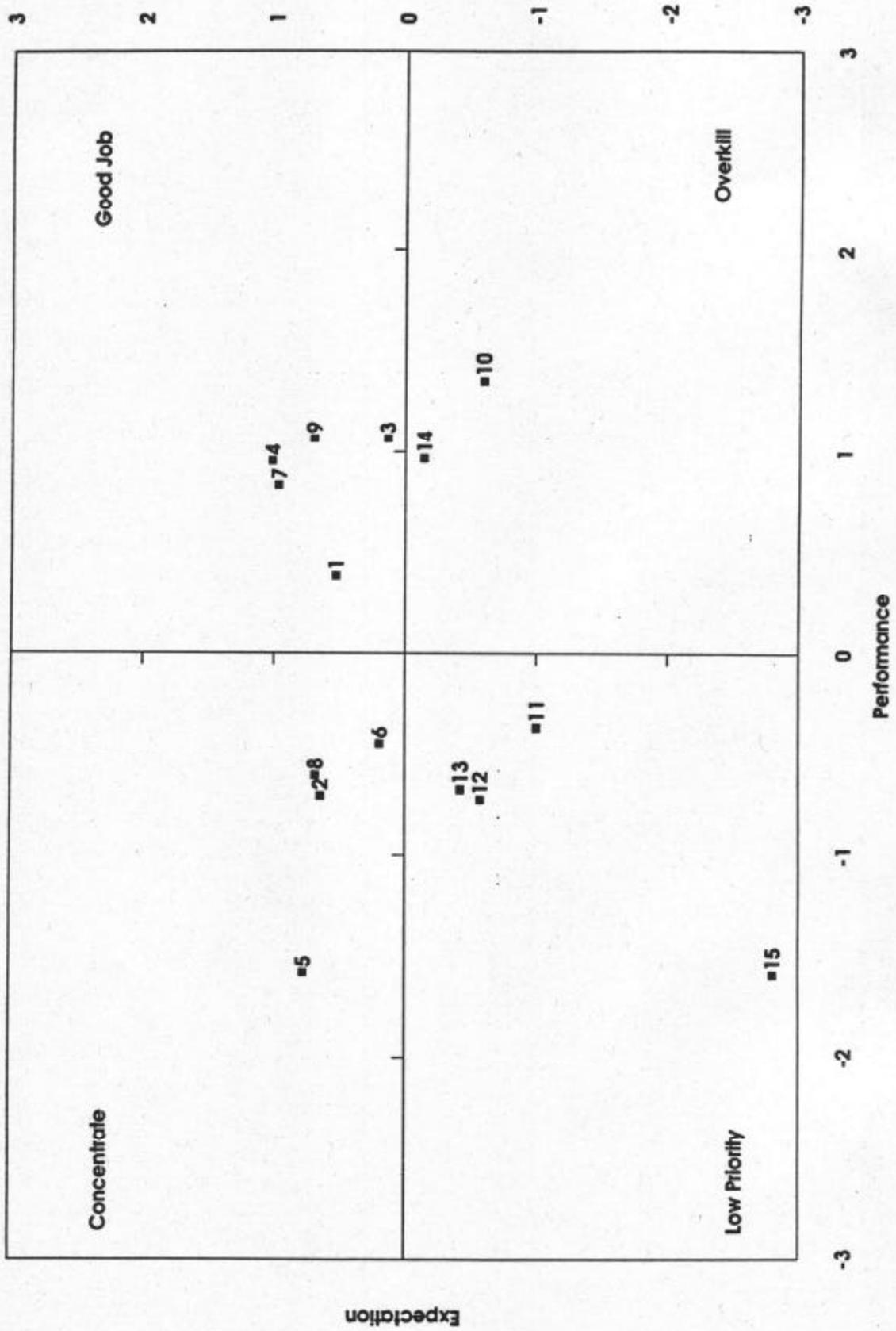


Chart 2

Expectation/Performance Comparison

Overall



Interaction with executive-level sponsors. The National Survey of Users of Employment and Unemployment Statistics is an "infrastructure" project on the part of the Employment and Unemployment Statistics (EUS) Quality Council. The Council itself is the executive-level coordinating body for the EUS Quality Improvement Program. The Council took its first look at conducting a customer survey at its August 1991 meeting. I was assigned the task of pulling together a few ideas on how such surveys were conducted and how they might apply to the EUS quality program. At their November 1991 meeting, the Council approved moving forward.

For the next several months, the minutes don't do the meetings justice. "...council members expressed concern," "...requested that a full proposal be prepared before proceeding," "...discussed the difficulties...." "After some discussion, it was agreed...." I am sure all of you are aware of what lurks behind these bland formulations.

After laboriously negotiating final approval, the survey team administered the instrument to the Quality Council itself, with the instruction, "Complete the questionnaire as if you were the 'average' customer." This exercise had the twin goals of further educating the Council on the survey and developing a baseline measure of the Council's awareness of customer needs and knowledge of the customers' perceptions of our performance.

As a baseline exercise, the Quality Council identified the same order of priority among the major quality factors that customers did. Interestingly enough, however, the absolute sizes of the perceived performance gaps were actually larger among Quality Council members than among customers. The primary source of the larger gaps, as shown below, was lower performance ratings by the Quality Council.

	Customers	Council
Average performance	4.08	3.80
Average expectation	4.46	4.36

How this exercise worked as an educational tool is a good question. My subjective evaluation is that the Council members themselves perceived it fairly narrowly in its baseline setting role, and would be surprised to find out about its covert objective of preparing them to more fully understand the survey's results.

When the final report of the project was drafted, the Quality Council was briefed on its contents and provided with copies for comment and approval. Once approval was obtained, final reports with more extensive technical documentation were published and circulated to the Council and the staff of the employment and unemployment statistics activity. The Business Research Advisory Council to the Bureau also expressed an interest in the survey and its results. Members of the Quality Council attended that briefing as well.

Interaction with customer-contact staff. Another set of customers for the survey and its results is the front line customer-contact staff. This is the group of our colleagues that provided roughly 3,500 customer names and addresses representing over 5,000 direct contacts over a 3-month period. These are also the people whose work product was put under scrutiny by the survey and the upon whom much of the burden of improvement would be likely to fall.

From the outset of the survey, five senior professionals from customer-contact units were assigned to the project. Their substantive contributions were critical and they also served as a "backchannel" of informal feedback between the staff and the survey team. That channel was, during the universe-building phase, our best means of helping the staff focus on keeping a complete log of contacts. (As a result of our debriefing of the representatives, one improvement we are likely to make in future surveys is a shorter log-in.) While the survey was in the field, the backchannel kept the staff informed about our progress.

Other interactions with the customer-contact staff included formal training sessions on the objectives, concepts, and methods of the survey and the procedures they would follow maintaining the universe log. In addition, the log procedures were documented on the forms themselves along with explanations of the purposes of the survey itself and of some of the most critical pieces of universe information--e.g., telephone numbers.

In general, interaction with this group of customers is something we should improve on. Some regional office information staffs had virtually no training or documentation of the survey or their role in it until what might be generously called the last minute. Interaction with the Inquiries and Correspondence Branch of the Office of Publications, while more timely, never reached the extent or intensity needed.

Reaction of sponsors. The most important measure of the success of a project such as the National Survey of Users of Employment and Unemployment Statistics is the action it prompts the organization to take. On this score the results are promising, but not overwhelming. Actions "moved onto higher priority time paths" to improve on the critical data quality factor include:

- Advancing the review and release of State-wide Local Area Unemployment Statistics (LAUS) data by 2-3 weeks (from a baseline of roughly 9 weeks after the reference period).
- Converting 20,000 late respondents to the Current Employment Statistics (CES) survey to automated self-reporting using an advanced touch-tone telephone data collection technology to improve the timeliness of these reports.
- Hosting the International Occupational Classification Conference to provide a forum for discussing new ideas and alternative approaches to the details of occupational categorization.
- Expanding service-sector detail for Current Employment Statistics by adding 108 new series to our most detailed publication and 20 seasonally-adjusted series to the employment news release.

In general, these have been projects that were on various burners to start with--the most the survey can claim is that some were completed more quickly and with more fanfare. There have been a few initiatives to start addressing some of the issues of service quality:

- Increasing the clarity of hard copy information sent to customers
- Resource book for information calls
- Developing new-employee training module for customer service
- Cross-program briefing on data availability.

My personal evaluation of the impact of the survey is that it was useful, but not nearly in proportion to the skills exercised or to the resources expended. That sense of disproportionality of effort leads me to the point of the paper--how well served are the "other" customers?

User-survey-users' surveys. To find out, I conducted a pair of informal surveys of the two groups of "other" customers. The survey of the Quality Council asked for their evaluation of the importance of the strategic directions the results pointed to, an evaluation of the communication processes between the Council and the project team, and an evaluation of the team's effectiveness at communicating the results. A similar survey was conducted among the front-line staff.

The results of the executive survey indicated that the group found that they rated the importance of the 3 strategic themes quite closely together between 5.3 and 5.8 on a scale of 1-to-7 (Not important at all to Extremely important). The highest score went to the Clarity, Clarity, Clarity theme.

The executives' evaluations of the effectiveness of our communication of the concepts and methods of the survey and of the results of the survey were devastatingly frank. On 5-point scales, the scores were 3.43 on effective communication of concepts (between "fairly" and "very" effective) and 3.57 on clear communication of results (again between "fairly" and "very"). These low scores, and remember they came from colleagues, probably reflect the reason the response to the survey was not overwhelming--the credibility of the product was not established and the results were not clearly communicated to the sponsors.

Reaction of front-line staff. The front-line's reaction to the survey is neatly summarized by the response rate to the survey included in the individual copies of the final report--almost 8 percent. Obviously, our efforts to engage this group fell short. The open question is why did it happen? I fear that the real reason was a fundamental failure to convince the front line that the customer satisfaction survey was serious. This may be the most significant quality issue for the National Survey of Users of Employment and Unemployment Statistics.

For what it is worth, the front-line survey found that among six respondents the importance ratings of the themes ranged from 4.7 to 5.7, with the clarity issue highest once again. The scores for effective, clear presentation of the concepts, methods and results of the customer survey--3.2 for effective presentation of objectives, concepts, and methods and 3.3 on clear communication of the results-- were even lower than those given by the sponsors.

Conclusions. Customer satisfaction is the "outcome" of any statistical or information service. This must often be measured quite separately from the "output" of programs. Output measures too often tell more about what is important to us than what is important to the customer. We in the Federal statistical community have always been concerned about hard measures of the output, "accurate data". We have only just now become aware that the soft outcome, "satisfaction with promptness and dependability of service," is perhaps even more important. That is why customer satisfaction surveys are useful--they are tools to measure and manage that outcome.

To be taken seriously as management tools, however, customer satisfaction surveys must be credible to the "other customers"--the people who should respond to the results. My point is that to obtain that outcome, the customer survey manager must establish credibility in advance and not think that good output--a clever report based on sound data--will suffice.

Note: All material in this article is solely the responsibility of the author. The views expressed here do not necessarily reflect the policy of the Bureau of Labor Statistics or the views of other BLS staff members.

**COMPARABILITY IN CUSTOMER SATISFACTION SURVEYS:
PRODUCTS, SERVICES, AND GOVERNMENT AGENCIES**

by

Michael D. Johnson

National Quality Research Center
The University of Michigan
School of Business Administration
Ann Arbor, MI 48109-1234

Phone: 313-764-1259

Fax: 313-763-9768

April, 1994

To appear in *New Directions in Statistical Methodology*, Washington, DC: Office of Management and Budget, forthcoming. Do not quote or reproduce without permission. The author thanks the Swedish Post Office, the National Quality Research Center at the University of Michigan's School of Business Administration, and the International Center for Studies of Quality and Productivity at the Stockholm School of Economics for providing the data described in the paper.

**COMPARABILITY IN CUSTOMER SATISFACTION SURVEYS:
PRODUCTS, SERVICES, AND GOVERNMENT AGENCIES**

ABSTRACT

This paper describes recent advances in customer satisfaction surveys and their implications for government agencies. Many agencies are in the process of implementing customer satisfaction monitoring systems and identifying appropriate private sector benchmarks. Satisfaction models and survey methods currently being used to produce national customer satisfaction indices are described. These efforts illustrate a number of important steps that should help government agencies produce meaningful measures of satisfaction and identify private sector industries that provide realistic agency benchmarks.

INTRODUCTION

Customer satisfaction has emerged as an important benchmark for gauging the performance of various economic agents over the past decade. Manufacturers of durable and nondurable products, retailers, service providers, utilities, and government agencies alike have implemented, or are in the process of implementing, customer satisfaction measurement systems. At a micro-level, these systems monitor a firm's or agency's primary asset - their customers - and provide important diagnostic information needed to improve or maintain satisfaction. At a macro-level, Sweden and Germany have implemented national customer satisfaction indices to monitor the major sectors of their economies while the United States, Taiwan, and New Zealand are in the process of doing the same.

There is disagreement, however, among psychologists, economists, consumer researchers, public policy makers, and others regarding the merits of comparing satisfaction across individual and industries. The ever broadening arena of customer satisfaction, in conjunction with recent advances in how satisfaction is surveyed and operationalized, shed light on this long-standing debate. The goal of this paper is to describe these developments and discuss their implications for government agencies. Many agencies now find themselves, for the first time, asking such questions as, "who are the customers served by our agency, who should be our customers, what standards should we use, and what comparable businesses should we benchmark on?"¹ Recent advances in satisfaction survey methods provide important insights into how government agencies should survey their customers. Sweden's experience with a national Customer

¹ Section 1 of President William J. Clinton's executive order, dated September 11, 1993, begins as follows:

In order to carry out the principles of the National Performance Review, the Federal Government must be customer-driven. The standard of quality for services provided to the public shall be: Customer service equal to the best in business. For the purposes of this order, "customer" shall mean an individual or entity who is directly served by a department or agency. "Best in business" shall mean the highest quality of service delivered to customers by private organizations providing a comparable or analogous service.

Satisfaction Barometer (the SCSB) also illustrates which private sectors businesses provide benchmarks or standards of comparison for these agencies. First, however, the nature of the debate over the comparability of satisfaction and the issues involved are described.

SATISFACTION AND THE HAPPY SLAVE PROBLEM

Customer satisfaction is a customer's evaluation of their overall experience with a product or service to date (Johnson and Fornell 1991; Johnson et al. 1994). This definition of satisfaction is consistent with existing views in economic psychology, where satisfaction is often equated with notions of subjective well-being (Van Raaij 1981), and economics, where satisfaction is equated with post-purchase consumption utility (Meeks 1984). Because it describes the customer's total consumption experience, satisfaction predicts customer loyalty and a firm's subsequent "profitability." In the private sector this "profit" is bottom line return on assets (Anderson et al. 1994). For government agencies, the benefits of increased customer satisfaction range from budget considerations, to more efficient use of taxpayer dollars, to the creation of a more positive image, to compliance (e.g., for the Internal Revenue Service).

There is a long standing debate in economics over the comparability of satisfaction across individuals and industries (see Hammond 1991 for a review and extensive bibliography). Bentham (1802) defended the comparability of satisfaction as both possible and necessary from a public policy standpoint, though not without error. Subsequent economic theorists sought to eradicate satisfaction measurement and comparisons as value laden and unnecessary (Hicks 1939; Robbins 1938). Recently, satisfaction has again emerged as a basis for making meaningful comparisons across people and products. Virtually all policy recommendations require comparisons of welfare which is proof enough that they are possible (Scitovsky 1951). The important question has become how comparisons of satisfaction or well-being are and should be made (Hammond 1991; Jorgenson 1990; Sen 1979; Simon 1974; Tinbergen 1991).

Our interest is specifically with customer satisfaction. Economic theorists are more often concerned with comparisons of more global economic well-being, which includes not only customer satisfaction but job satisfaction and income evaluation (Poiesz and Grumbkow 1988). Broad based comparisons of customer satisfaction are not exactly new. Andreasen and Best (1977) report meaningful comparisons of customer satisfaction and complaint behavior across a variety of product and service categories, while Pfaff (Lingoes and Pfaff 1972; Pfaff 1977) has used subjective measures of satisfaction to construct an index for the purpose of comparing various food product categories. Wikström (1983) has even compared subjective levels of customer satisfaction across countries (Sweden and the U.S.) and argued that the observed differences can be traced to underlying differences in market performance between the two countries.

Yet some policy researchers have concluded that subjective measures of customer satisfaction are incapable of revealing any meaningful differences (Hunt 1988; Ölander 1988). Most notably, Ölander (1977a, 1977b, 1988) argues persuasively that subjective measures of customer satisfaction are fraught with problems. Foremost among these is the so-called "happy slave" problem. Because customers adapt to the levels of product and service performance available to them, no meaningful differences in satisfaction should emerge. Individual differences in the degree of adaptation within and across industries further compounds the problem. Other problems include the notion that customers may have different yardsticks by which they judge satisfaction. Even if they used the same standards, consumers may have very different product or service alternatives available to them, and/or differ in their knowledge of these alternatives. Finally, customers may fail to express true dissatisfaction or strategically express false dissatisfaction in hopes of receiving some retribution.

At some level Ölander's concerns are very real, as when one might compare satisfaction between customers in relatively wealthy and impoverished countries. At the same time, several considerations suggest that the concerns expressed over comparing

subjective measures of customer satisfaction are overstated. As Pfaff (1977) argues, who is in the best position to evaluate customer satisfaction but customers themselves? People are more similar than different, and these similarities are growing in an increasingly "global" economy. We live in an information age in which consumers from different countries and socio-economic strata of our society are increasingly aware of alternative products and services available in the marketplace.

Equally if not more important are recent advances in satisfaction survey methods and modeling which facilitate our ability to compare subjective measures of satisfaction. Sweden's Customer Satisfaction Barometer (SCSB) embodies these advances and is serving as the prototype for the American Customer Satisfaction Index (ACSI).

THE SWEDISH CUSTOMER SATISFACTION BAROMETER

Established in 1989, the SCSB was the first national customer satisfaction index for domestically purchased and consumed products and services (Fornell 1992). The index is constructed using survey measures obtained from representative customers in each of 32 major Swedish industries which themselves represent approximately 70% of Sweden's gross domestic product (GDP). Those companies that account for approximately 70% of combined industry sales are selected to represent each industry. In cases where a company sells multiple products or services, the "flagship" brand (the product or service with the highest sales in kronor) is chosen to represent the company. For example, Saab Scania is represented by the sales of its 9000 series automobiles while banks are represented by their money lending activities. Each year approximately 100,000 customers are contacted by telephone and screened to obtain a sample that has experience with the products and services in the index. The number of customers who pass the experience screen and agree to participate is approximately 25,000 each year. Industry level sample sizes range from about 250 to over 4000 depending on the number of competitors.

Comparability in the Model

The first step in assuring comparability in the SCSB involves the choice of satisfaction related constructs and how they are modeled. Survey respondents are asked a variety of questions to operationalize five key constructs: (1) customer perceptions of product or service performance, (2) their expectations regarding performance, (3) customer satisfaction, (4) whether they have complained ("voice"), and (5) customer loyalty. The SCSB model, which is presented in Figure 1, posits six relationships among these variables. These relationships are relatively universal in that they cut across all of the products and services in the barometer and are described briefly here (for more extensive descriptions and discussion see Fornell 1992).

- insert Figure 1 here -

Satisfaction is posited to be a function of two antecedent variables, perceived performance or quality and customer's expectations regarding performance (Fornell and Johnson 1993; Johnson and Fornell 1991). Customer satisfaction should increase with the degree to which a product or service provides net benefits that customers value (i.e., perceived performance). Because expectations embody past quality or performance information, they too should positively affect satisfaction. Expectations serve to anchor overall evaluations of satisfaction in the vicinity of the expectations (Oliver 1977, 1980). The size of this anchoring effect depends on the relative strength of the expectations versus performance information (Johnson, Nader, and Fornell 1994). As experienced customers can predict, to some degree, what levels of performance they will receive, expectations should also show a positive relationship to perceived performance.

There are two primary behavioral consequences of satisfaction. Increased customer satisfaction should reduce the incidence of customer voice or complaining behavior. Satisfied customers are also loyal customers, which is the key to the satisfaction-profitability linkage (Anderson, Fornell and Lehmann 1994). Finally, voice may increase loyalty. The size of this relationship reflects the degree to which customers

are allowed to voice their complaints and a firm's ability to address these complaints. That is, the relationship is positive when a firm can turn a complaining customer into a loyal customer. Overall the relationships in Figure 1 are well supported and appear to generalize across Swedish industries (Fornell 1992).

Comparability in Satisfaction Survey Items

The next step in assuring the comparability of satisfaction in the SCSB is to use a survey instrument whose questions are themselves universally applicable and help control for industry differences. This is quite different from what typically occurs in the context of a particular product category or industry where perceived performance is operationalized using customer ratings of a product or service on quality dimensions or attributes that are idiosyncratic to the industry (e.g., attributes of an automobile). In the SCSB, performance is operationalized using two measures of perceived value, the customer's perception of quality received relative to the price or prices paid (benefits relative to costs) and their perception of the price or prices paid relative to quality received (costs relative to benefits). Research demonstrates that this "value" is a common denominator that consumers use to compare even very dissimilar or "noncomparable" products and services (Johnson 1984, 1989). Using value perceptions to measure performance also controls for differences in income and budget constraints across respondents (Hauser and Shugan 1983; Lancaster 1971) which allows us to compare very high and very low priced products and services.

Satisfaction is also surveyed using comparable items. These include the customer's rating of overall satisfaction, how well the product performs relative to an ideal product or service in the industry, and whether performance fall short of or exceeds customer expectations. Theoretically, all three of these rating should reflect the underlying level of satisfaction independent of the particular product, firm, or industry involved (Johnson 1994; Johnson, Anderson and Fornell 1994). Customer voice is, meanwhile, measured in two ways: the incidence of formal complaints to company or

agency managers, and the incidence of informal complaints to personnel or service providers. As for performance and satisfaction, both measures are flexible and apply to a variety of organizations.

Finally, customer loyalty is measured using questions regarding repurchase likelihood and sensitivity to price increases. While very applicable to competitive product and service industries, these loyalty measures are more problematic in the case of government agencies and monopolies. The solution used in the SCSB is to make the questions hypothetical. That is, assuming some other organization could provide an agency's services, how likely would you be to use the agency again and how much more would the agency have to "charge you" before you would switch to the hypothetical competitor? Over time, these questions are becoming less hypothetical and more realistic as government agencies are being reinvented and subjected to increased competition and market pressures.

Comparability in Satisfaction Model Estimation

The third step in assuring the comparability of subjective satisfaction centers on just how the survey items described above are used to operationalize the constructs and estimate the relationships in Figure 1. An important aspect of the Swedish index is that satisfaction (as well as performance, voice, and loyalty) is operationalized as a latent variable within a system of equations. Johnson and Fornell (1991) argue that satisfaction, as a theoretical concept, is a common denominator on which very different people and products may be compared. As a latent theoretical construct, satisfaction is empirically measurable as a weighted average of multiple satisfaction indicators. As Ölander and others have argued, any individual rating or measure which uses a particular yardstick is at best an indirect proxy for satisfaction. Operationalizing satisfaction as the shared variance among a set of multiple satisfaction survey measures provides a more direct measure of latent satisfaction.

This latent variable is estimated with a system of equations, or causal model framework, using the SCSB model in Figure 1. The particular estimation method used to operationalize latent satisfaction and estimate the model is partial least squares or PLS (Fornell 1989; Lohmöller 1989; Wold 1982). PLS is an iterative estimation procedure that corrects for routine least-squares measurement problems and does not impose distributional assumptions on the data. This is particularly attractive in a satisfaction context where distributions are often highly skewed. As a result, PLS is better suited to causal model estimations involving small samples than is, for example, covariance structure analysis using LISREL. It also allows the researcher to operationalize latent variable scores and hence calculate an index value.

Another important feature of PLS is that it aims to explain variances at the observed (measurement) level while LISREL aims to account for observed covariances. In Figure 1, latent satisfaction should ultimately explain variance in loyalty across customers. PLS weighs the individual satisfaction survey items in the satisfaction index so as to maximize the index's ability to explain loyalty. This, in turn, provides a satisfaction index that is comparable in the following sense. In each industry, the satisfaction index explains an endogenous, dependent variable that is universally applicable across industries. The satisfaction index is itself explained by two antecedents that should affect satisfaction in a similar fashion across industries. If the satisfaction index behaves as it should behave according to the model in Figure 1, then its validity and value as a benchmarking source is supported.

Empirical Evidence of Comparability

The ultimate test of the "happy slave" problem and other questions raised regarding the comparability of satisfaction is an empirical one. A recent study by the author and Claes Fornell (Fornell and Johnson 1993) using the SCSB data explicitly examines this issue. In the study we argue that if one can explain differences in satisfaction across industries using some underlying difference in the industries, then the

observed differences are meaningful. Specifically, we argue that product or service differentiation in an industry is one logical basis for explaining differences in expectations, performance, and satisfaction across industries.

Differentiation, in this context, refers to the availability of predictably different options to customer. Differentiated industries offer predictably different options that more directly meet the needs of a heterogeneous population of customers. In contrast, undifferentiated industries offer "no choice." The automobile industry in Sweden is, for example, highly differentiated. Customers can choose among a wide variety of options and are confident in their ability to evaluate differences among them. Police, telecommunications, and public postal services are, in contrast, relatively undifferentiated due to the lack of variety from which to choose. At a more intermediate level of differentiation are banks and insurance companies, where alternatives exist yet customers have difficulty judging their differences. The study found that the level of differentiation across the industries explained fifty-percent of the variance in aggregate perceived industry performance. This performance, in turn, explained over half of the variance in aggregate industry customer satisfaction.

This study has important implications for government agencies who must now benchmark their customers' satisfaction to that observed in private sector industries. In the past, public utilities, monopolies, and government agencies had no competitors on which to benchmark satisfaction levels. Because the industry level differences in satisfaction are meaningful, national indices such as the SCSB provide these agencies and firms with useful benchmarks. The Satisfaction Index scores for the Swedish industries are presented in Table 1. Government owned industries include the pharmacies, local police services, business post, public post, railroads, business telecommunications, public telecommunications, and state sponsored TV broadcasting. To illustrate the differences among industry types, the industries in Table 1 were grouped into three classes: (1) products and product retailers, (2) services, and (3) government owned agencies and

businesses. The average satisfaction indices for each of these three groups from 1989 to 1993 are plotted in Figure 2.

- insert Table 1 and Figure 2 here -

The figure illustrates several interesting points. First, following Fornell (1992; Fornell and Johnson 1993), products and product retailers show systematically higher levels of satisfaction than do competitive services and government owned agencies and businesses. Both of the latter groups are service-oriented, which makes it inherently more difficult to meet specific customer needs. While products meet customer needs largely through their physical means of production, the production of services involves more of the human resources of the firm and customers themselves. This creates greater heterogeneity, on average, in the production of services versus products and lower average performance (Fornell and Johnson 1993; Zeithaml et al. 1988). In Figure 2, products and product retailers show the highest satisfaction and it stays relatively stable over time. Competitive services are below the products and retailers, which is consistent with the nature of service production. The drops in service satisfaction in 1992 and 1993 are due primarily to the recent poor performance in the banking sectors. Finally, the government owned agencies and businesses are generally lowest in satisfaction. This is due both to their service orientation, which makes it difficult to provide consistent quality, and monopoly positions, which limits customer choice.

More important from a benchmarking standpoint is the steady increase in satisfaction for the government sector over the five years in which the index has been in operation. Average satisfaction has increased from 54 to 61 (on a 0 to 100 scale) in this five years. Some of this increase is due to the addition of the high performing state pharmacies in 1990. Even without the pharmacies, however, there is a steady increase in this sector (from 54 to 59). Importantly, the differences between competitive and government owned services is decreasing over time. Following Fornell and Johnson (1993), this suggests that competitive services provide government agencies with a useful

benchmark for industry satisfaction that were not previously available. It appears that government agencies in Sweden are using these attainable benchmarks to improve performance. It would be more unrealistic to expect agencies, on the whole, to achieve the satisfaction levels that we observe for competitive products where the means of production is quite different. A second implication is that individual government owned or regulated businesses, such as the pharmacies, are capable of achieving even higher satisfaction levels. Overall, the SCSB results thus provide government agencies in Sweden with both attainable benchmarks and role models for setting satisfaction goals.

THE AMERICAN CUSTOMER SATISFACTION INDEX

The SCSB serves as the prototype for the American Customer Satisfaction Index (ACSI) which will be released for the first time in October of 1994. The ACSI is a quarterly, national index of customer satisfaction. Sponsored by the University of Michigan, the National Quality Research Center at the Michigan Business School, and the American Society for Quality Control, the index will, in its first year, survey approximately 50,000 customers of approximately 200 companies and government agencies which comprise about 49 percent of U.S. Gross Domestic Product.

There are important differences between the ACSI and the existing SCSB. First, the ACSI is larger in scope given the greater size of the U.S. economy. American firms are also more diverse in that a single firm is more likely to compete in multiple industry sectors. Sampling is, therefore, being done at the "firm level" rather than the "product or service" level. Finally, the ACSI (and future versions of the SCSB) include an expanded set of survey items. In addition to the original SCSB questions, customers will be asked both their expectations and perceptions of performance regarding two key quality components: (1) "fitness for use," or the degree to which a product or service provides those things that the customer personally requires from the product or service, and (2) "things gone wrong," or the degree to which a product or service is free from defects.

Both factors are germane to quality across all U.S. industries and will provide interesting bases for comparison.

Conclusions: Implications for Government Agencies

Our recent experience in the development of national customer satisfaction indices illustrates a number of important principals and concepts that should help government agencies as they actively implement satisfaction measurement systems. The first is that there is a relatively "universal" model of the antecedents and consequences of customer satisfaction. The challenge that agencies face is in translating the constructs in Figure 1 to the particular agency context. Customer loyalty, for example, may be "repurchase" in some agencies (e.g., the buying of Census Bureau data) and "compliance" in others (e.g., with an IRS regulation or rule). A second implication is that there are universal ways of asking the survey questions needed to operationalize such things as performance, satisfaction, and loyalty. This involves a focus on common denominators, such as "value" when operationalizing perceived performance, and using multiple standards of comparison, as when measuring satisfaction.

Once a flexible model and a set of survey measures are in place, the measures should be used to develop indices of the key constructs. This is especially true for satisfaction where any single survey item is at best a proxy for a customer's overall evaluation of their experience with a firm or agency. Ideally, the satisfaction index should be estimated within the context of a model (e.g., Figure 1) where, for example, performance and expectations explain satisfaction and satisfaction, in turn, explains customer voice and loyalty. These steps address many of the criticisms raised by consumer and policy researchers such as Ölander and Hunt toward the use of subjective measures of satisfaction. They help assure comparability in satisfaction measures across people and industries.

However, the ultimate test of this comparability is an empirical one. As the Swedish experience shows, customer satisfaction is empirically comparable. When

customer satisfaction is properly surveyed, measured and modeled, it allows one to compare "apples and oranges." The resulting comparisons provide useful benchmarks for government agencies as they improve quality. Competitive service industries provide a very straightforward benchmark that, based on the Swedish experience, appears attainable for government agencies as a whole. A national index also allows one to identify a particular agency or agencies to serve as role models and provide even higher satisfaction goals.

At another level, having established the comparability of satisfaction surveys, government agencies can use satisfaction index results to make better decisions and resource allocations. Existing productivity measures and price indices are limited in the way they account for quality changes (National Economic Research Associates 1991). Resources could be allocated more effectively by targeting industries or agencies that rate particularly low on satisfaction to help improve overall consumer welfare. For example, if the IRS rates particularly low on satisfaction, allocating resources toward improving customer satisfaction should more than pay for itself in terms of increased efficiency, compliance and resulting revenue generation. Finally, agencies will benefit by having a more complete picture of their organizations. Any comprehensive strategic plan for a public or private organization must integrate the organization's goals for achieving customer, employee, and owner satisfaction. In government agencies, taxpayers are the ultimate "owners." As agencies strive to meet customer needs, build customer loyalty, and save taxpayer dollars, these owners are the ultimate winners.

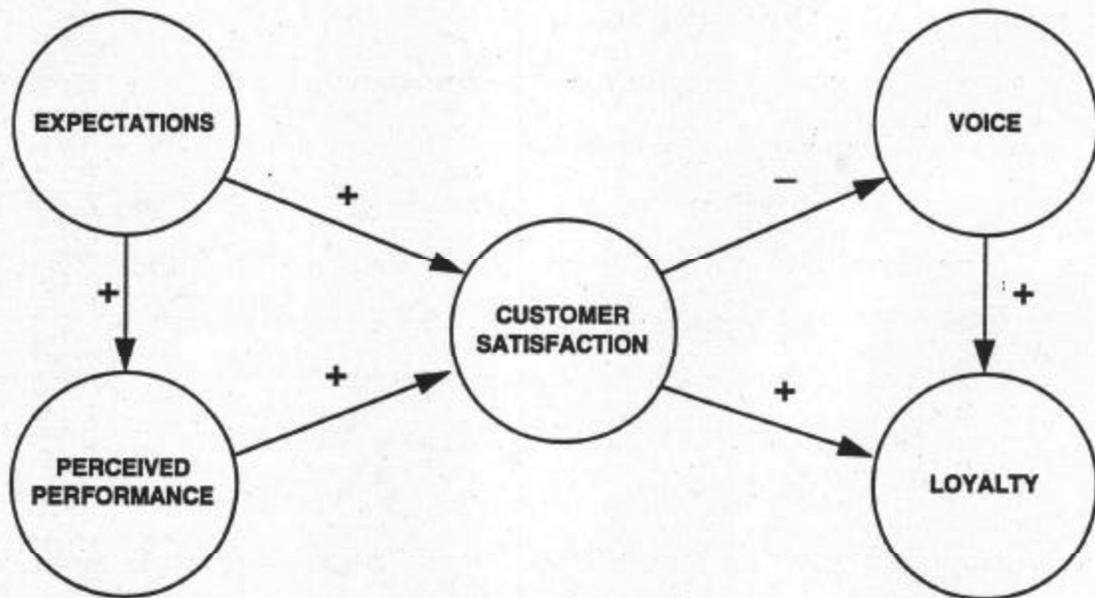


Figure 1. The SCSB Model

The SCSB: 1989 - 1993

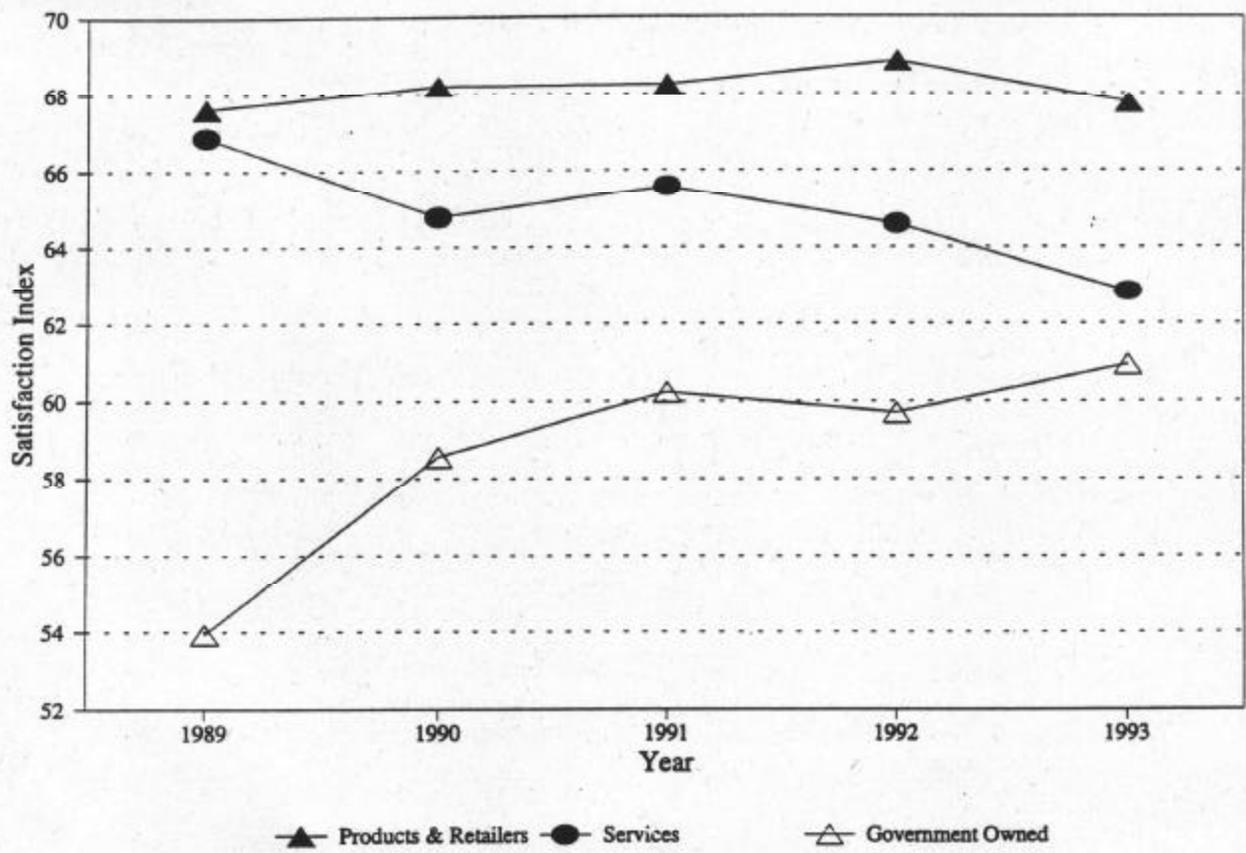


Figure 2. Year-to-Year Changes in the SCSB

Industry	Satisfaction by Year*				
	1989	1990	1991	1992	1993
Airlines	67	67	68	63	65
Automobiles	77	78	78	76	77
Banks (Business)	70	67	64	65	59
Banks (Public)	69	69	67	67	63
Clothing Retailers	63	63	62	63	63
Computers (Main Frames)	69	63	63	64	62
Computers (Business PCs)	70	66	66	67	64
Department Stores	62	63	61	61	60
Food Processors**	68	71	71	72	70
Furniture Retailers	64	63	65	65	64
Gas Stations	67	68	70	70	70
Grocery Stores	66	68	65	67	66
Insurance (Business)	64	62	64	62	61
Insurance (Automobile)	66	63	66	64	62
Insurance (Life)	65	65	63	61	54
Mail Order	na	64	63	64	64
Newspapers	na	60	64	63	62
Pharmacies	na	76	73	72	74
Police	56	55	58	59	58
Postal Service (Business)	59	62	65	61	66
Postal Service (Public)	65	61	67	63	65
Railroad	44	55	54	54	54
Shipping	na	65	69	67	69
Travel (Charter)	68	67	68	68	68
Telecom. (Business)	53	57	57	61	61
Telecom. (Public)	55	59	61	59	61
TV Broadcasters	44	43	47	48	49

* Satisfaction index is on a 0 to 100 scale.

** The averages for Food Processors include six separate food industries (basic foods, candy and coffee, baked goods and dairy products, beer, meat products, and canned and frozen foods).

Table 1. Swedish Customer Satisfaction Barometer Results

REFERENCES

- Anderson, Eugene W., Claes Fornell, and Donald R. Lehmann (1994), "Customer Satisfaction, Market Share, and Profitability: Findings from Sweden," *Journal of Marketing*, forthcoming.
- Andreasen, Alan R. and Arthur Best (1977), "Consumers Complain - Does Business Respond?," *Harvard Business Review*, July-August 1977, 93-101.
- Bentham, J. (1802), in E. Dumont, *Theory and Legislation*, London: Trubner, 1871.
- Fornell, Claes (1989), "The Blending of Theoretical and Empirical Knowledge in Structural Equations with Unobservables," in H. Wold (ed.), *Theoretical Empiricism*, New York: Paragon House, 153-173.
- _____ (1992), "A National Customer Satisfaction Barometer: The Swedish Experience," *Journal of Marketing*, 56 (January), 6-21.
- _____ and Michael D. Johnson (1993), "Differentiation as a Basis for Explaining Customer Satisfaction Across Industries," *Journal of Economic Psychology*, 14 (4), 681-696.
- Hammond, Peter J. (1991), "Interpersonal Comparisons of Utility: Why and How they are and Should Be Made," in J. Elster and J. E. Roemer (eds.), *Interpersonal Comparisons of Well-Being*, Cambridge: Cambridge University Press, 200-254.
- Hauser, John R. and Steven M. Shugan (1983), "Defensive Marketing Strategies," *Marketing Science*, 2 (4), 319-360.
- Hicks, J. R. (1939), *Value and Capital*, London: Oxford University Press.
- Hunt, H. Keith (1988), "Consumer Satisfaction/Dissatisfaction and the Consumer Interest," in E. Scott Maynes (ed.), *The Frontier of Research in the Consumer Interest*, Columbia, MO: American Council on Consumer Interests. 731-747.
- Johnson, Michael D. (1984), "Consumer Choice Strategies for Comparing Noncomparable Alternatives," *Journal of Consumer Research*, 11 (December), 741-753.

- _____ (1989), "The Differential Processing of Product Category and Noncomparable Choice Alternatives," *Journal of Consumer Research*, 16 (December), 300-309.
- _____ (1994), "The Four Faces of Aggregation in Customer Satisfaction Research," Working Paper, Ann Arbor, MI: National Quality Research Center.
- _____, Eugene W. Anderson and Claes Fornell (1994), "Rational and Adaptive Performance Expectations in a Customer Satisfaction Framework," Working Paper, Ann Arbor, MI: National Quality Research Center.
- _____ and Claes Fornell (1991), "A Framework for Comparing Customer Satisfaction Across Individuals and Product Categories," *Journal of Economic Psychology*, 12 (2), 267-286.
- _____, Georg Nader and Claes Fornell (1994), "Expectation, Perceived Performance, and Customer Satisfaction for a Complex Service: The Case of Bank Loans," Working Paper, Ann Arbor, MI: National Quality Research Center.
- Jorgenson, D. W. (1990), "Aggregate Consumer Behavior and the Measurement of Social Welfare," *Econometrica*, 58, 1007-1040.
- Lancaster, Kelvin (1971), *Consumer Demand: A New Approach*, New York: Columbia University Press.
- Lingoes, James C. and Martin Pfaff (1972), "The Index of Consumer Satisfaction: Methodology," in M. Venkatesan (ed.), *Association for Consumer Research: 3rd Annual Conference Proceedings*, 689-712.
- Lohmöller, J.-B. (1989), *Latent Variable Path Modeling with Partial Least Squares*, Heidelberg: Physica-Verlag.
- Meeks, J. G. Tulip (1984), "Utility in Economics: A Survey of the Literature," in Charles F. Turner and Elizabeth Martin (eds.), *Surveying Subjective Phenomena, Volume 2*, New York: Russell Sage Foundation, 41-91.

- National Economic Research Associates (1991), *Developing a National Quality Index: A Preliminary Study of Feasibility*, Report prepared for The American Quality Foundation, White Plains, NY: NERA, Inc.
- Ölander, Folke (1977a), "Can Consumer Dissatisfaction and Complaints Guide Public Consumer Policy?" *Journal of Consumer Policy*, 1 (Spring), 124-137.
- _____ (1977b), "Consumer Satisfaction - A Skeptic's View," in H. Keith Hunt (ed.), *Conceptualization and Measurement of Consumer Satisfaction and Dissatisfaction*, Cambridge, MA: Marketing Science Institute, 409-452.
- _____ (1988), "Consumer Satisfaction/Dissatisfaction and the Consumer Interest," in E. Scott Maynes (ed.), *The Frontier of Research in the Consumer Interest*, Columbia, MO: American Council on Consumer Interests, 753-759.
- Oliver, Richard L. (1977), "Effect of Expectation and Disconfirmation on Post-exposure Product Evaluations: An Alternative Interpretation," *Journal of Applied Psychology*, 62, 480-486.
- _____ (1980), "A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions," *Journal of Marketing Research*, 17 (November), 460-469.
- Pfaff, Martin (1977), "The Index of Consumer Satisfaction: Measurement Problems and Opportunities," in H. Keith Hunt (ed.), *Conceptualization and Measurement of Consumer Satisfaction and Dissatisfaction*, Cambridge, MA: Marketing Science Institute, 36-71.
- Poiesz, Theo B. C. and Jasper von Grumbkow (1988), "Economic Well-Being, Job Satisfaction, Income Evaluation and Consumer Satisfaction: An Integrative Attempt," in W. F. van Raaij, G. M. van Veldhoven and K. E. Wärneryd (eds.), *Handbook of Economic Psychology*, Dordrecht, the Netherlands: Kluwer Academic Publishers, 570-593.

- Robbins, L. (1938), "Interpersonal Comparisons of Utility: A Comment," *Economic Journal*, 48, 635-641.
- Scitovsky, Tibor (1951), "The State of Welfare Economics," *American Economic Review*, 41, 303-315.
- Sen, A. K. (1979), "Interpersonal Comparisons of Welfare," in M. J. Boskin (ed.), *Economics and Human Welfare: Essays in Honor of Tibor Scitovsky*, New York: Academic Press, 183-201.
- Simon, John L. (1974), "Interpersonal Welfare Comparisons can be Made - and Used for Redistribution Decisions," *Kyklos*, 27, 63-98.
- Tinbergen, Jan (1991), "On the Measurement of Welfare," *Journal of Econometrics*, 50 (1), 7-13.
- Van Raaij, W. Fred (1981), "Economic Psychology," *Journal of Economic Psychology*, 1 (1), 1-24.
- Wikström, Solveig (1983), "Another Look at Consumer Dissatisfaction as a Measure of Market Performance," *Journal of Consumer Policy*, 6 (1), 19-35.
- Wold, Herman (1982), "Systems Under Indirect Observation Using PLS," in Claes Fornell (ed.), *A Second Generation of Multivariate Analysis: Methods*, New York: Praeger, 325-347.
- Zeithaml, Valerie A., Leonard L. Berry, and A. Parasuraman (1988), "Communication and Control Processes in the Delivery of Service Quality," *Journal of Marketing*, 52 (April), 35-48.

Session 3 Customer Surveys

Discussion

Robert M. Groves

University of Michigan and Joint Program in Survey Methodology

In addressing the notion of customer service standards and customer measurement programs, the U.S. government is attempting to import a set of ideas tried in the commercial sector. It is useful to note that the ideas, once tried, do not always prove themselves to be cures of the ailments of modern commercial organizations. Why they work sometimes and why they don't other times is the topic of much current debate. We are now living through the period of time when most of you in the audience are determining whether this is the management philosophy of the week or the beginning of a new perspective on agency functioning.

First, let's take a minute to review the recent history of the commercial sector: the common lessons of the "customer satisfaction movement" are:

- external threats help shock organizations into paying attention to their customers
- customer orientation succeeds only when top management forces it, repeatedly, in every forum, relentlessly
- measurement of satisfaction only once is nearly useless
- measurement of satisfaction without simultaneous measurement of production/service activities related to satisfaction is nearly useless

Now let's see whether these lessons are relevant to the papers we have heard presented today.

The Devens Paper

There are really three parts to the Devens paper -- a commentary on the customer-orientation movement, the description of a survey, and a review of the feedback loop to managers.

Commentary on customer-orientation movement

Devens notes that managers do "want" to improve their operations, one way or another. Clearly, the question is whether the operations *are* being improved. It seems clear that all change is difficult to induce in government agencies, but if anything, continuous improvement changes may be more difficult in ongoing statistical operations than in other areas. The problem stems from the need to maintain comparable measurement

systems over time in ongoing series. The concern is change that affects the bias properties of estimates, not just the variance properties (yet even changes that theoretically affect only variance properties (eg. a new sample) can affect bias).

Why is that a concern? Rarely do we have information that bias change goes in the right direction. Exceptions are the higher victimization rates of the NCS and higher unemployment rates of CPS, where there is a strong model of tendencies to underreport those phenomena, but even there the model can be easily challenged.

So incremental change in statistical operations may be harder than incremental change in other fields because the product of today has more value if it is comparable to the product of yesterday.

Devens notes that many managers are skeptical about customer focus because customers aren't qualified to judge the quality aspects of statistical series. This comment on the surface sounds familiar to those in charge of the design of the 1994 Chevrolet Caprice, but statistical agencies may have greater challenges than manufacturers of other products. The General Motors managers did have access to many marketing research studies about the concerns and interests of their customers, but apparently discounted them. Most statistical agencies have no equivalent of the market research function, and thus customer desires are only indirectly and erratically communicated. As with automobiles it is easy to confuse the fact that only the customers know what information they need, but only the statisticians may know how best to produce it.

Description of the Survey

In this section there is very little concern about issues concerning the sampling frame and inference. This unfortunately is a serious lacuna in most of the literature about the Total Design Method. Despite its name it does not address issues of coverage error in surveys. These issues are complex and largely uncharted in surveys of customers of statistical agencies because the target population of customers has usually not been fully enumerated at the time of a survey. Even the definition of "customer" becomes a complex one, when considering information as a product.

The most important point of this section is that careful planning of a mailed questionnaire can yield high response rates. For this reason alone, this is an important paper. When government agencies are telling others and are being told that response rates in the 20% range are the highest to be expected, this work has shown that careful planning and execution can obtain high participation rates. High response rates are doubly important in this area, because of the finding that nonrespondents to satisfaction surveys tend to be dissatisfied with the service or product.

Review of Feedback Loop to Managers

Despite some commentary that the survey did not yield clear findings, conclusions were indeed drawn from the data collection -- the need for more timely faster products, one stop service for questions, and clearer presentations of information. Of these, it appears that the organization has addressed timeliness of products most directly. It is noteworthy

that this area was not the one with the largest gap between expectations and perceived performance. This might be an example of management either ignoring the empirical findings or management supplementing the empirical work with other external information about performance. The paper does not reveal which is the appropriate interpretation.

The Johnson Paper

This paper describes a large effort to construct a useful measure of consumer satisfaction across all sectors of the economy. It is conceptualized as another macroeconomic indicator, measuring an outcome of production-- in one term "post-consumptive utility" The paper is divided into three sections: a) Can satisfaction be compared across sectors/industries?; b) a description of the Swedish Customer Satisfaction Barometer; and c) an announcement of the American Customer Satisfaction Index

Can satisfaction be compared across sectors/industries?

How would you know whether you'd have the answer to this question? Would it depend on the ability to predict behavior? What behavior? The evidence of comparability presented includes the finding that 50% of the variance in perceptions of performance is explained by the amount of differentiation in the industry, and performance explains 50% of variance in satisfaction. Clearly, one would like to assemble more evidence: behavioral outcomes measured on same persons over time, stable relationships between satisfaction levels and growth rates, complaint rates, etc.;

Swedish Customer Satisfaction Barometer

This is a large data collection and estimation series, running since 1989, covering 70% of sales in each industry, and measuring one product or service per company. Clearly the process of sampling firms and products/services is a nontrivial problem, as those in the consumer and producer price index measurement process know. The sampling problem facing this index is multi-level (sectors, firms, products/services, customers) there are important sources of variance at each level and important sources of information about customer satisfaction.

The concepts measured include: perception of performance, expectations regarding performance, satisfaction, reports of whether the customer has complained about the product or service, and customer loyalty. These are difficult measurement issues. For example, the approach is forced to use hypothetical questions on loyalty, using words like "if another agency could provide the same service."

The paper presents findings from the Swedish effort that are stimulating, given the current effort at measuring customer satisfaction in U.S. government agencies. For example, there is the finding that Swedish government agencies (police, pharmacies, post office, railroads, telecommunications, tv broadcasting) started with lower satisfaction and rose in satisfaction faster than other sectors. One wonders whether that finding will be duplicated in the US. The finding itself illustrates one of the challenges to the measurement process. To what extent is measurement of satisfaction with services of a government agency affected by general feelings of civic pride, trust in government,

political efficacy?

American Customer Satisfaction Index

The last section of the paper sketches out the plans for a U.S. customer satisfaction index. This effort will be different from the Swedish experience in that the U.S. population of firms offer more diverse products and services.

In both this and the Swedish index description there seems to be most emphasis placed on the psychometric properties of the measurement and little concern with traditional survey issues of coverage of the target population (telephone surveys are planned), nonresponse error, and measurement errors associated with social desirability, mode of data collection, etc.

Summary

These two papers, although seemingly disparate in the focus, can serve to remind us of two important debates in customer satisfaction:

- Is satisfaction merely a function of the difference between reported expectations and performance ratings?
- Do expectations cause perceived performance?

Let me summarize my reactions to the papers:

1. In government agencies, we are at the beginning of the customer measurement process. Its value rests on repeated measurement, empirical assessment of relationship between actions of employees and satisfaction, and change in satisfaction over time. We are a long way from this status of measurement and innovation.
2. Both of these papers appear to miss the connection to actual activities of the units to increase satisfaction. They are more heavily focused on measurement than how measurement can lead to change and then later to improved satisfaction.
3. The papers flow from different conceptual bases; a debate that is not joined by the two. One stems from the notion that satisfaction is in some sense the gap between expectations and performance. The other attempts to add another concept, perceptions of an ideal service or product, in order to calibrate the gap between expectations and performance. These conceptual differences are part of the debate now ongoing in the satisfaction measurement field. These are important issues for the practical import of satisfaction measurement. If, for example, performance at time 1 sets expectations at time 2, then poor performance lowers expectations, and in one perspective, would yield higher satisfaction, as expectations and performance were in sync. From the other perspective, departures between performance and the concept of the ideal, would be larger at time 2 and lead to large "performance gaps."

If government agencies take seriously the measurement of customer satisfaction, they will

inevitably be forced to attend to such issues. They are key to the meaningful tracking of how satisfaction can change with improved performance of organizations. We are at the beginning of this process for government agencies, and we are in the debt of these two papers for alerting us to such issues.

DISCUSSION

Elizabeth Martin
Bureau of the Census

Both papers presented in this session raise issues which are relevant to the current efforts by federal agencies to respond to a presidential order to survey their customers, measure satisfaction, and use the results to set service standards and provide customers with greater choice in services. . . . Significantly, the executive order further states that "as information about customer satisfaction becomes available, each agency shall use that information in judging the performance of agency management and in making resource allocations."

If this aspect of the executive order comes to pass, then the issue of the comparability of customer satisfaction measurements among agencies and across diverse products and services is of more than academic interest. I'd like to start by focussing on the issue of comparability of measurement as addressed in the Johnson paper, which describes a customer survey conducted across 32 industries in Sweden, including government-run industries such as the postal service and railroads. I'll be drawing on my recent involvement in an effort to design a generic customer satisfaction questionnaire for use by all of the agencies of the Department of Commerce. Next, I'll discuss the Devens paper, which discusses a customer satisfaction survey targetted much more narrowly, to users of employment and unemployment statistics produced by the Bureau of

Labor Statistics. It raises some interesting issues about the utility and consequences of customer satisfaction surveys.

Johnson and his colleagues are concerned with a very ambitious effort to develop a customer satisfaction index (the Swedish Customer Satisfaction Barometer) which can be applied across industries, products, and, ultimately, in different countries. Customers were sampled from 32 Swedish industries, including several which were government-run, which represented 70 percent of Sweden's gross domestic product. The companies that accounted for approximately 70 percent of combined industry sales were selected to represent each industry. Each company was represented by its product or service with the highest sales. Each year, 100,000 customers were contacted by telephone and screening questions were asked to determine if they had experience with the products and services chosen to represent the sample companies. About one quarter were eligible and were asked questions to measure their perceptions of performance, their expectations about performance, their satisfaction, whether they ever complained, and their loyalty.

Johnson argues that the survey measurements may be used to compare satisfaction and performance across industries on the following basis: that the model of customer satisfaction which informs their measurements is universally applicable across industry; the measurements used in their survey are universally applicable across all industries; and that there exist meaningful differences in satisfaction between industries which can be explained by industry-level differences in degree of product or

service differentiation--the amount of choice offered customers. He also suggests that that the index provides meaningful and comparable information about customer satisfaction for government agencies and industries as well as private sector ones.

Although it may be possible to design a survey which yields comparable measurements of satisfaction for customers of diverse products and services produced by different industries, Johnson and his colleagues have not satisfactorily made the case for the comparability of their measurements. They need to address the following questions: First, is the definition of a customer comparable across different industries? Second, are the sampling frame and response rates comparable across industries and over time? Finally, are their satisfaction measurements comparable and meaningful across different industries and over time?

The first issue, of what is a customer, usually is not terribly ambiguous in the private sector, but it bedevils attempts to measure customer satisfaction in a government setting. To most of us, a customer is someone who purchases a commodity or service, usually by choice or voluntarily. In a government setting, many products and services are not purchased directly by their users, but subsidized in whole or in part by taxes. Many government products and services are not received voluntarily on the part of the "user" or recipient. Many "customers" of police services or tax collection services no doubt would, if they had the choice, choose not to obtain the service at all. The Johnson paper does not address this issue. In their survey, they defined as customers persons who had experience with the products and services surveyed.

This is a reasonable strategy, but it is important to include measures in the survey to permit the analyst to separately identify voluntary customers of government services, who obtained a product or service by choice, and involuntary customers, who did not choose to obtain the product or service. Comparisons of the former group with customers of private industry may be meaningful, but comparisons involving the latter group probably would not be.

Another definitional problem arises because many services offered by the government are not intended to benefit those who experience them directly, but to protect or benefit others, such as the public, who may not even be aware of their existence. For example, one service provided by a Department of Commerce agency is the inspection of fisheries. Presumably this service is ultimately intended to benefit fish-eaters by ensuring the quality of fish, and only indirectly benefits the fisheries themselves. In this example, it would be difficult to measure the satisfaction of customers who may be unaware that a service exists, much less that they are recipients of it. This sort of issue can make it difficult to identify who should be regarded as the customers of government services, and this category of "customer" would be ruled out by Johnson's screening criterion of "having experience" with a product or service.

A second set of very difficult issues affecting comparability of data has to do with the identification and sampling of customers. In order to make comparisons across industries, one must be certain that the samples are comparable. Johnson reports that each year customers were identified in surveys of the public in

which 100,000 individuals were asked about their experiences with the target products. It is not clear how or whether the sample represents organizational customers. For many of the industries being evaluated--such as banks, railroads, main frame computers--much if not most of their business would be with other organizations or businesses, not with individual consumers. Organizations, and their experiences as customers, do not appear to be represented in the index. Their absence reduces the meaningfulness of customer satisfaction measures for industries in which organizational customers represent a large share of all customers, and reduces comparability of measures across industries which differ in their customer base.

Even if one accepts the limitation that only individual customers are represented in the sample, it is still unclear what universe the results represent. Johnson surveyed customers of the leading products of companies representing 70 percent shares of each of a set of industries which together accounted for 70 percent of the Swedish GDP. One would expect the companies, products, and industries included in the index to change over time with changes in the economy. This implies that there are two potential sources of change in the value of the index: changes in customer satisfaction for a given set of products, and changes in the composition of products, companies, and industries which make up the index. Given the uncertain interpretation which could be put on any given change in its level, it is not clear how a customer satisfaction index defined this way can provide useful information about trends. Moreover, Johnson's definition appears to leave out most customers,

since 75 percent of his sample was ineligible for the survey. The limited and rather peculiar constraints on the set of customers included in the survey seem to reduce its usefulness as a general index of customer satisfaction with wide applicability across different countries.

The quality and comparability of the results of customer satisfaction surveys depends not only on the quality of the sampling frame, but also on response rates. Johnson presents no information about response rates in the customer survey he reports. In order to make comparisons among industries, one would want reasonably high response rates for all the industries being compared. If response rates varied among industries, then artifactual differences in satisfaction may result from greater nonresponse bias for some than others.

In general, the construction of sampling frames for customer surveys is problematic. If there exist records of purchases, orders, or logs of telephone or other contacts, then these may be used as a sampling frame. However, for many services and products, there are no records which identify customers, especially if no formal or recorded transaction takes place. Customers who pay cash for a product or service, or who listen to the weather station or look up information in a census publication in the library cannot be readily identified. Thus, sampling from records or logs of customer transactions is likely to provide uneven coverage of customers depending on the nature of the industry, how it conducts business with its customers, and the quality and completeness of the records it keeps about customers or transactions. In some

customer surveys, samples are drawn from lists of customers provided by an agency or firm specifically for the survey. Such lists can be very vulnerable to selection bias, since organizational representatives who know that customer satisfaction is to be evaluated are likely to overrepresent satisfied customers in their lists. This selection bias may vary among agencies or organizations and could have a very adverse effect on the comparability of satisfaction measures across agencies or organizations.

The third issue which needs more attention in the Johnson paper is the comparability and meaningfulness of the measurements, especially when applied in a government context. Performance was measured as value, or benefits relative to costs, which seems not to apply very well to products or services which have no specific or direct cost attached to them, as is the case for many government products and services. The key construct of customer loyalty was operationally measured by intention to repurchase and insensitivity to price. Because these measures do not fit government's transactions with its customers, Johnson and his colleagues changed the measures to hypothetical ones for government agencies. It is questionable whether this modification yields results comparable to the original measure. Finally, customer expectations were measured retrospectively, that is, customers were asked to report what their expectations had been at the time of purchase. Retrospective reports of past attitudes are notoriously biased toward present attitudes, and it is highly likely that this measure of "expectations" is contaminated by respondents' subsequent experiences with

a product. This flaw would make it impossible to test the effect of prior expectations on satisfaction or perceived performance.

In summary, customer surveys which aim to compare across diverse industries and products (such as the SCSB discussed by Johnson) potentially are affected by very serious problems of data comparability, including lack of comparability arising from sample design, differential nonresponse, and the measurements themselves. There appears to be a considerable amount of careful methodological and statistical work that still needs to be done to ensure that customer surveys are designed to yield meaningful comparisons of customer satisfaction across industries and over time. Until that groundwork is done, such comparisons should be made cautiously.

The Devens paper raises a different set of issues. The Customer Satisfaction Survey he reports on was much narrower in scope and purpose than the satisfaction index discussed by Johnson, and the issues of data comparability are not nearly as serious, if they exist at all. The survey of customers of employment and unemployment statistics was well done, and obtained an admirably high response rate (88 percent) using reminders and mailings of follow up questionnaires. The survey assessed several aspects of the statistical product, including data quality, ability of staff to answer technical questions, etc. Devens reports that the survey results moved the agency to take several actions to improve the timeliness of the release of its statistics, which the survey showed customers thought was very important.

The paper includes a couple of telling comments by the author, who personally evaluates the survey as "useful but not nearly in

proportion to the skills exercised or the resources expended." He also notes the very low response rate--8 percent-- obtained from a survey of front-line employees upon release of the customer survey, and voices his suspicion that this low rate is due to "failure to convince the front line that the customer satisfaction survey was serious."

Devens' remarks remind us of a couple of key points about customer surveys of this type. First of all, there needs to be clear specification of the goals of the survey, and an understanding of how the information from the survey will be used, in order for the survey to be useful. (This point applies to any survey, not just customer surveys.) To be taken seriously by employees, a customer survey should be designed to address questions to which managers and employees need or want the answers. In the case of the survey Devens reports, it appears that the survey was not credible to managers, and was used in a very limited way by them, reducing the meaningfulness of the survey.

The second point is that customer surveys can themselves affect the expectations of customers and employees. Carrying out a customer survey may raise the expectations of customers (and employees) that a company or agency is going to do something to improve service. If that doesn't happen, and if the survey turns out to be an empty exercise, then the indirect effect of a customer survey may be to reinforce the cynicism of customers or employees or both.

Taking these two points together, and returning to the earlier discussion, we can draw several general conclusions about customer

surveys: Unless an agency or company plans to actually use the results of a customer survey, it shouldn't conduct the survey. The survey should be planned with clear goals and uses in mind. It should be designed to provide fairly specific information that represents useful feedback to managers and employees, and that has implications for action. If the intent is to compare customer satisfaction over time, among products, or among industries and agencies, then the survey should be designed and data evaluated to ensure that results are comparable and can support the comparisons to be made.

Session 4

ADVANCES IN DATA EDITING

IMPROVING OUTLIER DETECTION IN TWO ESTABLISHMENT SURVEYS

Julia L. Bienias,¹ David M. Lassman, Scott A. Scheleur, and Howard Hogan
U. S. Bureau of the Census

I. Introduction

One step in producing estimates from survey data is editing. In many settings, trained analysts examine the data to find unusual or unexpected values, which may be the result of errors made by the respondent or in the data-capture processes. Having found a questionable case, the analyst then tries to verify its accuracy by checking the original form, obtaining related data from other sources, and/or contacting the respondent. One would like to correct as many errors as possible within the time limitations for a given survey. Thus, accurately identifying the cases whose values are most likely to be the result of errors is an essential part of efficient editing.

Previous researchers have successfully used various graphical methods to improve both the efficiency and accuracy of the editing process (e.g., Esposito, Fox, Lin, & Tidemann, in press; Granquist, 1990; Houston & Bruce, 1992; Hughes, McDermid, & Linacre, 1990). We describe the application of graphical methods from exploratory data analysis to the task of identifying potentially incorrect data points. Our report is the result of a working group of analysts, research statisticians, and programmers devoted to this effort.² We illustrate the methods with data primarily from the Annual Survey of Communication Services and the Monthly Wholesale Trade Survey. We first describe the two surveys and the current methods used for editing.

2. Descriptions of the Two Surveys

2.1 The Annual Survey of Communication Services

The Annual Survey of Communication Services (ASCS) is a mail survey covering all employer firms that are primarily engaged in providing point-to-point

¹This paper reports general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau. Address correspondence to: Julia L. Bienias, Economic Statistical Methods and Programming Division, Bureau of the Census, FOB 3015-4, Washington, DC 20233.

²We thank the other members of our working group, without whom the work described here would not have been possible: Thomas Bell, Willard Caldwell, Vicki Garrett, Imelda Hall, Donald Hundertmark, Jennifer Juzwiak, William Knowlton, David Stachurski, and Georgeann Wright. We also acknowledge the many other members of Business Division and of Economic Statistical Methods and Programming Division who have supported this effort and who continue to support it.

communication services (e.g., telephone, television, radio), as defined in Major Group 48 of the 1987 edition of the *Standard Industrial Classification Manual*. The ASCS provides detailed revenue and expense statistics from a sample of approximately 2,000. The Census Bureau introduced the survey in 1991 to track the explosive growth and change in the industry. The Bureau of Economic Analysis is the primary federal user of the data collected; other users are the Bureau of Labor Statistics and private industry (U.S. Bureau of the Census, 1992.)

2.2 The Monthly Wholesale Trade Survey

The scope of the Monthly Wholesale Trade Survey (MWTS) is all employer firms engaged in wholesale trade, as defined by Major Groups 50 and 51 of the 1987 edition of the *Standard Industrial Classification Manual*. Particularly, the survey covers merchant wholesalers who take title to the goods they buy and sell, collecting sales and inventory information. The MWTS, conducted since the 1940's, is a mail survey of approximately 7,000 firms, of which 3,500 receive forms in a given month.

As with the ASCS, the Bureau of Economic Analysis is the primary federal user of the data. (See U.S. Bureau of the Census, 1994.)

3. Issues Involved in the Current Editing Procedures

After the data from the questionnaires are keyed, a computer program flags cases failing completeness, internal consistency, and/or tolerance edits. Editing review is divided among several analysts for a given survey. Each analyst finds which edits have failed for a case through an interactive correction system or a paper listing, on a case-by-case basis. They can also use a database query system to try to find problem cases that have not already been identified.

There are several disadvantages to this approach. Examining one case at a time does not permit the analyst to obtain a broad view of the behavior of the industry as a whole, and such a view can be of great benefit in determining the impact of an individual unit on the aggregate estimate. In addition, it undoubtedly leads analysts to examine more cases than necessary. Finally, for a few of the ASCS tolerance edits, constant parameter levels derived from previous surveys have been hard-coded into the programs. This implicitly assumes the relationships among the variables are static over time, which may not be the case.

4. Application of Exploratory Data Analysis Methods

4.1 Background

Exploratory data analysis (EDA) can be described as "a set of tools for finding what we might have otherwise missed" in a set of data (see Tukey, 1977). These tools, combined with the analysts' subject-matter expertise, are particularly well-

suited to the task of data editing. In this setting, we are not interested in ascertaining the truth of a postulated economic model or a similar estimation or hypothesis testing problem. Rather, our goal is to determine which cases are unusual with respect to the bulk of the cases and to follow up those cases. In addition to providing methods for displaying data in a variety of ways, EDA emphasizes fitting data using methods that are relatively insensitive to the presence of outliers in the data ("resistant" methods). Such fitting is a way to define and then account for (remove) certain aspects of the data so the analyst can concentrate on other aspects. (See Hoaglin, Mosteller, & Tukey, 1983; Velleman & Hoaglin, 1981.)

EDA fits well with the survey processing environment. Because in the editing stage we expect to find wild observations that might be off by orders of magnitude from the bulk of the data, transformations and resistant techniques are particularly useful in helping us find order amid the chaos. In addition, these techniques allow for efficient examination of large amounts of information at once, an aspect that is particularly valuable in the time- and resource-constrained survey production environment.

From the arsenal of tools collectively called "exploratory data analysis," we considered both univariate boxplots and the more general bivariate fitting. We describe boxplots first, followed by scatter plots and some methods for fitting. In addition, although transformations are applicable to all tools, we describe them in the context of scatter plots, because that is where we used them most.

4.2 Boxplots

Boxplots allow quick visual analysis of the location, spread, and shape of a distribution. Our boxplot has its box spanning the lower and upper quartiles, with whiskers extending from the box to the furthest data point within a distance of one-and-one-half times the interquartile range from the box. We considered data values beyond the whiskers as potential outliers. If the data are reasonably symmetric, then these cutoffs provide a good working definition of cases which may need review. See Tukey (1977) for a discussion of boxplots in general, and Hoaglin, Mosteller, and Tukey (1983) for a discussion of the expected number of outliers for different sample sizes. Note that the whisker definition could be modified to suit the needs of a particular survey operation (e.g., one could use 2 times the interquartile range instead of 1.5).

Figure 1 demonstrates the use of the boxplot for operating ratio (expenses/revenue) data from the ASCS.³ The boxplot shows that the median operating ratio is .7978 and fifty percent of the points lie between .7269 and .9811. The left and right whisker values are .3760 and 1.3401. The cases flagged by the

³To protect the confidentiality of our data, we have not provided details about the particular subset of data analyzed in each plot, nor have we labeled axes when such information could be revealing.

use of the boxplot are different (and fewer in number) than the cases that would have been flagged by the current hard-coded edit parameters, .9 and 1.1. Those parameters fail to help us isolate the "true" outlier cases, as they result in too many cases being flagged. Alternatively, we could flag cases that would appear beyond the whiskers as in our boxplot, an approach that is "dynamic" in that it relies on incoming data to set parameters. At minimum, we could use values from Figure 1 as new hard-coded edit bounds, noting that these revised bounds would no longer be symmetric around one (consistent with the findings of Granquist, 1990).

4.3 Scatter Plots

A scatter plot of two variables is a simple and particularly useful technique. When the data are appropriately transformed, one can use a variety of methods to remove linearity in the scatter and then examine the residuals from the linear fit. This allows us to see patterns that we might otherwise miss when looking at the original data; looking at the residuals from a fit allows us to examine the data on a finer scale (see Section 4.5).

As a vivid illustration of the kinds of problems encountered in editing data, we used another survey for which we had raw responses to a particularly problematic question. One item in the Motor Freight Transportation and Warehousing Survey is the percent of revenue derived from local trucking, a question believed to be confusing to respondents may define "local" in different ways. Figure 2, a scatter plot of these unedited data for the current versus prior period, shows a weak linear relationship. Cases along the 45 degree line are companies whose year-to-year reports are consistent. The reports become more inconsistent the further they are from the 45 degree line. Some of the cases along the vertical axis are "births" to the survey (cases selected during the current period to reflect new firms). Births should be analyzed separately, because they have only current-year data.

4.4 Transformations

Transforming the data so patterns can be more easily discerned is a technique that is important to all graphical and data-fitting methods. It is used to obtain symmetry in the data, to promote linearity, and to equalize spreads between data sets. These properties are assumed, implicitly or explicitly, by many of the techniques we use to analyze data. For example, when we look for outliers by examining a boxplot, we are implicitly assuming the data are supposed to be symmetric. If the data are naturally skewed, many of the points in the tail that appear to be outliers are actually values that are consistent with the underlying distribution. Thus, "discovering" such outliers in the long tail would not be very meaningful. With skewed data, we want to spend our time investigating those data points that are particularly unusual, given that we expect many points far from the bulk of the data. If we transform skewed data to be generally symmetric, we can then find those points.

Because economic data are typically positively-skewed, transformations that lead to the expansion of lower data values and to shrinking the spread of larger data values are particularly useful. (See Hoaglin, Mosteller, & Tukey, 1983, for more details on types of transformations.)

Figure 3 is an example of the use of transformations for the ASCS. The scatter plot of untransformed revenue data (Figure 3a) reveals little, as one case is many times larger than the other cases. Hiding the large case was unsuccessful, as the next largest case was still many times larger than the remaining cases. Instead, taking logs of the data showed a useful scatter plot (Figure 3b). We see a strong linear relationship, which is what we expect for a plot of current and prior data. Cases that do not appear to be following this linear relationship would thus be considered unusual. We also see that the case that appeared to be an outlier in Figure 3a is, in fact, very much in line with the rest of the data.

For the MWTS, a scatter plot of the current inventory data against the current sales data shows that most of the data are bunched in the lower left corner (see Figure 4a). Because both variables are skewed, we first tried a natural log transformation ($\log(x + 1)$). (We added one because a value of 0 for inventory data does not indicate the case is a birth, and thus it may be of value to include such cases.) This overtransformed the data, skewing them in the opposite direction (Figure 4b).⁴ This is because there was a big gap in values between 0 and the next largest value. Such an effect would also occur if there were many establishments with very small reported data and a few with very large values. We then tried taking the square root (Figure 4c) and fourth root (Figure 4d). The latter resulted in the most useful transformation, as most of the data can be seen clearly.

4.5 Fitting

In this section we describe two approaches to fitting, ordinary linear regression and resistant regression. Both were useful, in different ways.

In analyzing ASCS data, we considered the relationship between revenue and payroll for current year data. Figure 5a shows the ordinary least squares regression of revenue on payroll; there are many points clustered near the origin and two cases in the upper right corner. First, we tried removing the two large cases. Again the distribution showed points clustered in the left corner. Such an approach, of iteratively hiding points and refitting, has the disadvantage of being subjective and of essentially requiring analysts to identify outliers first.

One alternative is to use ordinary least squares on transformed data. In this example, logs were useful. Figure 5b shows the fit to the logged data, depicting a

⁴If the cases with 0 reported inventory are ignored, as they might be for other variables, then the logarithm transformation provided a useable picture of the data.

strong linear relationship. The point labeled A is an obvious outlier. Examination of the residuals revealed a pattern, which allowed us to discover that tax-exempt cases were inadvertently being included in the analysis. Tax-exempt cases should be examined separately from taxable cases, because our revenue item only includes taxable receipts. Removing both those cases and point A and refitting the data (Figure 5c) led to the distribution of absolute residuals shown in Figure 5d. This plot can be used to detect outliers, as with a cutoff level C:

$$C = K * (\text{median absolute residual}).$$

We found $K=4$ (corresponding to $C = .7868$) to be the best. All cases above .7868 were examined and most were "true" outliers. For our example, this method was judged by the survey analysts to be excellent for finding outliers.

Unfortunately, ordinary least squares (OLS) can give great weight to fitting a few wild values. It may work well, as in our example, when there are only a few wild cases and the demarcation between usual and unusual is clear. As an alternative, we investigated resistant fitting using the biweight function developed by Tukey (Mosteller & Tukey 1977; McNeil, 1977). This widely-tested iterative weighted-least-squares fitting procedure uses a weighting function defined as:

$$w_i = \begin{cases} (1-u_i^2)^2, & u_i < 1 \\ 0 & \text{otherwise,} \end{cases}$$

where $u_i = (r_i / (c*s))$

r_i ≡ Residual from previous fit for point i

s ≡ mean absolute residual from previous fit

c ≡ scale factor.

Setting $c = 4$ is quite resistant, $c = 8$ is moderately resistant. We stopped iterating when the proportionate change in s was less than 0.01. This required few iterations; resistant regression is a very efficient and fast procedure.

We applied resistant regression to the MWTS, predicting logged current inventory data by logged inventory data from the prior year. We expect a linear relationship. Figure 6a shows the data and the line from the OLS fit, and Figure 6b shows the residuals from that fit. It is easy to see the OLS fit misses the central tendency of the point cloud. Figure 7a shows the fit resulting from resistant regression ($c = 4$). This fit more effectively removed the linearity from the data. The residuals now cluster around 0, as we would want (Figure 7b).

5. A Note on Using Ratios

In many instances, data review has relied on calculating ratios (e.g., sales/payroll) and looking for unusually large or small ratios. There is nothing wrong with this approach per se, but it would be wrong to rely too strongly on it.

The use of ratios assumes a rather simple model of the true relation between the two variables, specifically a straight line through the origin. The true relation may

differ markedly, there may be data clouds following different straight lines. For example, the relationship might be different for a small company than for a large company. It is essential that the data reviewer plot the data and look at the shape. Further, the "acceptable ratios" are often set from historic data, last year's or last census'. The relationships can change systematically throughout the business cycle. One could iterate, calculate the average ratio from the current survey, calculate its standard deviation, identify and remove outliers, and start again. However, given the existence of rather fine iterative resistant fitting tools, it is hard to see the advantage of this approach.

6. Summary and Extensions

We have described how principles and methods from EDA can be used to improve the efficiency and accuracy of editing, by helping analysts see patterns in the data and use that information to prioritize cases for follow-up. Building a successful editing system using this approach is more than just selecting the correct statistical tools. The system must be acceptable to the people who will use it. Creating such acceptance requires training the analysts in the methods described here, as well as incorporating the tools into the current production environment and existing computer systems. To date, we have been successful in getting many people to try the methods on several surveys. In addition to the surveys described previously, these methods are currently being applied to the Motor Freight Transportation and Warehousing Survey, the Service Annual Survey, and the Commodity Flow Survey.

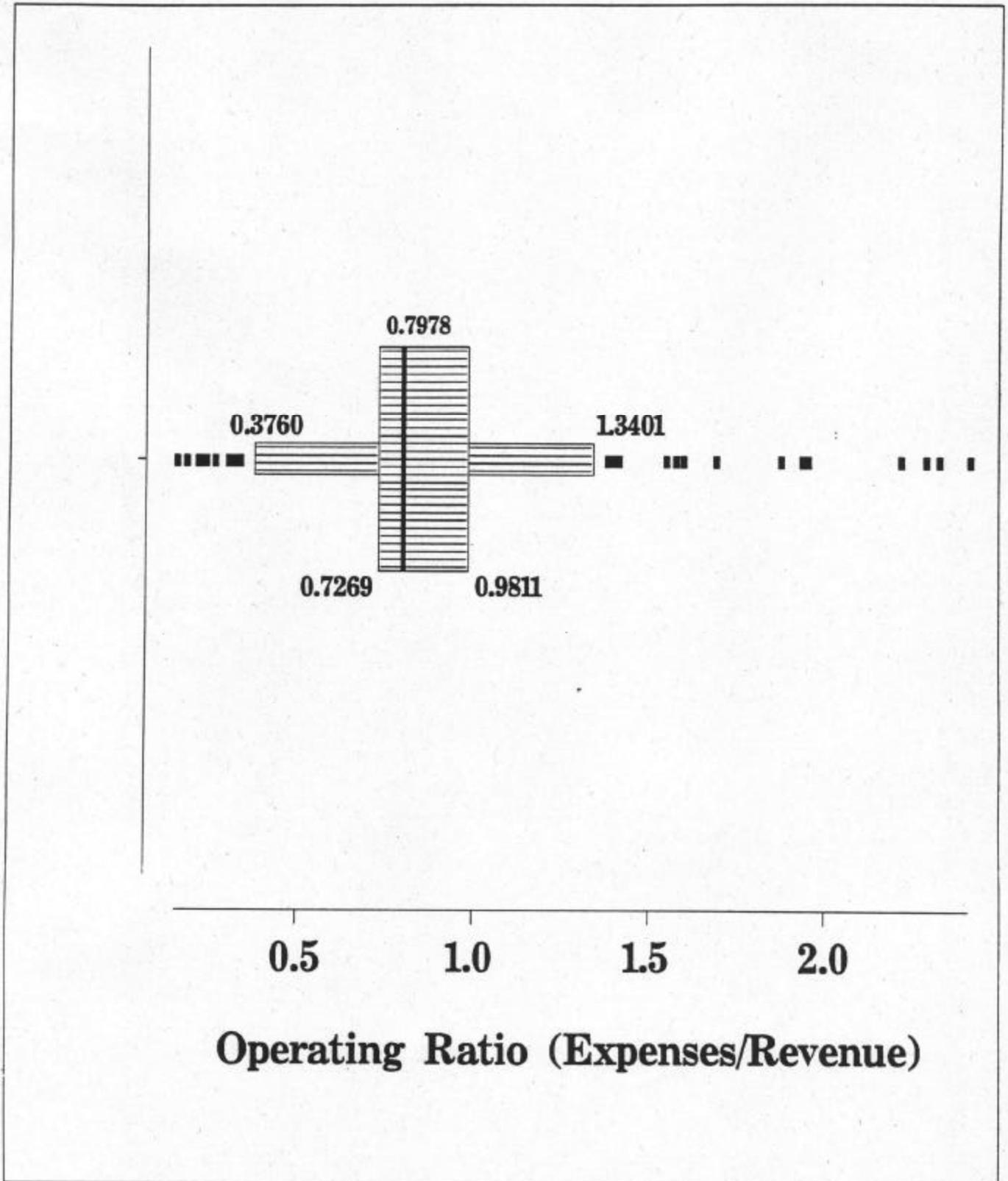
Analysts for these surveys reported that being able to ascertain the effect of a given case on the estimate was quite useful. Other specialized programs written for data editing provide this feature (e.g., Esposito, Fox, Lin, & Tidemann, in press; Houston & Bruce, 1992). Incorporating sampling weights in the procedures described here provides a similar utility.

The EDA approach can be combined with batch-type edits (e.g., SPEER, Draper, Greenberg, & Petkunas, 1990; Lee, in press). One could examine the data flagged from a batch program along with the unflagged data using the tools described here. Or, the graphical-based methods could be the basis for batch-type dynamic edits. For example, a program could transform the data to be more symmetric and then flag all cases that would be beyond the whiskers of a boxplot. Finally, in settings in which hard-coded edit parameters must be used, these methods can be used on a subset of data to help find or evaluate such cutoffs.

7. References

- Draper, L., Greenberg, B., & Petkunas, T. (1990). On-line capabilities in SPEER (Structured Programs for Economic Editing and Referrals). *Proceedings of Statistics Canada Symposium 90: Measurement and Improvement of Data Quality*, pp. 235-44. Ottawa: Statistics Canada.
- Esposito, R., Fox, J. K., Lin, D., & Tidemann, K. (in press). ARIES: A visual path in the investigation of statistical data. *Computational and Graphical Statistics*.
- Granquist, L. (1990). A review of some macro-editing methods for rationalizing the editing process. *Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality*, pp. 225-34. Ottawa: Statistics Canada.
- Hoaglin, D.C., Mosteller, F., & Tukey, J. W. (Eds.) (1983). *Understanding Robust and Exploratory Data Analysis*. NY: Wiley.
- Houston, G., & Bruce, A. G. (1992, February). Graphical editing for business and economic surveys. Technical report, New Zealand Department of Statistics, Mathematical Statistical Branch.
- Hughes, P.J., McDermid, I., & Linacre, S. J. (1990). The use of graphical methods in editing (with discussion). *Proceedings of the 1990 Bureau of the Census Annual Research Conference*, pp. 538-54. Washington, DC: U.S. Department of Commerce.
- Lee, H. (in press). Outliers in survey sampling. In B. Cox et al. (Eds.), *Survey Methods for Business, Farms, and Institutions*. NY: Wiley.
- Mosteller, F., & Tukey, J. (1977). *Data Analysis and Regression*. Reading, MA: Addison Wesley.
- McNeil, D. R. (1977). *Interactive Data Analysis*. NY: Wiley.
- Office of Management and Budget. (1987). *Standard Industrial Classification Manual*. Available from National Technical Information Service, Springfield, VA (Order no. PB 87-100012).
- Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.
- U.S. Bureau of the Census. (1992). *Annual Survey of Communication Services: 1992*. Washington, DC: U.S. Government Printing Office (Current Business Reports, Item BC/92).
- U.S. Bureau of the Census. (1994, April). *Combined Annual and Revised Monthly Wholesale Trade, January 1987-December 1993*. Washington, DC: U.S. Government Printing Office (Current Business Reports, Item BW/93-RV).
- Velleman, P.F., & Hoaglin, D. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury Press.

Figure 1



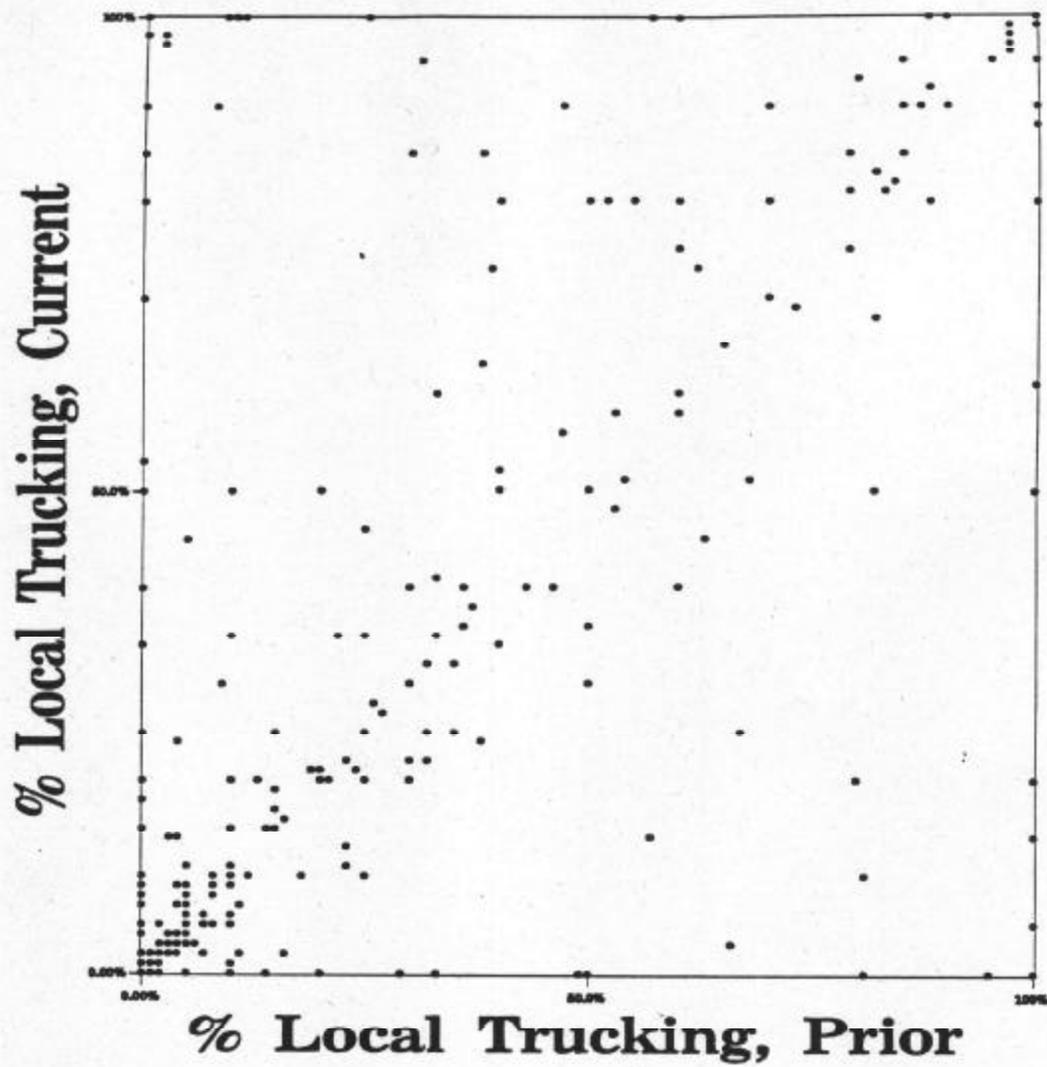
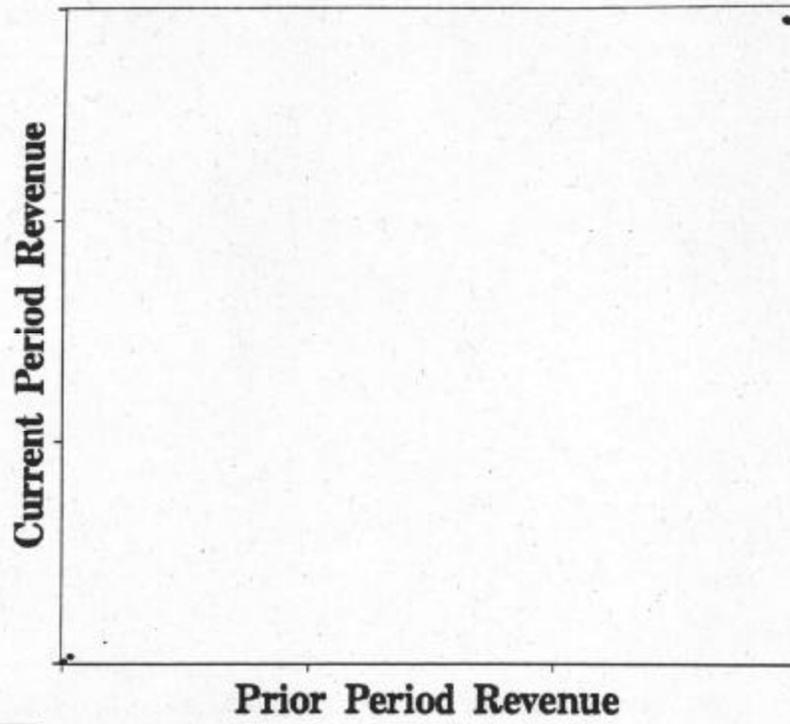
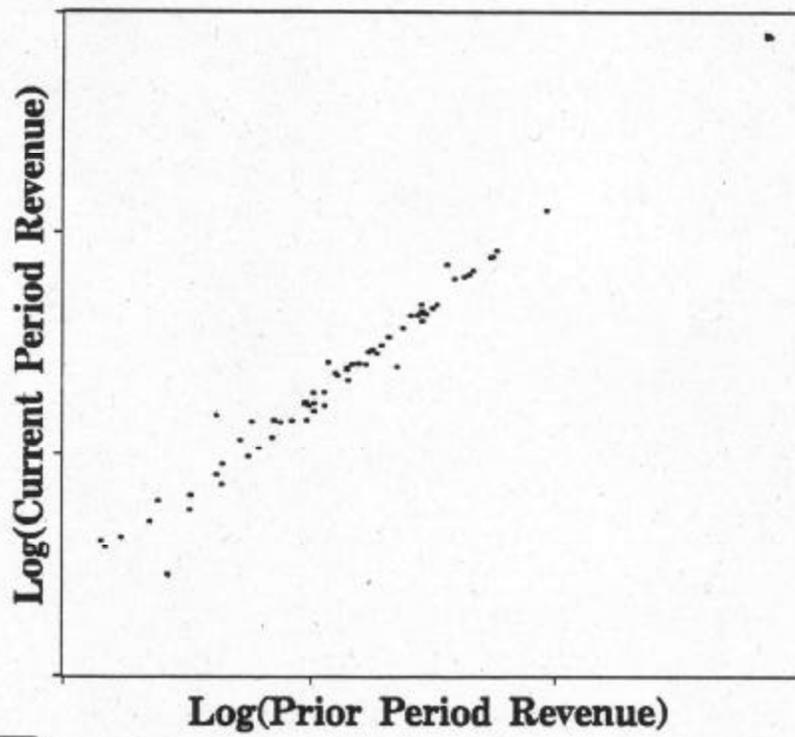


Figure 2

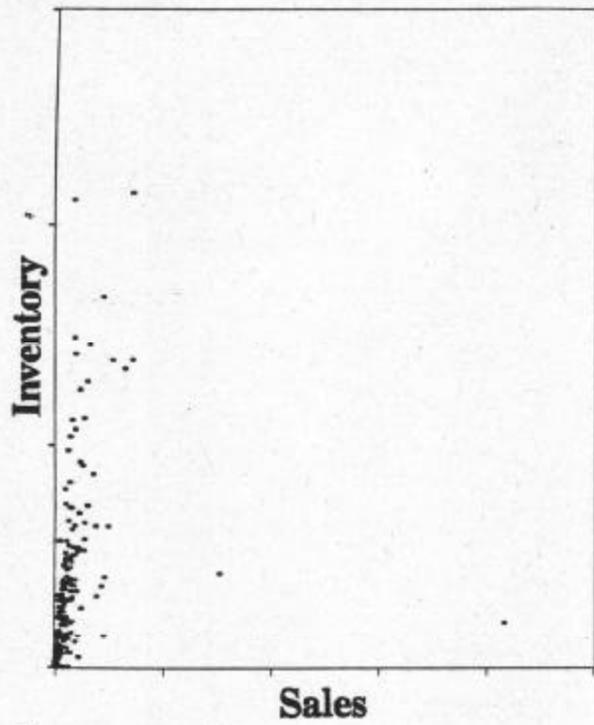


a

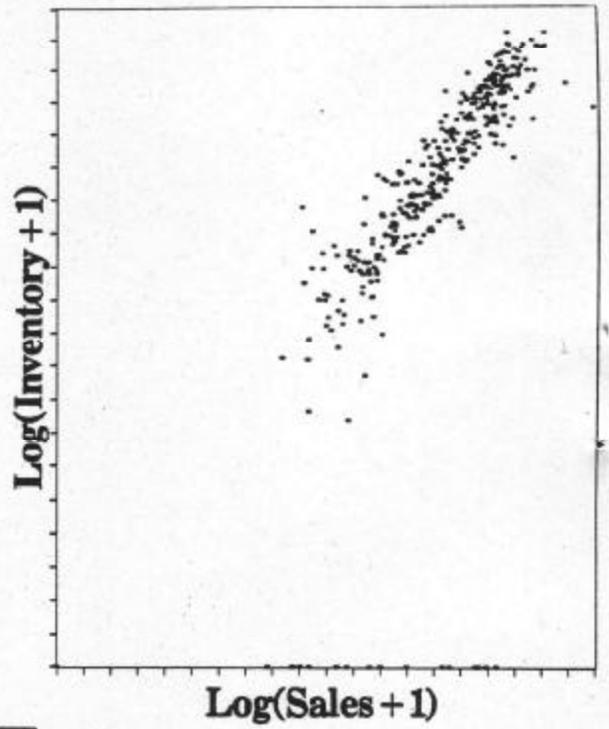
Figure 3



b

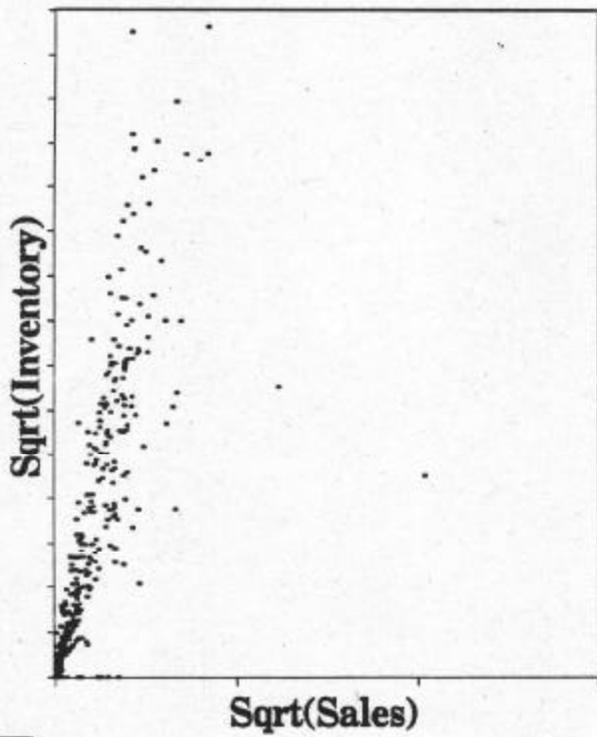


a

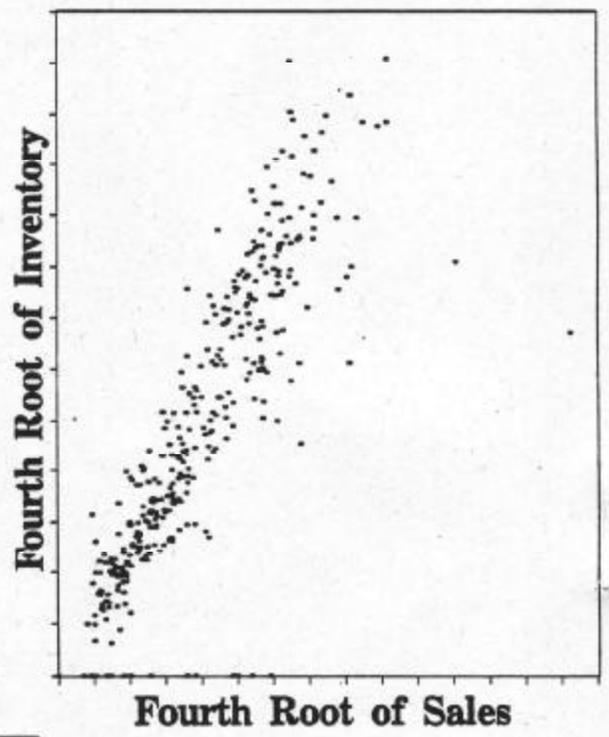


b

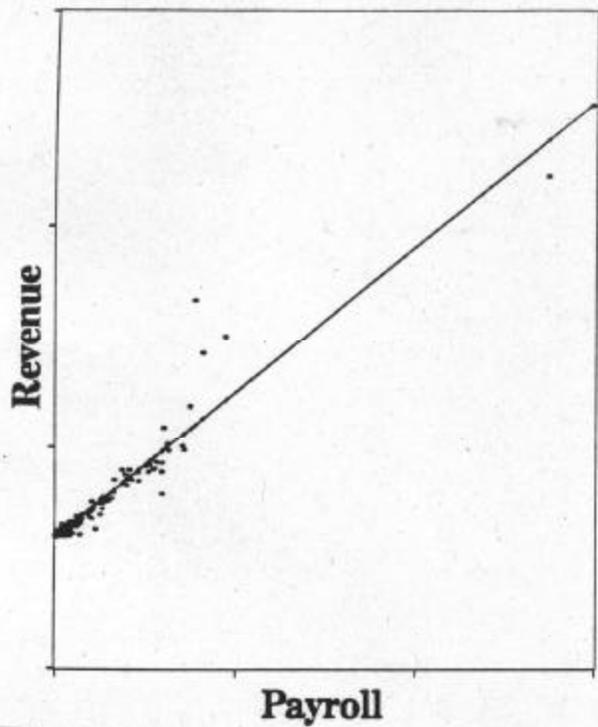
Figure 4



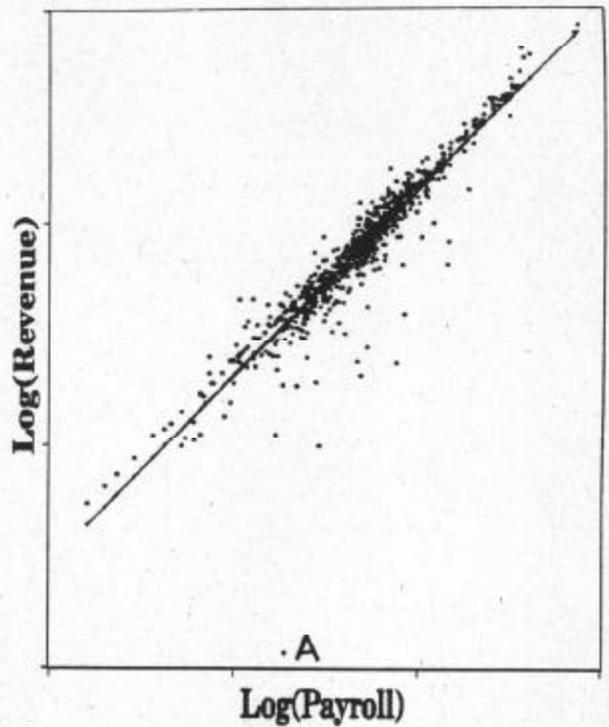
c



d

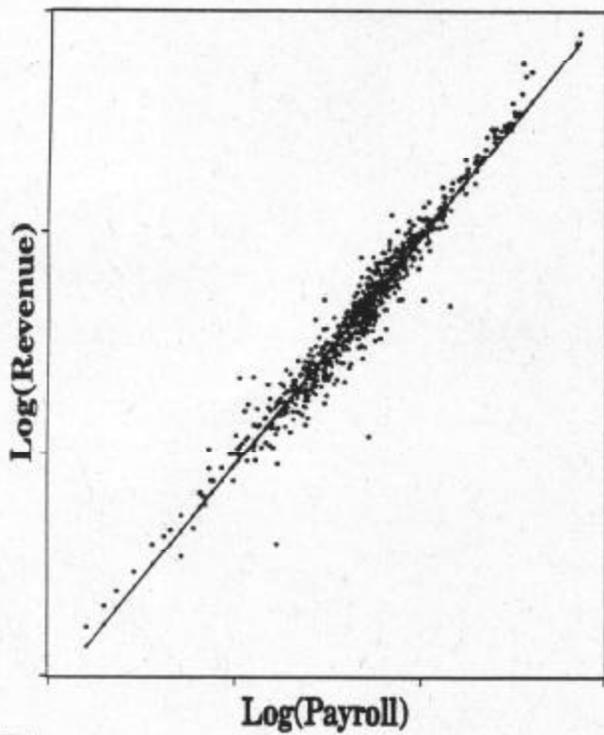


a

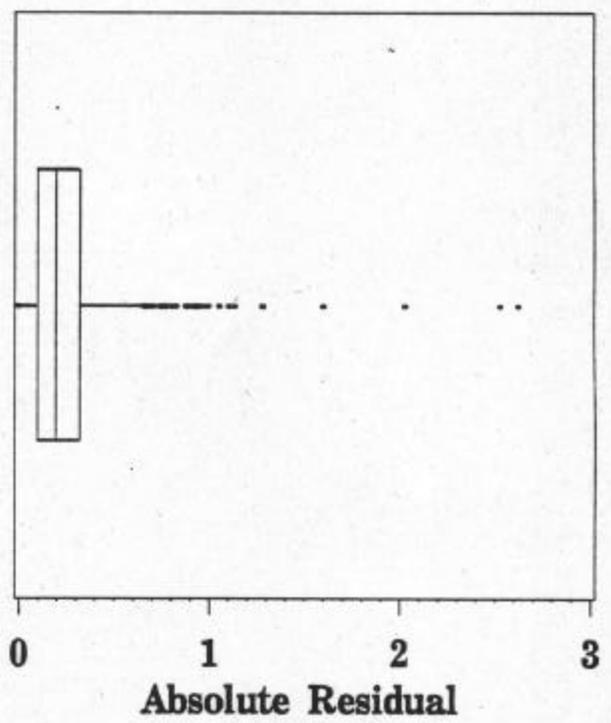


b

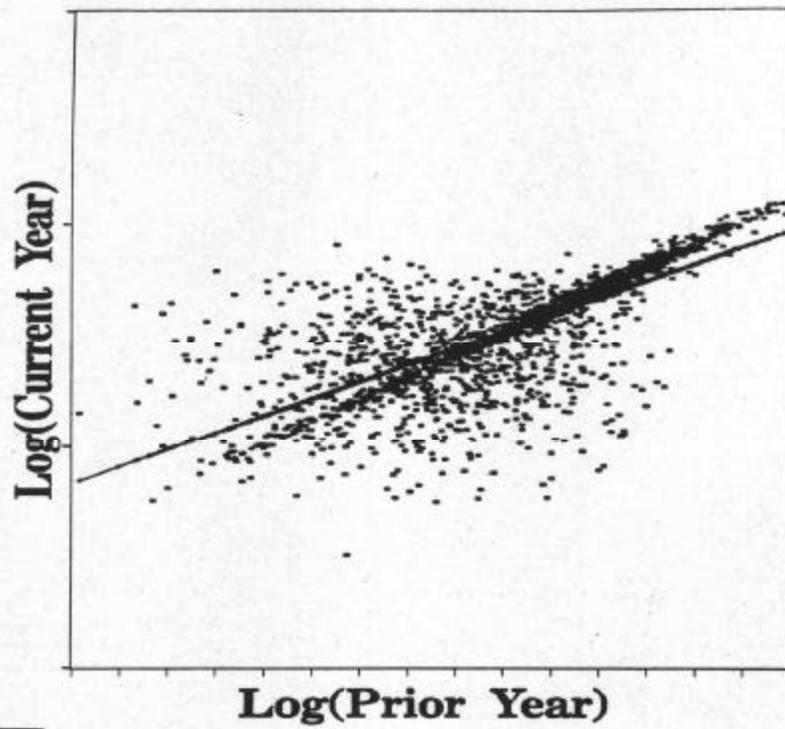
Figure 5



c

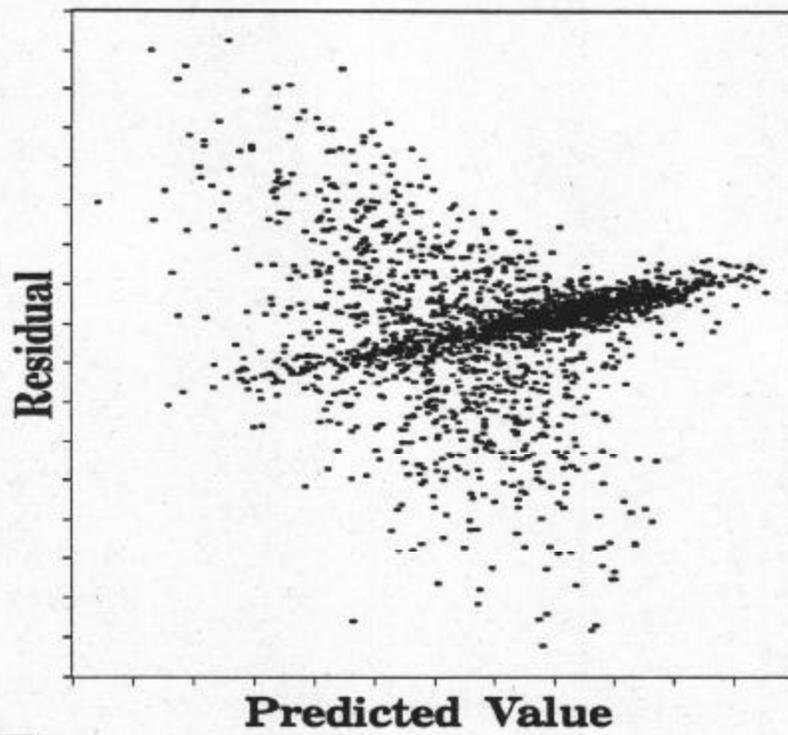


d

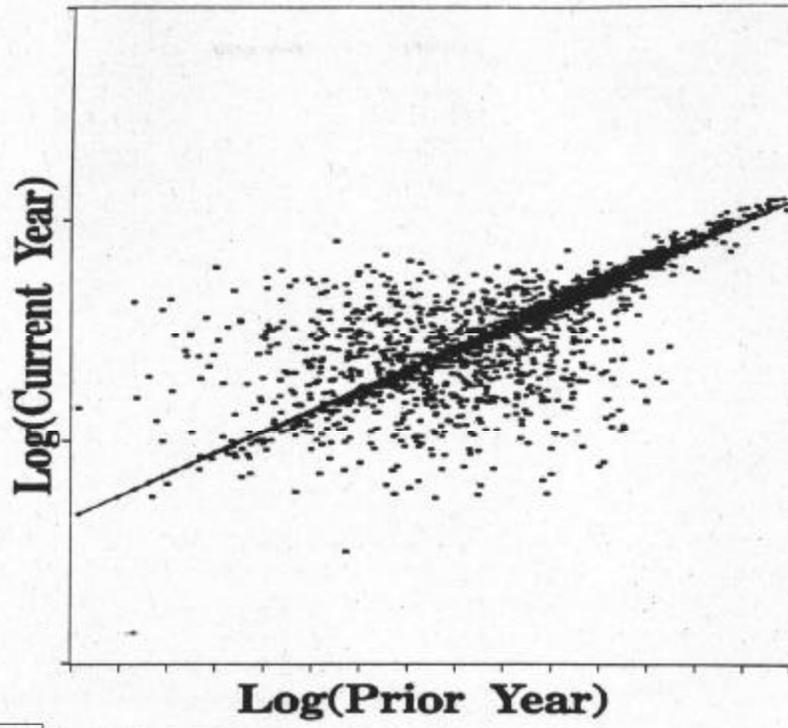


a

Figure 6

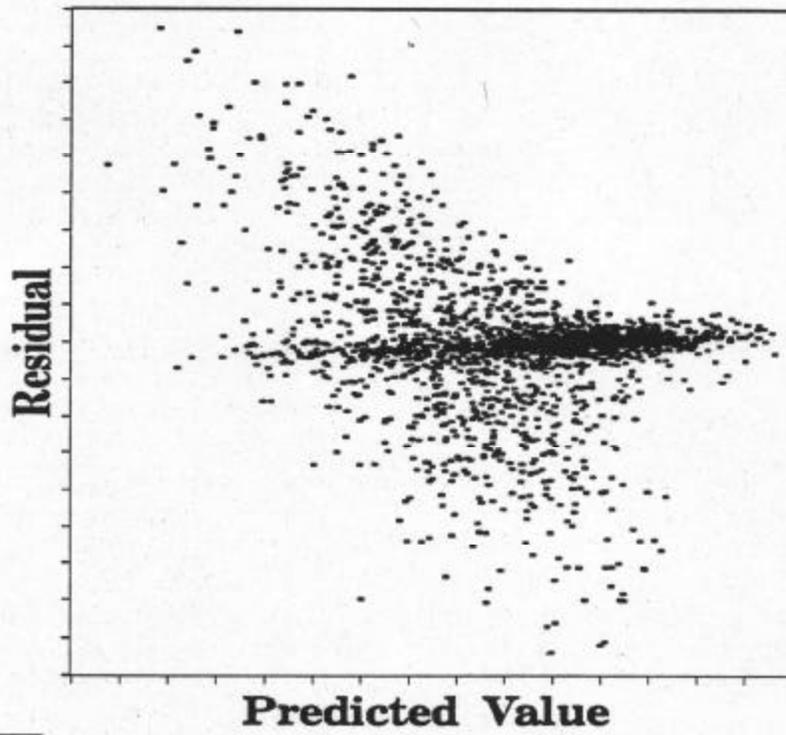


b



a

Figure 7



b

TIME SERIES AND CROSS SECTION EDITS WITH APPLICATIONS TO FEDERAL RESERVE DEPOSIT REPORTS

David A. Pierce and Laura Bauer Gillis¹
Federal Reserve Board

ABSTRACT

Currently data from the major deposit reports submitted by commercial banks to the Federal Reserve System are edited by comparing the incoming value for a variable to that variable's value for the previous week, using a set of published *tolerances*. The previous value represents an estimate or forecast of what the current value would be in the absence of error or unusual circumstance. This paper investigates two generalizations of this editing method, which both involve incorporating information beyond that contained in the previous week's value. One of these is to base this estimate on the item values from a *cross section* of similar institutions in the current time period which have already reported, and the other is to calculate a forecast based on the *time series* of past values of the item. A composite estimate combining these two methods is also presented. Edit simulations are performed to measure the improvement from this approach (in terms of fewer edit exceptions which are correct and/or increased detection of errors), which is found to be substantial for some items and size groups. Efforts thus far to implement these enhancements are described, and possible further generalizations are mentioned.

1. INTRODUCTION AND SUMMARY

Data for the U.S. Money Supply are regularly transmitted to the Federal Reserve System by commercial banks and other financial institutions at weekly and other intervals. A major vehicle for this transmission is the "Report of Transactions Accounts, Other Deposits and Vault Cash", or simply the "Report of Deposits", on which banks and other financial institutions report weekly data for 25 deposit categories and related items. Based on these data and on similar information contained in other reports, the money supply measures are constructed and reserve requirements are maintained.

The money and reserves figures are important both as barometers of economic activity and in enabling the Federal Reserve to perform its economic stabilization and bank regulatory functions, and it is essential that the data submitted on the Report of Deposits and other reports be reliable and of high quality. To ensure their accuracy, all such data are subjected to numerical edits to detect unusual or deviant values. These edits are to two general types, *validity* edits to ensure that adding-up and other logical constraints are satisfied, and *quality* edits based on statistical or distributional aspects of the data.

¹ The authors are respectively Senior Statistician and Statistician, Division of Research & Statistics, Federal Reserve Board, Washington, DC 20551. The valuable assistance of Mia Johnson is gratefully acknowledged. Any views expressed do not necessarily reflect those of the Federal Reserve System.

The most commonly used quality edit involves the comparison of an incoming weekly figure to the previous value of that variable (in both dollar and percentage terms), using a tolerance band constructed about that value. The *tolerances*, or half-widths of the tolerance bands, are determined from previous estimates of the variable's distribution, in particular measures of spread, and are published in a Technical Memorandum or "Tech Memo"². An edit "exception" occurs if the incoming value falls outside this tolerance band; when this happens, the reporting bank or other institution may be contacted for verification or correction. All tolerance-table comparisons are made (and edit exceptions generated) by machine, whereas the decision to contact the respondent is made by data analysts. The editing is done at both the Federal Reserve Board and the 12 Federal Reserve Banks.

Edits are in essence hypothesis tests, and errors of both kinds can occur. A major task in setting edit tolerances is to ensure adequate sensitivity without generating unnecessarily large quantities of "false positive" edit exceptions. It is because of the large number of exceptions currently generated that editing at both the Reserve Banks and the Board is currently quite labor intensive. All exceptions are reviewed by data analysts who must decide which are to be referred to the respondent institution for verification or revision. At the same time, a large majority of the data errors are not caught by these edits, based on the historical record of revisions submitted by respondents (they may be detected by other edits at a later date). There is consequently a need both to increase the sensitivity of the edits and to streamline the data editing process.

The previous value of the variable being edited, to which the tolerances are applied, in effect represents an estimate or forecast of the current figure in the absence of error or unusual circumstance. By basing this forecast or estimate on information beyond that contained in the previous week's value for that variable or item, we obtain the generalizations of the current editing method that are investigated in this paper. One generalization is to base this estimate on the item values from a *cross section* of similar institutions in the current time period which have already reported, intending to capture economic, institutional or calendar movements which tend to affect similar respondents in a similar manner. The other is to calculate a forecast based on the *time series* of past values of the item for that respondent, including possibly last month's or last year's figures in addition to the one for last week as in the current procedure. A composite estimate combining these two methods is also investigated, the idea being that each method may incorporate information not captured by the other. (We also generated a composite of the cross section and current edits).

The paper's focus is on the data submitted on the Report of Deposits, also known as the Edited Data Deposits System (EDDS) Report. We investigated four of the more important items on this report, total transactions deposits, total savings deposits, and large and small time deposits. The study was motivated by the desire for greater automation in the Federal Reserve Board's Division of Information Resources Management, which carries out the edits. The improvements resulting from the study are being incorporated into a new software package called DEEP (Distributed EDDS Editing Project), for interactive editing on the PC.

Our results vary greatly according to item, entity type (e.g. commercial bank, credit union, etc.), and the amount of data in an institution group -- the latter being important for reliable cross-

²"Processing Procedures for the Report of Transaction Accounts, Other Deposits and Vault Cash (FR2900), Technical Memorandum No. 16, Publications Section, Federal Reserve Board (December 1993).

section estimates. In most cases we find that, with sufficient data, the cross section approach is as reliable as the current editing procedure. For total transactions deposits almost uniformly, and for total savings deposits for most commercial bank categories, time series modelling plays a significant role in the edits.

The following section of the paper discusses in greater detail the methodology underlying the different data editing approaches investigated. Section 3 then describes a set of edit simulations we performed with each of the five types of edits studied, and presents the results of these. Based on the simulation results, we provided a set of recommendations for experimental edits for DEEP, for each entity type and item, which have recently become operational.

2. METHODOLOGY

Given a variable or item of interest, many data editing procedures can be characterized as first generating a forecast (a point estimate) of the incoming value for that item, next applying a tolerance to the forecast to form a tolerance interval (an interval estimate) for the incoming value, and then flagging that value if it is outside the tolerance interval. In the current editing framework, that forecast is taken to be the previous week's item value, and the tolerance is as given by the Tech Memo (footnote 2). In this section the two generalizations to the forecast noted in Section 1 are presented, along with composite procedures, after first describing the data and framework used.

2.1 Choice of Items and Statistical Form

The current approach to editing data from financial institutions is to subdivide them into homogeneous "cells", which are combinations of an institution's size group, entity type, geographic location. There are six size groups for commercial banks and a smaller number of size groups for each of the other entity types, which are credit unions, S&Ls, savings banks, agencies and branches of foreign banks, and Edge and Agreement Corporations. The geographic locations are defined in terms of 12 Federal Reserve districts.

There are thus a great many edit cells, and to make our task manageable, and to achieve comparability with the current edits, we have simplified this study in the following ways:

1. Staying with the *same cells* of the current EDDS edits. This will facilitate assessing the effects of the cross section estimates, model forecasts, and composite procedures. We recognize that more sophisticated groupings into cells may enhance the performance of the edits and plan to work with these in the future. Also we have eliminated all acquisitions and mergers from the institutions studied and have placed "credit-card banks" in a separate group.
2. Maintaining the *same tolerance widths* as currently (applied, however, to the time series / cross section estimates that we generate, as well as to the most recent value as currently done). This may at first seem unnecessary, since standard deviations, percentiles, and other aspects of the distribution can be determined from either the cross section data or the historical model. However, such calculations can sometimes be unreliable, especially with cross sections without at least several hundred institutions

in a group, as we are working with the extremes of distributions. And as with the cells themselves, keeping the current cell tolerance-interval widths facilitates comparisons among procedures.

We have also confined our attention in this study to the smaller institutions ("Priority-3" or P-3 institutions), where there may be the greatest potential for human resource savings from this approach. (Essentially this excludes the largest three size groups for commercial banks and a portion of the largest size group for other entity types). For these institutions, we have examined the following items:

Total transactions deposits	Large time deposits
Savings deposits	Small time deposits.

Current EDDS editing is performed with both dollar and percentage changes of the item being edited, with both required to exceed tolerances ("and" condition) for an exception to occur. The modifications outlined in this report are only for percentage changes; the Tech Memo tolerances continue to be applied to the dollar changes. There are several reasons for choosing percentage changes as the focus. Since they are used in current edits, the present edit cells and tolerances can be employed, and comparisons with current procedures can be made. They (or their annualized versions, growth rates) are also used in other analyses, such as with the Small Bank Sample of early reporting institutions. They are more homogeneous than dollar changes among different sized institutions, so that fewer edit groupings should eventually be needed. Percentage changes were found to be more sensitive to reporting and other errors than ratios to other items such as total deposits, which change with the denominator as well as the numerator and moreover present difficulty when the denominator was zero.

2.2 Cross Section Edits

Period-to-period edits compare an institution's current value for an item to the previous period's value. However, useful additional information may be contained in the current values of that item for other institutions that are similar to the one being edited. For example, if most of the institutions in a group experience a surge in large time deposits in a given week, then it would probably be inaccurate to list them as exceptions simply because they were outside the EDDS tolerances. Conversely, a very small change that week in large time deposits for a particular institution in that group may be suspicious even though current period-to-period tolerances would not be exceeded.

Cross section edits are carried out by examining the distribution of values (here, of percentage changes) for institutions within a homogeneous group, and listing as exceptions any values that were unusual compared to that distribution. Ordinarily one would calculate the mean and standard deviation of the percentage changes and flag those that were farther away from the mean than (say) two or three standard deviations; but in the present study we modified this set-up in two ways. First, because extreme values (the ones we hope to detect) would themselves influence the mean to which they would be compared, we "trimmed" the mean by eliminating the largest and smallest 5% of the values before calculating the estimated mean. Second, more observations are required to form a reliable estimate of the standard deviation than of the mean, and since most of the cells or groupings of institutions were too small for this, we chose to use multiples of the current EDDS tolerances as proxies for the standard deviations. As noted earlier, an additional advantage of this practice is to facilitate comparisons with the current edits.

One difficulty in using a cross section edit is that the data for an editing group need to be available in order to calculate such quantities as the average percentage change for that group. But the data for Priority-3 institutions are not due at the Board until nine days after the as-of date; and since timely estimates of the monetary aggregates and required reserves are needed, the editing process cannot be postponed this long. Our solution to this is to wait until a large enough fraction of the institutions have reported, and to form the distributional estimates (the trimmed means in this case) from the data available at that time.

For the EDDS data, more than half of the P-3 institutions' records are received by the Federal Reserve Board on the Thursday night following the as-of date (the previous Monday, on which the statement week ends), with the majority of those outstanding arriving by Friday night and the few remaining ones by the following Wednesday. For this study it was therefore decided to start the cross section editing on Friday morning, although work in progress is comparing this with the alternative of beginning on Monday morning. In either case, the trimmed mean estimates initially formed are not modified when more institutions have reported, in order not to confuse the editing process.

Some of the editing cells contain only a small number of respondents (and an even smaller number reporting by Friday), so that the estimated mean for those cells may not be very reliable. We required a minimum of 50 available observations in order to use the cross section estimate by itself. If the number of available observations is less than 50 but at least 20, a composite (see Sec. 2.4) of that estimate and the previous week's value for the institution is employed, and with less than 20 the previous week's value alone is used.

The cross section edit is performed by comparing the deviation between the observed and the estimated percentage changes to the current EDDS edit tolerance for the item. As noted earlier, if the percentage-change condition is violated, then a second comparison of the magnitude of the dollar change versus its tolerance is performed, and the item is flagged only if both sets of tolerances are exceeded. An exception to this is that, as is done with the current edits, when the item changes from zero to a nonzero value or vice versa, the current dollar-change edit tolerances are applied without any adjustment.

2.3 Time Series Edits

These edits are based on time series models, which predict or explain an item's present value in terms of its past history. This usually involves the immediately previous value, on which the current edits are based, and often additional values as well, such as last year's. To the extent that these more distant values are important in predicting the incoming value, more sensitive edits should result from taking them into account.

Editing using a time series model for generating forecasts of percentage changes implies that a historical relationship exists between the item and its previous values. The "random walk" model is a time series model in which the best forecast of the current value is simply last week's value. Thus, the random walk model is implied by the current period-to-period change edits, which take last week's value as the current-period forecast around which the tolerances are applied. More complicated time series models yield forecasts which are weighted averages of several past values of the percentage change.

We first investigated the fitting of time series models for each institution separately. Some

institutions' data fit the models quite well, with reductions in the standard deviation of the forecast errors (a key to the effectiveness of tolerances of a given width) of 50% or more, while other institutions exhibited only weak fits, or only the random walk behavior that the current editing framework already captures. Although fitting individual models is the preferable method for forecasting, it was not feasible to maintain over 8000 models for each item edited within the DEEP framework - at least not at this time. Thus, at this stage and for the P-3 institutions, a single time series model was fit to each editing cell's aggregate, and the coefficients from that estimated model were used to obtain an individual bank's forecast using its own previous values. While the benefits of time series modelling are reduced by doing this, the method can be easily implemented, and updated when necessary. Another constraint at present is that, because of data storage limitations, we only utilized terms in the model at lags of 1, 2, 3, 52 and 53 weeks, thus capturing nearby effects and annual seasonal influences but not, say, monthly or quarterly effects.

As an example of the model-fitting results, Table 1 provides information on time series models fit to cell aggregates of Total Transactions Deposits for three of the editing cells. Notice the highly statistically significant seasonal effect (lag 52, and in some cases lag 53). The strength of the fit declines going down the page, with the third one (Edges & Agreements, a root MSE reduction of only 9.2%) being not much different from the random walk model underlying current edits. On the other hand the results suggest that model-based editing may be valuable for certain commercial bank cells, for total transactions.

As with cross section edits, the deviation between the actual percentage change and the forecasted change from the time series model is compared to the edit tolerances. A tolerance exceedance both here and on the dollar change (also using current EDDS tolerances) triggers an edit exception for the record.

2.4 Composite Edits

The cross-section and time series edits are based on different sets of information, past values of the institution being edited and present values of similar institutions. Thus a forecast which combined these two estimates, thereby utilizing both sources of information, may be more accurate than either one separately, and edits derived from such forecasts correspondingly more sensitive.

For a given institution (e.g. bank) and a given item, if T denotes a time-series estimate (forecast) for a given week, C represents a cross-section estimate, and A the actual value that is reported, then the composite estimate is a weighted average of T and C which is of the form

$$\omega T + (1-\omega)C.$$

The weights ω and $1-\omega$ depend on the relative sizes and the correlation between the estimation / forecast errors of T and C. If these errors are given by

$$ET = A - T \text{ and } EC = A - C,$$

then

$$\omega = [\text{Var}(EC) - \text{Cov}(ET, EC)] / \text{Var}(ET - EC).$$

A composite forecast is thus a weighted average of individual component forecasts where the relative weights are chosen to minimize the sum of the squared forecast or estimation errors, and where the

sum of the weights is one.

Using past data, we investigated a composite estimate of the cross section and the time series forecasts, denoted "CSTS", for each editing cell and each item. The composite forecast defaults to the time series forecast with fewer than 20 available observations in the cell average. (With exactly 20 and using the 90% trim, 18 observed changes would be used in the cell estimate).

The other type of composite edit we considered combines the cross section and the random walk forecasts (CSRW). We employed this edit when a CS edit was indicated but the sample size - the number of observations available on Friday morning when the cell means are formed -- was insufficient (less than 50) to obtain an adequately reliable cross section estimate. For very small sample sizes (less than 20), our procedure is to revert to the use of only the RW edit.

3. MODELLING AND SIMULATIONS

To examine the relative performance of different types of edits, we conducted simulations of these edits over the 1991-92 time period. For each cell (choice of item, entity type, size group and geographic region), we performed five sets of simulations, corresponding to the different types of edits under consideration: current (random walk), cross section, time series, cross section/time series composite, and cross section/random walk composite.

3.1 *Simulation Procedure*

Data preparation was a time consuming task. First, all Priority-3 reporters' weekly average data were compiled for the period from January 1986 through December 1992. While the edits were simulated only for the most recent two years, the additional data were used for fitting time series models with potential annual patterns. To avoid distortions, we eliminated all banks involved in mergers during this period. We next partitioned the data set into the editing groups or cells. We found that not all cells had a sufficient number of reporters to fit a model or to obtain reliable cross section estimates, and so some of them were combined. For commercial banks of size group 3 (total deposits between \$1B+ and \$3B), there were too few P-3 reporters to employ any of the new approaches. In addition, we added an editing category for known credit card banks. In total there were 40 edit cells, 37 of which were involved in the simulations.

Once the data were prepared, time series models were fit to the percentage changes in each cell's aggregate, as described in Section 2.3. Using the fitted model for a cell, predicted values for the last two years were generated for each institution in the cell. (Although forecasted values of the percentage change were generated for all periods, those in which a change of zero to a value or a value to zero were edited using the current special tolerances). Both the model-based and the zero-valued random walk forecasts were assigned to each observation in the cell. The 10% trimmed mean of the percentage changes was also calculated for each cell and each week of the two year simulation period, for use in the cross section edits. (Since the cross section simulation employed all the data within a cell to calculate the current-period forecast, rather than the available data as of Friday morning when editing begins, the simulated results will differ from those in practice). In order to generate the two composite forecasts, the prediction errors from the original three forecasts were computed and the formulas in Section 2.4 applied by institution. A cell root mean square prediction

error (RMSE) was also computed.

Since the composite forecast combines the component forecasts in such a way as to minimize the sum of the squared prediction errors, we chose to estimate the appropriate weights for each bank in a cell and then to average those weights over the cell in order to obtain the composite for editing. Since the composite is a weighted average of the individual forecasts, the sum of the weights must equal one. For some institutions, where the prediction errors were very highly correlated between methods, we obtained pairs of weights with one value less than zero and the other greater than one. Evidently it only requires a small number of observations away from that correlation structure to cause such disproportionate weights. In calculating the average pair of composite weights for each cell, therefore, we first screened out those sets of weights not within the (0,1) range. After the two composite weighting schemes were determined for each cell, the mean square prediction errors were computed for these two forecasts as well.

For each of the five edit methods, Table 2 presents the root mean square prediction errors and composite weights for the commercial bank cells for total transactions and total savings, and Table 3 presents the same information for the other entity cells, for total transactions. We anticipate the method with the smallest forecasting error to have the best potential as an edit, but until our tolerances are better tailored to the actual editing method, this potential may not be realized.

To apply the edits, we first looked for percentage changes that differed from the forecasted percentage changes by more than the appropriate tolerance (whether taken from the Tech Memo or generated as described in this paragraph), and for those ascertaining whether the dollar change tolerance was also exceeded. Since total savings and large time deposits are currently edited items, their current tolerances can be used. However, for total transactions and small time deposits, current tolerances do not exist. We therefore generated tolerances in a manner similar to that used for the creation of the current ones. This involved iterative steps with the intent of flagging approximately 0.3% of the observations per cell on average (the maximum percentage of observations flagged using current editing methods for other items, for the year 1991). Using the components of total transactions and items that were related to small time, such as total and large time, we first compiled a range of feasible values for the tolerances. We then examined where these values occurred on the distribution of percentage changes over each cell for the two-year period. Given a reasonable proportion of the changes exceeding the initial values, we then examined the dollar change distribution for the subset of percentage change exceptions. Appropriate percentiles of this distribution were then determined to obtain the expected 0.3% edit failures under the current random walk model. These percentiles became the dollar change tolerances.

Once all the forecasts and tolerances were in place, the editing experience for the 1991-92 period was simulated for each of the five forecast methods. For each method we observed which observations were flagged as edit exceptions. Then based on a history of weekly revisions to the EDDS file maintained by the Federal Reserve's Statistical Services branch, we were able to determine the rate of type I and type II errors for each method. [A type I error (a "false positive") refers to an item that was flagged but not in error, or at least not revised. A type II error occurs when an item is not flagged but is erroneous (as evidenced by a later revision)].

3.2 Simulation Results

For reference in this section, Table 4 shows our recommended edits based on these simulations. As mentioned in Section 1, these are currently being implemented as part of the Federal Reserve Board's DEEP editing software. In Table 4, the left column lists the entities (with the included size groups in parentheses), followed by the chosen edit for each item.

Turning to the results on which this table is based, Table 5 summarizes the editing simulations for commercial banks; those for other entity types were similar and are given in an earlier report³. To assess the magnitudes and the implications of errors caught and errors missed by the editing schemes, the tables break down these errors in terms of their size (i.e. the size of the revision—we assume, however accurately, that revised data are correct and the revision is the error in the unrevised data). Each section of these tables compares the current (random walk) method with an alternative editing strategy. It is clear from these simulations that there is room for improvement, especially regarding the type II error probabilities, which range from 98% to 99%. And although the type I error probabilities appear small, the number of flagged items that are not in error is quite large (between 87% and 94%).

Wherever the fitted time series model indicated a potentially substantial payoff relative to the random walk model (as in the first model in Table 1), the time series edit tended to be the most accurate, yielding the smallest number of edit exceptions and with fewer errors missed that were captured by other methods than vice versa. The reduction in the number of edit exceptions was not as great for the CS and CSTS composite methods, but often the composite method caused less of an increase in the type II error probability. The CS and the CSRW composite often mimicked the current RW results. Where there was doubt regarding the preferable edit method, we tended to favor the CS or CSRW -- even when the reduction in RMSE and the number of edit exceptions was small relative to the current (RW) method -- since cross section edits would allow possibly large shifts in behavior for a given week to be incorporated into the editing norm, and the DEEP software is well-suited to this type of edit. Also, we gave some preference to a uniformity of editing method across related cells (e.g. adjacent size groups within an FR region, or like size groups between regions).

For commercial banks, the alternative edits on the whole did quite well. The time series edits for total transactions and total savings were effective in reducing the total number of exceptions while missing only 3 small revisions and actually finding an additional error of over \$25M.⁴ For the other entity types, total transactions was the only item that allowed for an alternative other than the CSRW method (CSRW was selected for these entity types in place of CS in order to accommodate smaller sample sizes in the preliminary data). Those credit unions and savings institutions which would have more activity in transactions accounts than the other entity types, do exhibit cyclical patterns which the time series model was able to capture (See Table 3.A). Agencies and branches also exhibited

³ "Editing in DEEP: Utilizing Time Series and Cross Section Information", Laura Bauer Gillis and David A. Pierce, Federal Reserve Board, 1993 (preliminary report). Available from the authors.

⁴ This revision was generated either by an outside source or by an edit of another report that is not being considered here. This occurrence brings to light that some errors are detected by other sources - not the Reserve Banks or the Board. What we gain from this additional edit exception an earlier detection of the error; it would not necessarily go undetected permanently.

improved editing results with the CSTS method. As mentioned, this combination of alternative strategies yielded an 11% reduction in both the type I error probability and the number of edit exceptions, with only a very slight increase in the type II error likelihood (about 0.1%).

All of these results are based on simulations using 1991 and 1992 EDDS data. Any errors caught before the data arrived at the Board are not reflected in these data, nor are errors undetected by Banks or Board that do not show up in the revision files. And as previously mentioned, the other factor to be monitored is the use of preliminary data in cross section estimates of the mean percentage change. Depending on how and where the preliminary data fall in the distribution of all percentage changes for an item, the operational results based on the CS, CSTS, or CSRW methods may differ significantly from what is expected based on the simulation results. The data availability and timing issue for cross section estimates is currently being studied.

This investigation is still in progress, and further generalizations of the work are underway or planned. Among these are examining time series models with regression components to account for such phenomena as tax dates, calendar effects or related variables, alternative groupings of the data according to size or geographic region, modelling larger banks individually, and examining additional items or variables.

**Table 1. Percentage Change Models for Total Transactions Aggregates,
Selected Editing Cells**

-----Cell = CB, Size Group 4, Region I-----

Root MSE(orig.) = 0.0383 Root MSE(model) = 0.0211
Reduction in Root MSE = 44.9%

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN _{t-1}	-0.4349	0.0483	-9.005	0.0001
TRN _{t-2}	-0.0341	0.0329	-1.039	0.2996
TRN _{t-3}	-0.1510	0.0338	-4.467	0.0001
TRN _{t-52}	0.6494	0.0318	20.391	0.0001
TRN _{t-53}	0.4668	0.0440	10.606	0.0001

-----Cell = CU, Size Group 2, Regions II&III-----

Root MSE(orig.) = 0.1067 Root MSE(model)=0.0809
Reduction in Root MSE = 24.2%

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN _{t-1}	-0.2450	0.0546	-4.486	0.0001
TRN _{t-2}	-0.1160	0.0474	-2.444	0.0151
TRN _{t-3}	-0.2200	0.0486	-4.525	0.0001
TRN _{t-52}	0.4922	0.0477	10.312	0.0001
TRN _{t-53}	0.1866	0.0533	3.498	0.0005

-----Cell = EA, All-----

Root MSE(orig.) = 0.0564 Root MSE(model)=0.0512
Reduction in Root MSE = 9.2%

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN _{t-1}	-0.3776	0.0569	-6.632	0.0001
TRN _{t-2}	-0.1547	0.0586	-2.642	0.0087
TRN _{t-3}	-0.0449	0.0553	-0.815	0.4181
TRN _{t-52}	0.2432	0.0524	4.638	0.0001
TRN _{t-53}	0.1057	0.0540	1.955	0.0514

Table 2. Root Mean Square Errors for Forecasts: Commercial Bank Cells

<i>A. Total Transactions</i>		Root Mean Square Error					Weight of CS in Composite		
<u>Cell</u>	↓	<u>RW</u>	<u>CS</u>	<u>TS</u>	<u>CSRW</u>	<u>CSTS</u>	↓	<u>CSRW</u>	<u>CSTS</u>
Region 1									
-Size 4		0.077	0.073	0.077	0.074	0.071		0.72	0.51
-Size 5		0.096	0.094	0.097	0.094	0.089		0.73	0.55
-Size 6		1.190	1.190	1.276	1.190	1.204		0.70	0.58
Region 2									
-Size 4		0.064	0.059	0.236	0.060	0.121		0.77	0.58
-Size 5		0.210	0.209	0.223	0.209	0.212		0.62	0.55
-Size 6		0.331	0.330	0.344	0.330	0.333		0.68	0.57
Region 3									
-Size 4		0.102	0.099	0.108	0.100	0.100		0.75	0.51
-Size 5		0.054	0.048	0.051	0.050	0.046		0.74	0.58
-Size 6		0.067	0.063	0.071	0.064	0.062		0.70	0.60
<i>B. Total Savings</i>		Root Mean Square Error					Weight of CS in Composite		
<u>Cell</u>	↓	<u>RW</u>	<u>CS</u>	<u>TS</u>	<u>CSRW</u>	<u>CSTS</u>	↓	<u>CSRW</u>	<u>CSTS</u>
Region 1									
-Size 4		0.042	0.042	0.045	0.042	0.042		0.64	0.73
-Size 5		0.054	0.054	0.056	0.054	0.054		0.64	0.67
-Size 6		0.048	0.048	0.055	0.048	0.048		0.60	0.72
Region 2									
-Size 4		0.038	0.038	0.099	0.038	0.043		0.65	0.76
-Size 5		0.235	0.234	0.244	0.234	0.236		0.64	0.64
-Size 6		0.055	0.055	0.067	0.055	0.055		0.64	0.66
Region 3									
-Size 4		0.051	0.051	0.998	0.051	0.274		0.68	0.74
-Size 5		0.041	0.040	0.041	0.040	0.040		0.63	0.66
-Size 6		0.055	0.055	0.065	0.055	0.055		0.61	0.75

Table 2. Root Mean Square Errors for Forecasts: Commercial Bank Cells (Continued)

<i>C. Large Time</i>		Root Mean Square Error					Weight of CS in Composite		
<u>Cell</u>	↓	<u>RW</u>	<u>CS</u>	<u>TS</u>	<u>CSRW</u>	<u>CSTS</u>	↓	<u>CSRW</u>	<u>CSTS</u>
Region 1									
-Size 4		0.067	0.067	0.069	0.067	0.067		0.53	0.61
-Size 5		0.110	0.110	0.117	0.110	0.110		0.52	0.75
-Size 6		0.160	0.160	0.184	0.160	0.161		0.49	0.81
Region 2									
-Size 4		0.089	0.088	0.093	0.088	0.089		0.54	0.68
-Size 5		0.063	0.063	0.065	0.063	0.063		0.48	0.62
-Size 6		0.099	0.099	0.109	0.099	0.100		0.46	0.76
Region 3									
-Size 4		0.047	0.047	0.051	0.047	0.047		0.55	0.70
-Size 5		0.075	0.075	0.076	0.075	0.076		0.54	0.59
-Size 6		0.120	0.120	0.141	0.120	0.120		0.51	0.80
<i>D. Small Time</i>		Root Mean Square Error					Weight of CS in Composite		
<u>Cell</u>	↓	<u>RW</u>	<u>CS</u>	<u>TS</u>	<u>CSRW</u>	<u>CSTS</u>	↓	<u>CSRW</u>	<u>CSTS</u>
Region 1									
-Size 4		0.064	0.064	0.098	0.064	0.068		0.58	0.70
-Size 5		0.143	0.143	0.156	0.143	0.144		0.52	0.70
-Size 6		0.110	0.110	2.274	0.110	0.409		0.55	0.83
Region 2									
-Size 4		0.468	0.468	0.516	0.468	0.470		0.59	0.79
-Size 5		1.363	1.363	1.420	1.363	1.369		0.61	0.68
-Size 6		0.034	0.034	0.036	0.034	0.034		0.58	0.77
Region 3									
-Size 4		0.062	0.062	0.068	0.062	0.063		0.61	0.67
-Size 5		0.017	0.017	0.018	0.017	0.017		0.58	0.65
-Size 6		0.063	0.063	0.072	0.063	0.063		0.57	0.80

Table 3. Root Mean Square Errors for Forecasts: Other Entity Types, Total Transactions

A. Agencies and Branches

<u>Cell</u>	<u>Root Mean Square Error</u>						<u>Weight of CS in Composite</u>	
	<u>RW</u>	<u>CS</u>	<u>TS</u>	<u>CSRW</u>	<u>CSTS</u>	<u>CSRW</u>	<u>CSTS</u>	
-All Regions Size 1	1.366	1.363	1.378	1.364	1.364	0.48	0.54	
-Region 1 Size 2	2.700	2.696	2.794	2.698	2.715	0.45	0.53	
Size 3	5.061	5.061	5.974	5.061	5.198	0.38	0.62	
-Region 2 Size 2	2.248	2.240	2.406	2.245	2.317	0.38	0.35	
Size 3	4.158	4.154	4.965	4.156	4.248	0.43	0.67	
-Region 3 Size 2	0.250	0.247	0.250	0.248	0.243	0.34	0.44	
Size 3	4.300	4.289	5.416	4.295	4.544	0.42	0.58	

B. Credit Unions

<u>Cell</u>	<u>Root Mean Square Error</u>						<u>Weight of CS in Composite</u>	
	<u>RW</u>	<u>CS</u>	<u>TS</u>	<u>CSRW</u>	<u>CSTS</u>	<u>CSRW</u>	<u>CSTS</u>	
-All Regions Size 1	0.106	0.073	0.075	0.076	0.069	0.75	0.53	
-Region 1 Size 2	0.093	0.069	0.059	0.075	0.059	0.57	0.37	
Size 3	0.122	0.112	0.120	0.114	0.110	0.59	0.43	
Size 4	0.084	0.075	0.078	0.077	0.073	0.57	0.45	
-Regions 2 & 3 Size 2	0.084	0.062	0.054	0.065	0.052	0.64	0.57	
Size 3	0.099	0.080	0.078	0.083	0.073	0.63	0.37	
Size 4	0.082	0.069	0.060	0.072	0.059	0.56	0.40	

**Table 3. Root Mean Square Errors for Forecasts: Other Entity Types,
Total Transactions (Continued)**

C. Edges and Agreements

<u>Cell</u>		Root Mean Square Error						Weight of CS in Composite	
		<u>RW</u>	<u>CS</u>	<u>TS</u>	<u>CSRW</u>	<u>CSTS</u>		<u>CSRW</u>	<u>CSTS</u>
-ALL		17.21	17.20	18.94	17.20	17.72		0.44	0.47

D. Savings Institutions

<u>Cell</u>		Root Mean Square Error						Weight of CS in Composite	
		<u>RW</u>	<u>CS</u>	<u>TS</u>	<u>CSRW</u>	<u>CSTS</u>		<u>CSRW</u>	<u>CSTS</u>
-Region 1									
Size 1		0.057	0.048	0.049	0.049	0.047		0.79	0.67
Size 2		0.187	0.185	0.193	0.185	0.185		0.74	0.57
Size 3		0.744	0.743	0.965	0.743	0.779		0.71	0.61
Size 4		0.627	0.626	0.645	0.626	0.627		0.66	0.63
-Regions 2 & 3									
Size 1		0.073	0.065	0.068	0.065	0.064		0.73	0.68
-Region 2									
Size 2		0.132	0.129	0.153	0.129	0.131		0.73	0.62
Size 3		0.077	0.072	0.079	0.073	0.072		0.78	0.66
Size 4		0.066	0.062	0.069	0.062	0.061		0.75	0.66
-Region 3									
Size 2		0.077	0.069	0.077	0.070	0.068		0.78	0.58
Size 3		0.309	0.308	0.766	0.308	0.377		0.72	0.66
Size 4		0.370	0.369	0.568	0.369	0.408		0.63	0.59
-Region 4									
Size 1		10.73	10.73	11.44	10.73	10.77		0.67	0.75

Table 4. Experimental Edits for DEEP

	Total Transactions	Total Savings	Large Time	Small Time
Commercial Banks (3,Ccd)	RW	RW	RW	RW
Commercial Banks (4,5,6)	CSTS	TS	CS	CS
Credit Unions (1,2,3,4)	TS	CSRW	CSRW	CSRW
S&Ls, Coops, Sbs (1,2,3,4)	$\frac{\text{RI}}{\text{TS}}$ $\frac{\text{RII-IV}}{\text{CSTS}}$	CSRW	CSRW	CSRW
Agencies & Brs.(1,2,3)	CSTS	CSRW	CSRW	CSRW
Edges & Agr. (1,2)	CSRW	CSRW	CSRW	CSRW

The numbers in parentheses are the size groups, with "Ccd" denoting credit card banks. CB size groups 1 and 2 are omitted, as they are priority 1 and 2 institutions. RI denotes the FR Region, as in TM#16. The other entries in this table have the following explanations:

TS: The time-series model-based forecast, utilizing the institution's past percentage changes (of 1,2,3,52 and 53 weeks ago).

CS: The cross-section forecast, or estimate of the average percentage change over all the institutions in the editing group or cell. Uses only the data received by the Friday after the as-of date and is calculated as the 90% trimmed mean of the individual percentage changes in the cell.

CSTS: A weighted average of the TS and CS percentage-change forecasts, with statistically determined weights. When the number (n) of institutions in the group available on Friday for calculating the mean is less than 20, the weights are 1 and 0 (only the TS forecast is used).

RW: The forecast based on the "random walk" model, or the time series model giving a zero period-to-period change as the best forecast – and is thus the implicit model underlying the current edits. This translates into a percentage-change forecast of zero.

CSRW: The forecast based on a composite of the CS and RW estimates of the percentage change, again depending on the number n of available observations in the cell. Thus:
 if $n \geq 50$, use CS only;
 if $20 \leq n < 50$, use weighted average of the CS and RW estimates;
 if $n < 20$, use the RW estimate (zero percentage change forecast).

Table 5. Editing Simulation Results: Commercial Banks

A. Total Transactions

1. Random Walk (Current Editing)

Frequency/ Percent	Not Revised	<\$5M	\$5M <\$10M	\$10M < \$25M	> \$25M	Total
Not Flagged	557,166 97.76	9,732 1.71	791 0.14	508 0.09	168 0.03	568,365 99.73
Flagged	1,444 0.25	75 0.01	17 0.00	12 0.00	6 0.00	1,554 0.27
Total	558,610 98.01	9,807 1.72	808 0.14	520 0.09	174 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.26%
 Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.0%
 Pr(Item not in error | Item Flagged) = 92.9%

2. Cross Section - Time Series Composite

Frequency/ Percent	Not Revised	<\$5M	\$5M <\$10M	\$10M < \$25M	> \$25M	Total
Not Flagged	557,326 97.78	9,743 1.71	792 0.14	509 0.09	167 0.03	568,537 99.76
Flagged	1,284 0.23	61 0.01	16 0.00	11 0.00	7 0.00	1,382 0.24
Total	558,610 98.01	9,807 1.72	808 0.14	520 0.09	174 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.23%
 Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.1%
 Pr(Item not in error | Item Flagged) = 92.9%

Reduction in edit exceptions = 11.1%
 Reduction in type I error probability = 11.5%
 Increase in type II error probability = 0.1%

Table 5. Editing Simulation Results: Commercial Banks (Continued)

B. Total Savings

1. Random Walk (Current Editing)

Frequency/ Percent	Not Revised	<\$5M	\$5M <\$10M	\$10M < \$25M	> \$25M	Total
Not Flagged	557,547 97.83	8,772 1.54	723 0.13	375 0.07	181 0.03	567,598 99.59
Flagged	2,176 0.38	91 0.02	22 0.00	18 0.00	14 0.00	2,321 0.41
Total	559,723 98.21	8,863 1.56	745 0.13	393 0.07	195 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.39%

Pr(type II error) = Pr(Do not Flag Item | Item in error) = 98.6%

Pr(Item not in error | Item Flagged) = 93.8%

2. Time Series

Frequency Percent	Not Revised	<\$5M	\$5M <\$10M	\$10M < \$25M	> \$25M	Total
Not Flagged	557,743 97.86	8,775 1.54	723 0.13	376 0.07	181 0.03	567,798 99.63
Flagged	1,980 0.35	88 0.02	22 0.00	17 0.00	14 0.00	2,121 0.37
Total	559,723 98.21	8,863 1.56	745 0.13	393 0.07	195 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.35%

Pr(type II error) = Pr(Do not Flag Item | Item in error) = 98.6%

Pr(Item not in error | Item Flagged) = 93.4%

Reduction in edit exceptions = 9.8%

Reduction in type I error probability = 10.2%

Increase in type II error probability = 0.0%

Table 5. Editing Simulation Results: Commercial Banks (Continued)

C. Large Time

1. Random Walk (Current Editing)

Frequency/ Percent	Not Revised	<\$5M	\$5M <\$10M	\$10M < \$25M	> \$25M	Total
Not Flagged	558,956 98.08	8,494 1.49	601 0.10	345 0.06	179 0.03	568,575 99.76
Flagged	1,248 0.22	68 0.01	19 0.00	8 0.00	1 0.00	1,344 0.24
Total	560,204 98.30	8,562 1.50	620 0.10	353 0.06	180 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.22%

Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.0%

Pr(Item not in error | Item Flagged) = 92.8%

2. Cross Section

Frequency Percent	Not Revised	<\$5M	\$5M <\$10M	\$10M < \$25M	> \$25M	Total
Not Flagged	558,967 98.08	8,494 1.49	601 0.10	345 0.06	179 0.03	568,586 99.76
Flagged	1,237 0.22	68 0.01	19 0.00	8 0.00	1 0.00	1,333 0.24
Total	560,204 98.30	8,562 1.50	620 0.10	353 0.06	180 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.22%

Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.0%

Pr(Item not in error | Item Flagged) = 92.8%

Reduction in edit exceptions = 0.8%

Reduction in type I error probability = 0.0%

Increase in type II error probability = 0.0%

Table 5. Editing Simulation Results: Commercial Banks (Continued)

D. Small Time

1. Random Walk (Current Editing)

Frequency/ Percent	Not Revised	<\$5M	\$5M <\$10M	\$10M < \$25M	> \$25M	Total
Not Flagged	556,637 97.67	9,869 1.73	1,007 0.18	479 0.08	215 0.04	568,210 99.70
Flagged	1,496 0.26	117 0.02	42 0.01	36 0.01	18 0.00	1,709 0.30
Total	558,138 97.93	9,986 1.75	1049 0.18	515 0.09	233 0.05	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.27%

Pr(type II error) = Pr(Do not Flag Item | Item in error) = 98.2%

Pr(Item not in error | Item Flagged) = 87.5%

2. Cross Section

Frequency Percent	Not Revised	<\$5M	\$5M <\$10M	\$10M < \$25M	> \$25M	Total
Not Flagged	556,637 97.67	9,869 1.73	1,008 0.18	479 0.08	215 0.04	568,211 99.70
Flagged	1,496 0.26	117 0.02	41 0.01	36 0.01	18 0.00	1,708 0.30
Total	558,138 97.93	9,986 1.75	1049 0.18	515 0.09	233 0.05	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.27%

Pr(type II error) = Pr(Do not Flag Item | Item in error) = 98.2%

Pr(Item not in error | Item Flagged) = 87.6%

Reduction in edit exceptions = 0.0%

Reduction in type I error probability = 0.0%

Increase in type II error probability = 0.0%

DISCUSSION

Sandra A. West
U.S. Bureau of Labor Statistics

Let me first commend both sets of authors on very interesting and informative papers. Let me start with the David Pierce and Laura Bauer Gillis paper, "Time Series and Cross Section Edits with Applications to Federal Reserve Deposit Reports."

I enjoyed this paper very much; it was nice to see editing formally enter the realm of statistical inference. One might think of Imputation as point estimation, and editing as interval estimation--or perhaps as multiple imputation. I'd like to focus on one of the editing techniques in terms of imputation, but first let me briefly summarize the study.

In the paper there are:

5 methods for editing percent changes

1. Assuming no change from last week--current method-random walk, **RW**,---would be called Carry Over in nonresponse literature.
2. Using a cross section, **CS**, of similar respondents in the current time period which have already reported. Underlying assumption here that the previous time period values are available. (For surveys that do have nonresponse, only those entities that have reported in both time periods would be used.)
3. Using a time series, **TS**, of the past values of the respondent.
4. Composite of 1 & 2, **CSRW**.
5. Composite of 2 & 3, **CSTS**.

Several entity types-Respondents

Commercial Banks (There were two categories of this type.)

Agencies and Branches

Credit Unions

Edges and Agreements

Savings Institutions

Although there are 25 variables collected, the following 4 were studied.

Variables Collected from each Respondent

Total Transactions

Total Saving

Large Time

Small Time

Edits are performed weekly for a span of two years

Edits are performed in homogeneous cells which are combinations respondents' size, type and geographic location.

I'd like to discuss the cross sectional estimator, CS, since I've had some experience with a similar one using BLS data in terms of imputation. First, I need some notation. In a given cell, let

$Y_{t,i}$ = level of item for entity i at time t .

$\hat{Y}_{t,i}$ = predicted level for entity i at time t .

Editing is performed with percentage changes of the item; that is,

$$D_i = \frac{Y_{t,i} - Y_{(t-1),i}}{Y_{(t-1),i}}$$

(and for the current method the actual changes are also required to be in the tolerance interval).

For the CS edits, the empirical distribution is formed for the percentage changes, and the trimmed mean is computed, where 5% of each tail is trimmed. (Later I will say something about the trimmed mean.) Multiples of the current tolerances are used for proxies for the standard deviation. We could write the trimmed mean as

$$\bar{D} = \frac{1}{m} \sum_{i \in M} \frac{Y_{t,i} - Y_{(t-1),i}}{Y_{(t-1),i}} = \frac{1}{m} \sum_{i \in M} \frac{Y_{t,i}}{Y_{(t-1),i}} - 1$$

where M denotes the set of entities for which the percentage changes are in the middle 90% of the distribution, and m is the number of elements in M .

Using this technique, I'd like to come up with an imputation method. If we let $\hat{Y}_{t,j}$ be the predicted value for the j th entity at time t , and we estimate the percentage change by the trimmed mean, we have the following formula:

$$\frac{\hat{Y}_{t,j} - Y_{(t-1),j}}{Y_{(t-1),j}} = \bar{D} = \frac{1}{m} \sum_{i \in M} \frac{Y_{t,i}}{Y_{(t-1),i}} - 1$$

which leads to $\hat{Y}_{t,j}$:

$$\hat{Y}_{t,j} = \bar{D} Y_{(t-1),j} + Y_{(t-1),j} = \frac{1}{m} \sum_{i \in M} \frac{Y_{t,i}}{Y_{(t-1),i}} Y_{(t-1),j}$$

Thus, for the j th entity, we would use his previous time period value, adjust it by the mean ratio of those entities that have already responded, to obtain his current value.

Now looking at this from a regression point of view, consider:

$$Y_{t,j} = \beta Y_{(t-1),j} + \varepsilon_{t,j} \text{ where } \varepsilon_{t,j} \sim N(0, \sigma^2 Y_{(t-1),j}^\delta) \\ \text{for } j \in M$$

Using a weighted least squares the predicted value is:

$$\hat{Y}_{t,j} = \hat{\beta} Y_{(t-1),j} \text{ for } j \in M$$

where

$$\text{if } \delta = 1 \quad \hat{\beta} = \frac{\sum_{i \in M} Y_{t,i}}{\sum_{i \in M} Y_{(t-1),i}}$$

$$\text{if } \delta = 2 \quad \hat{\beta} = \frac{1}{m} \sum_{i \in M} \frac{Y_{t,i}}{Y_{(t-1),i}}$$

Both are unbiased estimators of β but the one that is more precise depends on the value of δ . $\delta = 2$ is what underlies the CS method.

Under many situations one can show that the sum of the ratios has better properties than the ratio of the sums. However, in a study we did at BLS considering alternative imputation methods, we found that the model with $\delta = 1$ did the best. This was a study involving employment and wage variables for establishments on the Universe Data Base. We investigated many different methods; among them was a generalized Bayesian model, which led to multiple imputation. We also considered a time series going back a year, but only the prior month in this simple model was needed.

$$Y_{t,j} = \beta Y_{(t-1),j} + \varepsilon_{t,j} \text{ where } \varepsilon_{t,j} \sim N(0, \sigma^2 Y_{(t-1),j}^\delta) \\ \text{for } j \in M$$

Using a weighted least squares, the predicted value is

$$\hat{Y}_{t,j} = \hat{\beta} Y_{(t-1),j} \text{ for } j \in M$$

where

$$\hat{\beta} = \frac{\sum_{i \in M} Y_{t,i}}{\sum_{i \in M} Y_{(t-1),i}}$$

The M in this case was a set of homogeneous establishments that had reported values in both time periods. For establishment j in time period t , $Y_{t,j}$ denoted, in various studies, the reported employment, the reported $\ln(\text{wages})$, and the reported $\ln(\text{wages}/\text{employment})$. I'm not sure which model would work best with Bank type data, but I think it's worth writing down the underlying models so they can be tested.

In imputation studies, we have a problem similar to one that exists with the CS method. In imputation, when modeling the respondents to predict for the nonrespondents, one hopes the nonrespondents are missing at random; that is, the nonresponse mechanism is ignorable. If this is not the case, it is a difficult problem to model the response mechanism. A similar situation arises with the CS method, in that the edit criteria are set by the early arrivals, and it is hoped that the respondents that are due late, behave in a similar fashion.

I have a couple of observations from the Tables. There were 24 edit groups, consisting of the 6 types of respondents and the 4 variables. Of these 24, for more than half (13), the recommended procedure is the composite of CS and RW. In most of the cases, the CS had the larger weight than RW. Clearly, some form of the CS technique is worth pursuing.

I note from Tables 5 and 6 that the probability of a type II error is very large, and in some situations the probability of a type I error is also large, but not as large--70's versus 90's. I would think that the type II error, not flagging a value when it's in error, is more important than the type I error, flagging a value when it's not in error. But from an analysts point of view, I can see that the type I error would be more important.

Now let me discuss the Julia Bienias, David Lassman, Scott Scheleur, and Howard Hogan Paper, "Improving Outlier Detection in Two Establishment Surveys." I also enjoyed this paper. First a brief summary of the paper.

This paper uses Exploratory Data Analysis (EDA) to improve the detection of outliers in the following two establishment surveys.

1. **The Annual Survey of Communication Services**---2000 firms
2. **The Monthly Wholesale Trade Survey**---7,000 firms, only 3,500 receive forms in a given month

Techniques discussed:

Box Plots

Scatter Plots

Transformations

Fitting: Ordinary Least Squares
Weighted Least Squares

As I mentioned earlier, in our imputation study for wages, we found that if we first transformed the data by the natural logarithm, and used a weighted least squares, the imputation improved.

In general, I believe EDA should be part of any outlier detection system. There is an extensive literature on testing for outliers. A number of popular procedures have difficulty when a sample may contain multiple outliers. Problems include **masking**, in which the presence of other outliers makes each outlier difficult to detect, and **swamping** in which the

procedure tends to declare too many outliers. By using robust and resistant methods it is possible to minimize the effects of deviate observations. An example is given in the Hoaglin, Iglewicz, and Tukey, JASA, 1986 paper, "Performance of some Resistant Rules for Outlier Labeling". Here you have inner and outer fences with hinges formed by the lower and upper fourths. That is, using the lower and upper fourths, F_L and F_U , the inner rule labels as "outside" any observations below $F_L - 1.5(F_U - F_L)$ or above $F_U + 1.5(F_U - F_L)$. For the outer rule 1.5 is replaced by 3.

In comparing the two papers, I found that I would like to combine them. For example, in the cross sectional estimates of Pierce and Gillis, additional EDA techniques could be used. As an example, instead of trimmed means one might consider "adaptive trimmed means". Some "adaptive trimmed means" determine the amount of trimming according to a sample estimate of the tail heaviness of the underlying distribution. This is especially useful if the distributions are not symmetrical, which I assume is the case with bank data.

In closing, I'd like to compliment the authors for very fine papers.

DISCUSSION

Brian V. Greenberg
U. S. Bureau of the Census

In this discussion we attempt to relate these two fine editing papers to the broader issues in data editing and highlight what one can learn from them.

1. Introduction--Role of Editing

Broadly speaking, there are two primary reasons for editing survey and census data. First, we would like to remove erroneous values from micro-data sets. A second, and related objective, is to ensure that we can generate meaningful estimates from reported data.

For some programs, there is an emphasis on actual micro-level data. For example, when one establishes a longitudinal data file or when a public-use micro-data file will be the primary survey data product. An example of a longitudinal micro-data file is described in the Pierce and Gillis paper and their edit activities focus on the underlying data set.

On the other hand, for some surveys there is a single estimate (or small number of estimates) produced from a survey, and the underlying data file is less important than the single estimate. The Census Bureau's monthly report of wholesale trade, as discussed in the Bienias, Lassman, Scheleur, and Hogan paper, is an example of such a survey.

In any event, data editing does not exist in a vacuum, and in designing and evaluating an edit system, one should be mindful of the survey's data collection and release objectives.

2. Editing Stages in Data Collection and Tabulation

There are typically three stages of data editing for a typical survey or census: (1) data entry edit, (2) automated batch edit of individual data records, and (3) review of summary tabulations.

In the data entry stage, editing often consist of rudimentary checks that only attempt to detect keying errors and major reporting problems. There are, however, data entry programs which have sophisticated and extensive data edit capabilities. For some surveys, editing in the data entry stage, including on-the-spot follow-up with respondents, serves as the primary edit activity.

Batch editing of individual data records, referred to as micro-editing, has been the mainstay of many large-scale survey and census programs. For some surveys, the automated program alters suspicious values, while for others the automated edit only flags suspicious values for analyst review and action. In addition, automated batch edit systems often impute for missing values.

After preliminary editing, data are tabulated and estimates are edited against prior time periods, against information from other sources, or against one another. The process of editing tabulation cells is often referred to as macro-editing. If a tabulation cell looks suspicious, it is reviewed and the individual micro-level records contributing to the cell are examined. Some programs have very sophisticated macro-edit systems while other macro-edit programs are essentially manual.

It is important to note that even though potential data errors are detected at the macro-level in a tabulation cell edit, problems are typically resolved at the micro-level.

After data are processed through automated edit programs, there is typically analyst review of large and/or important cases which often include direct follow-up with respondents. For large programs, there are not sufficient resources to review all records, therefore records are ranked by importance and those most important to a program are reviewed by analysts. The ranking process is often informal, however, research at Statistics Canada to formalize this process (referred to as selective editing) seems to have met with success for their Annual Survey of Manufactures.

All three edit stages come into play in virtually all survey programs. Emphasis on one stage or another is typically embedded in the edit strategy of each program, bearing in mind the proposed uses of survey products.

3. Edit Tolerances

At each stage of the editing process, edit tolerances are required to target individual records or tabulation cells for review and, if necessary, correction. In many respects, the two papers discussed here focus on deriving edit tolerances and we discuss them from this perspective.

There are several steps to deriving meaningful edit tolerances. First, one defines edit cells; that is, groups of respondents whose behavior is fairly similar with respect to the edit criteria. One typically wants cells to be small enough so that respondents can be relatively homogeneous yet large enough so that parameters are not unduly influenced by a few nontypical responses.

One can generate (explicitly or implicitly) an anticipated value for data fields or a relation between fields. The anticipated value may be based on data from the current or prior time periods and can be modeled based on all respondents in the edit cell. Tolerance limits are applied to target records which have unacceptable deviations from anticipated values. For example, anticipated values can be based on a regression line and the tolerance limits can reflect the allowable band of values about this line. Under alternative approaches to deriving tolerances one directly determines a range of acceptable data values and designates response combinations outside that range as edit failures.

Tolerance parameters are derived and applied at each of the three stages of the edit process: data entry edit, batch edit of individual records and tabulation cell edit.

User-friendly systems to support the review of data in the development of edit tolerances can be extremely valuable. It is in this light that the work of Bienias, Lassman, Scheleur, and Hogan can be viewed. The graphics techniques which they present are particularly important because they can help users organize and systemize information and share findings with others. Such systems provide analysts access to methodology not otherwise readily available to them.

4. Striking a Balance in Edit Tolerances and Review Criteria

If edit tolerances are too tight, excessive data may be altered or sent for analyst review. In the first case, edit programs can distort estimates and force data to conform to expectations. In the second, too many referrals place a major burden on analyst resources.

If edit bounds are too loose, erroneous data gets into the system. Such errors in data limit the usefulness of micro-data and may lead to unreliable estimates. Broadly speaking, parameters which are too loose deprive us of the chance to identify a source of nonsampling error.

After automated edit programs have applied tolerances and targeted records as suspicious, one would like to select the most significant problematic records for analyst review. For each survey, one needs a reliable criterion as to what is significant and what is not. The notion of significant depends a great deal on the proposed user of the survey data.

Recently, the phrase "over-editing" has come into vogue to refer to spending too much time and money on editing and/or changing too much data. I feel somewhat uncomfortable with this phrase because it is unfocused and gives a misleading impression. It seems to imply that if we edited less--perhaps had looser edit bounds or reviewed less micro-data--we would be editing better. In fact, we want to edit more cleverly, not necessarily more or less. That is, we would like to target fields for change and/or review where change is needed and not target fields for change and/or review when not needed.

A more useful formulation of the issue can be couched in terms of Type I or Type II error for the edit process, as was done in the Pierce-Gillis paper. Namely, for their purposes:

A Type I error (a 'false positive') refers to an item that was flagged but was not in error, or at least not revised.

A Type II error occurs when an item is not flagged but is erroneous (as evidenced by a later revision).

We can broaden their definition a little to say:

Type I error refers to an item flagged for change or review but the time spent on it did not improve the data set or estimates for the survey.

Type II error occurs when an erroneous value, which adversely affect the quality of the data set or survey estimates, is not flagged for change or review.

The last conference on Statistical Policy Working Papers of the Federal Committee on Statistical Methodology was held in March, 1991. At that conference, there was a session based on Working Paper #18, "Data Editing in Federal Agencies." In that report, the focus was on development of multipurpose systems, software design, and edit methodologies. There was little discussion of parameter development. Since then, it has been increasingly clear that good parameter development is crucial in all stages of editing. It is also clear that we need to give greater attention to the interplay between subject-matter staff and automated programs in the resolution of edit failures and in the design of edit tolerances.

Even the best edit methodology embedded in the finest system will perform poorly if there are bad parameters. In fact, the choice of edit tolerances has a major influence on the Type I and Type II error for editing. We certainly need more investigation to highlight what methods and tools work well for the design of edit tolerances and we need to examine and learn from clearly presented case studies.

The two papers under discussion do an excellent job in addressing these related issues.

5. Bienias, Lassman, Scheleur, and Hogan Paper

The authors illustrate graphic techniques used in the spirit of exploratory data analysis as tools for subject-matter specialists in deriving edit parameters. They also describe how the simultaneous review of survey data can introduce advantages over a case-by-case analysis of report forms.

Box plots were used to review and summarize information and directly contributed to parameter development for the Annual Survey of Communication Services (ASCS). In particular, the box plot for parameters based on the expense/revenue ratio illustrates this use.

Graphics were also used to help uncover similarities or differences between establishments. By examining residuals in the relation between revenue and payroll in ASCS, they decided to remove tax-exempt establishments from edit cells for revenue and payroll. That is, they were able to design a more effective edit cell for subsequent analysis, which they describe.

And finally, the graphics and exploratory data analysis led to a more suitable editing model for current to prior inventory on the Monthly Wholesale Trade Survey. In this usage, the techniques they employed allowed subject-matter analysts to experiment with different models and to select the model that they felt best represented the data.

An important theme of this paper was the interplay between subject-matter analyst expertise and the use of graphical methods. These tools can provide a guide for analysts and allow them to see the impact of proposed models. They can contribute to the design process and help eliminate some of the more tenuous aspects of model description. In addition, the graphs provide a useful vehicle for improved communication and shared information among those working on a project.

By all accounts, the survey analysts and project managers who work on the surveys cited above found the contributions described in this paper extremely valuable. It will be through continued and expanded use that additional benefits and applications will arise.

6. Pierce and Gillis Paper

This paper is a superb case study for the development of effective edit cells and related tolerances. This report can be a textbook study. One of the author's primary objectives is stated clearly at the onset: "A major task in setting edit tolerances is to ensure sensitivity without generating unnecessarily large quantities of 'false positive' exceptions." They have an excellent test environment because there is an unequivocal response question and the "truth" can always be determined (by subsequent revision) so the appropriateness of the edit can be evaluated.

Note that the authors clearly have a quintessential micro-editing requirement. Namely, their intended product is a longitudinal file of individual records of bank deposits.

After describing the underlying survey environment, the authors described their step-by-step process to design effective edit tolerances. They described how they developed the definition of edit cell and how they had to combine cells to get the proper break between cell size and homogeneity. They next described the model to predict (forecast) reported deposits, discussed alternative models and provided cogent reasons for each decision along the way. Following that, the authors describe their procedure for setting cell edit tolerances. After details of the edit system were decided upon, they were able to test various options based on the 1991-92 edit experience.

They took a major step in couching their analysis in terms of Type I and Type II errors to evaluate findings. The authors provided extensive tables and descriptions of their analysis. It is interesting to note that the current system has too many Type II errors, and future work will introduce refinements to achieve a lower rate.

Although one rarely comes across such a well-suited environment to test edit procedures and evaluate performance, this report is valuable in describing how to proceed under ideal circumstances. Using this ideal as a guide, one can modify procedures and change directions based on information actually available when attempting to apply the methods described in this report to other surveys.

7. Concluding Remarks

Both of these papers have a great deal to offer the reader. The first clearly illustrates how graphics can be applied to actual editing issues. One would hope that the examples here can suggest methods which can be applied to other surveys. The second paper is an excellent case-study for developing edit tolerances and evaluating them. This paper also can serve as a guide in helping others plan their own evaluation projects.

Session 5

TIME SERIES REVISION POLICIES

Time Series Revisions: The Effects on Gross Domestic Product

Robert P. Parker and Teresa L. Weadock
U.S. Department of Commerce
Bureau of Economic Analysis

Gross domestic product (GDP) is the most widely used measure of a Nation's overall economic activity. In response to the need for timely estimates, the Bureau of Economic Analysis (BEA) releases an "advance" estimate of quarterly GDP one month after the end of the each quarter. This estimate is based largely on monthly survey data for the first two months of the quarter and BEA judgmental projections for the missing source data. In each of the next two months, revised estimates of GDP that incorporate newly available and revised monthly and quarterly source data are released. Annual and benchmark (comprehensive) revisions of GDP are released on a regular schedule as annual and less frequently collected census-type data become available.

To measure the reliability and accuracy of the quarterly GDP estimates, BEA has conducted a series of studies based primarily on revisions to the successive quarterly estimates. Such studies have been found to be very useful to both users of the GDP estimates and to BEA. For users of the estimates, the studies provide insights into the likely size of future revisions to GDP and its major components and identify components whose reliability they would like to have BEA improve. For BEA, the studies help to identify components with problems in the source data or estimating methodologies. This information helps BEA to work with the agencies who prepare the source data or to devote its own resources to developing improved procedures to reduce revisions. Information from these studies also enables BEA to analyze the impact on the reliability of the GDP estimates of its revision schedule.

In addition, the studies satisfy the requirement of the Office of Management and Budget that agencies producing the principal Federal economic indicators provide periodic evaluations of their performance. This requirement, Statistical Policy Directive Number 3, states that these evaluations will "include an analysis of the accuracy of the series, the effects of revisions, and performance relative to established benchmarks."

This paper, which is based primarily on measures of reliability for the quarterly GDP estimates for 1978-91, consists of four parts: An overview of the preparation of the quarterly GDP estimates and a discussion of the measures used by BEA to measure their reliability;

NOTE: The views expressed in this paper are those of the authors and not necessarily those of the Department of Commerce or the Bureau of Economic Analysis.

highlights of a recently completed study by Allan Young at BEA on the reliability of the first three, or "current" quarterly GDP estimates for 1978-91;¹ an extension of this study to the effects on the quarterly estimates of subsequent annual revisions; and a discussion of revision practices that agencies use for the source data used to prepare the GDP estimates and how these practices affect the accuracy of the quarterly estimates.

The studies reviewed in this paper have implications both for BEA and for the agencies that provide the source data used to prepare GDP. They indicate that BEA should review the need for three current estimates and its policy of not revising prior quarters except at the time of annual and benchmark revisions. They also indicate that other agencies should review their revision practices to provide more timely and accurate revised data.

Part 1. Overview of the Preparation of Quarterly GDP Estimates and Measures of Reliability

Estimating schedule for quarterly GDP estimates

For each quarter, GDP estimates are prepared on a schedule that consists of three successive "current" estimates--"advance," "preliminary," and "final"-- and of subsequent estimates prepared as part of annual and comprehensive NIPA revisions.

The advance estimate is prepared about 1 month after the end of the quarter. For most components, the estimate is based on source data for either 2 or 3 months of the quarter. In most cases, however, the source data are not final and are subject to revision by the issuing agencies. Where source data are not available, the estimate is based primarily on past trends and on BEA analysts' judgment.

One month later, the advance estimate is replaced by the preliminary estimate, which is typically based on source data for all 3 months of the quarter. In most instances, the source data used for the preliminary estimates, particularly the data for the third month of the quarter, are subject to further revision.

One month later, the preliminary estimate is replaced by the final estimate, the last of the current estimates, which incorporates revisions in source data for the third month of the quarter and quarterly source data for some components. For virtually all components, these source data are subject to further revisions by the issuing agencies.

¹ See Allan H. Young, "Reliability and Accuracy of the Quarterly Estimates of GDP" in the October 1993 issue of the Survey of Current Business.

Each quarterly estimate is subject to three successive annual revisions, usually released in July. The first annual revision incorporates further revisions in the monthly or quarterly source data and introduces some annual source data. The second and third annual revisions incorporate a broad range of annual source data. For example, the "final" estimate for the fourth quarter of 1993, which was released last month, will be revised in July 1994 (first annual revision), July 1995 (second annual revision), and July 1996 (third annual revision). Each quarterly estimate is also subject to one or more comprehensive revisions, in which information from the economic and demographic censuses is incorporated in the monthly, quarterly, and annual source data by the issuing agency or by BEA.

Source data

More complete and more accurate information is generally available on an annual basis than on a quarterly or monthly basis. In many cases, annual data are based on larger samples or represent a complete universe count. In addition, annual data often correspond more closely to the desired definitions and therefore require less adjusting, or they may contain more information for making the necessary adjustments. As a result of these factors, quarterly estimates are obtained either by interpolating between annual estimates or by extrapolating from the most recent annual estimate.

Similarly, the annual estimates in many instances represent interpolations or extrapolations of the more complete and accurate information available in economic and demographic censuses, which are conducted every 5 years and 10 years, respectively.

The quarterly and monthly indicators that are used as interpolators and extrapolators are based largely on monthly or quarterly sample surveys conducted by various Federal statistical agencies. Exceptions include budgetary data from the Treasury Department, tabulations of export and import documents filed with the Customs Service, and tabulations of several types on payroll and income tax returns. Another type of exception occurs if no monthly or quarterly data are available--for example, data for some types of consumer purchases of services and of State and local government purchases of goods and services. In such cases, the quarterly estimates are obtained by interpolation and extrapolation based on a BEA analyst's judgment or related information.

An updated summary of the source data used for the NIPA's is included each year in the Survey of Current Business article that presents the annual NIPA revision (see pages 31-42 of the August 1993 Survey). For a list of methodological papers and for additional information about the NIPA's, see "A Look at How BEA Presents the NIPA's" in the February 1994 Survey, pages 31-33.

Sources of error

The GDP estimates contain several kinds of error. The most obvious kind arises in the current estimates either from preliminary or incomplete tabulations of monthly or quarterly source data or, where source data are not yet available, from BEA's judgments. Error also arises in both the current and the latest available estimates because source data do not meet NIPA requirements in terms of timing, valuation, coverage, and definitions. For example, business firms report some types of data on a fiscal year, rather than a calendar year, basis; even though adjustments by BEA reduce the effect of fiscal year reporting, the results differ from those that would have been obtained with calendar year reporting.

Error also arises from the sampling errors and biases in the monthly, quarterly and annual surveys and from biases and other errors in the annual and periodic universe counts. Probably the most troublesome of errors are those due to the delayed recognition of births and deaths of business firms in sample surveys. (These types of errors that affect source data are discussed in part 4 of this paper.)

Seasonal adjustment is another source of error.² Even if the unadjusted source data were free of error, seasonal adjustment would introduce errors. Although some reduction in seasonal adjustment error appears to have been achieved over time in the current estimates through the use of concurrent seasonal adjustment and by combining ARIMA methods with the ratio-to-moving-average method of seasonal adjustment, such errors are still of considerable magnitude.

Measures of reliability

The term "reliability" used in BEA studies refers to the revisions in the estimates, which reflect the following: (1) Replacement of preliminary source data with revised or more comprehensive data, (2) replacement of judgmental projections with source data, (3) changes in definitions or estimating procedures, and (4) in the constant-dollar estimates, updating of the base year.

In its studies of the reliability of the quarterly GDP estimates, BEA uses six summary measures to describe the revisions: Dispersion, bias, relative dispersion, relative bias, upward revisions, and directional misses. (This paper focuses on the dispersion because this measure effectively summarizes the information provided by the other measures.)

² Quarterly and monthly NIPA estimates are seasonally adjusted if necessary. Seasonal adjustment removes from the time series the average impact of variations that normally occur at about the same time and in about the same magnitude each year--for example, weather, holidays, and tax payment dates.

The six measures are calculated as follows. Let P represent the percentage change in the current estimates, L the percentage change in the latest available estimates, and n the number of quarterly changes.

Dispersion is the average of the absolute values of the revisions:

$$\Sigma|P-L|/n$$

Bias is the average of the revisions:

$$\Sigma(P-L)/n$$

Relative dispersion expresses the dispersion as a percentage of the average of the absolute values of the percentage change in the latest available estimates:

$$\frac{\Sigma|P-L|/n}{\Sigma|L|/n}$$

Relative bias expresses the bias as a percentage of the average of the percentage change in the latest available estimates:

$$\frac{\Sigma(P-L)/n}{\Sigma L/n}$$

Upward revisions expresses the number of times that the current estimate of the quarterly change was revised up by the latest available estimate, as a percentage of the number of quarterly changes.

Directional misses expresses the number of times that the sign of the current estimate of the quarterly change differed from that of the latest available estimate, as a percentage of the number of quarterly changes.

In evaluating these measures, they should be viewed in light of two aspects of the estimation process. First, a change in source data or estimating procedures, which one may assume affects the accuracy of the estimates, is not necessarily reflected in the revision of estimates of a given vintage. For example, an improvement in the current estimates results in a permanent decrease in revision size. An improvement in the latest available estimates results in a permanent increase in revision size. Improvement in both the current and latest available estimates results in little change. However, improvement that is introduced retrospectively into the latest available estimates, as is often the case, results in an increase in revision size for a period of years until the improvement is also reflected in the current estimates. Second, the latest available estimates reflect different vintages. The latest estimates up to 1982 at present reflect the incorporation of the benchmark input-output (I-

0) tables, which are based on detailed information from the economic censuses; the latest estimates beginning in 1983 do not yet reflect the incorporation of the recently released benchmark 1987 I-O tables.³ Thus, the size of revisions beginning with 1983 estimates are most likely understated.

Part 2. Reliability of the Current Quarterly Estimates⁴

Summary of Young study, 1978-91

In the most recent BEA study, Young provided an overall evaluation of the reliability of the quarterly GDP estimates by comparing the successive current estimates of real GDP to the latest estimates and asking the following questions:

o Do the current estimates provide a correct indication of the direction of the change in aggregate economic activity?

o Do the current estimates provide a correct indication of whether the change in aggregate economic activity is larger (acceleration) or smaller (deceleration) than in the previous quarter?

Table 1 provides the summary answers to these questions for each of the current estimates. The record for 1978-91 shows that all three estimates correctly indicated direction of change almost 90 percent of the time. They correctly indicate acceleration and deceleration between 75 and 80 percent of the time. (If changes between -1 percent and +1 percent are disregarded, these early estimates correctly indicate direction over 90 percent of the time and acceleration about 85 percent of the time).

Young also found that for the same period, the incorporation of additional or more accurate source data in the second (preliminary) and third (final) quarterly estimates of GDP did not improve the reliability in comparison with the first (advance) estimate. He identified two factors that contributed to this finding. First, the data for second and third months of a quarter play only a small role in determining the change from the previous quarter. Second, the advance estimate is unaffected by certain sources of error in the preliminary and final estimates. In addition, the advance estimates

³ The benchmark 1987 I-O tables were published in the April 1994 issue of the Survey of Current Business. The results of these tables will be incorporated into the GDP estimates in a comprehensive revision presently scheduled for release in late 1995.

⁴ Most of the sections of this part of the paper, as well as several paragraphs of the previous section, were taken verbatim from Young's article in the October 1993 Survey of Current Business.

of GDP and its major components may benefit more from offsetting errors in the detailed components than the later current estimates; that is, the revisions of the advance estimates may be more negatively (or less positively) correlated than those of the preliminary and final estimates.

In the remainder of this part of the paper, Young's findings are presented in more detail.

Reliability of the quarterly estimates

Table 2 shows dispersion for quarterly changes in current- and constant-dollar GDP and its major components for 1978-82 and 1983-91. These measures show that the incorporation of additional or more accurate source data in the preliminary and final current estimates of GDP does not substantially improve the reliability in comparison with the advance estimates. Dispersion declines only slightly over the successive current-dollar estimates of GDP. For 1978-82, it declines from 1.93 percentage points in the advance estimates to 1.82 percentage points in the preliminary and final estimates. For 1983-91, it declines from 1.17 percentage points in the advance estimates to 1.14 percentage points in the preliminary estimates and 1.15 percentage points in the final estimates.

Dispersion actually increases slightly over the successive constant-dollar estimates of GDP. For 1978-82, it increases from 1.64 percentage points in the advance estimates to 1.72 percentage points in the preliminary estimate and to 1.75 percentage points in the final estimate. The corresponding figures for 1983-91 are 1.25, 1.27, and 1.33 percentage points.

A similar picture emerges for the major components of GDP. In many cases, the advance estimates provided a smaller measure of dispersion than did the preliminary or final estimates. In 1978-82, the advance estimates provided the smallest dispersion in 4 of the 11 current-dollar components--PCE nondurables and services, residential investment, and Federal Government purchases--which accounted for almost 60 percent of GDP. In 1983-91, the advance estimates provided the smallest dispersion in 3 components--PCE services, residential investment, and State and local government purchases--which accounted for over 40 percent of GDP. The record for the advance constant-dollar estimates is about the same as that for the current-dollar estimates, though the share of GDP for which the estimates perform the best is smaller for 1983-91. These results raise the question of whether one or both of the two later current estimates might be discontinued.⁵

⁵ Previous studies have also shown that the advance estimates perform well in comparison with the later estimates, but as shown in table 2, the results are not as clear cut in some periods as in others.

Table 2 also permits one to compare the size of the dispersion measure for GDP with that of its major components. In general, dispersion in the components was larger than that in GDP. The components with the smallest dispersion--about the same as that for GDP--were total PCE and PCE services. The components with the largest dispersion--roughly 6 to 8 times as large as that for total GDP--were gross private domestic investment and Federal Government purchases. The unusually large dispersion in these components reflected a change in the treatment of the Commodity Credit Corporation (CCC) that was introduced in the 1991 comprehensive revision, whereby the CCC was shifted from government enterprises to general government. This shift affected the timing and valuation of transactions and resulted in large, essentially offsetting revisions in Federal Government purchases and the change in business inventories. Dispersion was also quite large in current-dollar nonresidential structures in 1978-82 and in constant-dollar imports in 1983-91, reflecting statistical improvements introduced in the 1991 comprehensive revision.

Trends since 1978

Table 3 examines revisions year by year to see if reliability of the GDP estimates appears to have changed in recent years. The table shows annual averages of dispersion and bias in the quarterly revisions between the successive current estimates and between the current estimates and the third annual revision estimates. For the revisions between the current estimates, the measures are shown for 1978-92; for the revisions between the current estimates and the third annual estimates, the measures are shown for 1978-89.

In Young's study, estimates from third annual revisions were used in place of the latest available estimates in order to provide a more nearly comparable standard for the entire period against which to compare the current estimates. Use of third annual estimates abstracts from much of the effect of the economic census and other information that is used in the comprehensive revisions to revise previously prepared third annual estimates. However, it does not remove the effects of definitional changes in the comprehensive revisions, because for most quarters a comprehensive revision intervenes between the current estimates and the third annual estimates. (To more fully study the effects of the annual revisions,

An early study concluded that the advance estimate might be sufficient; see Rosanne Cole, "Errors in Provisional Estimates of Gross National Product," National Bureau of Economic Research Studies in Business Cycles No. 21 (1969). See also Stephen K. McNees, "Estimating GNP, The Trade-off Between Timeliness and Accuracy," New England Economic Review (January/February 1986): 3-10; and Joseph W. Duncan and Andrew C. Gross, Statistics for the 21st Century (The Dun and Bradstreet Corporation, 1993).

a different approach is used in part 3 of this paper.)

The good performance of the advance estimates

The absence of much improvement in the successive current estimates has puzzled both users and estimators for some time. Two seldom recognized factors contribute to the observed result: (1) The small role played by the data for second and third months of a quarter in determining the change from the previous quarter, and (2) certain sources of error in the preliminary and final estimates to which the advance estimates are immune. In addition, advance estimates of GDP and its major components may benefit more from offsetting errors in the detailed components than the later current estimates; that is, the revisions of the advance estimates may be more negatively (or less positively) correlated than those of the preliminary and final estimates.

This section first discusses the two factors and then addresses the problem of quantifying the total error introduced by the second factor, which embodies seasonal adjustment errors and errors related to the estimation process for certain components. The section concludes with a discussion of the implications for the future.

The role played by the data for second and third months of the quarter in determining the change from the previous quarter is small. The change from the second to the third month receives a weight of only one-ninth in the determination of quarterly change. The weight of the second and third months together is only one-third. The weight of the first month is another one-third, and the second and third months of the previous quarter receive the remaining one-third.⁶ Consequently, errors in neither the preliminary source data for the second and third months of a quarter nor in the judgmental projections used in lieu of source data affect the quarterly change as much as one might intuitively expect.

⁶ This may be demonstrated as follows: Let $Q_1 = X_1 + X_2 + X_3$ and $Q_2 = X_4 + X_5 + X_6$, where X_1, X_2, \dots, X_6 are successive months of source data.

Then, if $d_4 = X_4 - X_3$, $d_5 = X_5 - X_4$, and $d_6 = X_6 - X_5$, the months in Q_2 may be stated as $X_4 = X_3 + d_4$, $X_5 = X_3 + d_4 + d_5$, $X_6 = X_3 + d_4 + d_5 + d_6$, and

$$Q_2 = 3X_3 + 3d_4 + 2d_5 + d_6.$$

Therefore, the quarterly change is

$$\begin{aligned} Q_2 - Q_1 &= (3X_3 + 3d_4 + 2d_5 + d_6) - (X_1 + X_2 + X_3) \\ &= [(X_3 - X_2) + (X_3 - X_1)] + [3d_4 + 2d_5 + d_6]. \end{aligned}$$

Introducing the notation for monthly differences, the first bracketed term becomes $[d_3 + (d_2 + d_3)]$, and

$$Q_2 - Q_1 = d_2 + 2d_3 + 3d_4 + 2d_5 + d_6.$$

Normalizing the coefficients on the d 's provides weights of $1/9$, $2/9$, $3/9$, $2/9$, and $1/9$ for the five monthly changes that determine the quarterly change.

The seasonal adjustment of source data for the final current quarterly estimate introduces errors not present in the judgmental projections, which are developed on a seasonally adjusted basis. The seasonal adjustment factors for the current year are derived from the seasonal patterns of recent preceding years. (The concurrent seasonal adjustment method also includes the seasonal pattern of the current year.) The factors are revised as additional data become available, and they eventually reflect the average seasonal pattern of a period of years that extends symmetrically on either side of the given year. The difference between the initial estimate of the seasonal factor and the final estimate prepared some years later is an error that becomes part of the revision in the final current estimate. To the extent that they are based on judgmental projections, the advance and preliminary quarterly estimates do not contain this error.

Future work

The difference between the revisions to the advance estimate of a detailed component and those to the latest available estimate reflects three types of error: (1) The error in the preliminary monthly source data used for the advance estimates that is corrected in the revised monthly source data used for the final current estimate; (2) the error in the judgmental projections used in lieu of source data for the advance estimate; and (3) the error in the source data used for the final current estimate (including seasonal adjustment error) that replace the judgmental projections used for the advance estimate. The total revision in the advance estimate reflects the first two types of error; the total revision in the final current estimate reflects the third type. One should note that the second type of error, like the third, is determined with respect to the data as they stand several years later.

It would be desirable to determine the size of each of the three types of error at the detailed component level. It would also be desirable to determine the extent of correlation among the component revisions. Such analyses presently are not possible, because each vintage of each estimate and the associated source data are not available in a readily usable form. However, the database used by BEA to calculate the alternative measures of real GDP might be extended so as to retain not only the latest available estimates, but all the vintages of estimates at the detailed level at which deflation is carried out.⁷ Over time, this database would be useful in exploring the outcomes of the estimation process and in developing improved procedures. For example, it will facilitate the development of econometric projection techniques and their comparison with judgmental

⁷ For information about the alternative measures, see Allan H. Young, "Alternative Measures of Change in Real Output and Prices, Quarterly Estimates for 1959-91," in the March 1993 Survey of Current Business.

projections. In this respect, it should be noted that a recently completed study found that judgmental projections compared favorably with econometric techniques for certain import and export components.⁸ Thus, such procedures would be difficult to justify if they do not lead to smaller revisions than do judgmental projections.

The question of whether the reliability of the seasonal adjustments on which the current quarterly estimates are based can be improved merits attention. An analysis at a fairly high level of aggregation suggests that revisions in seasonal factors may be large enough to contribute significantly to the observed results. Specifically, in some of the series examined, seasonal-factor revisions are as large as the variation in the irregular component.⁹ Given that the error introduced by a judgmental projection is likely to be smaller than the irregular variation, this result suggests that for some detailed components, seasonal-factor revisions may play a significant role in causing the revision in the final estimate to be as large as that in the advance estimate. In addition, the seasonal adjustments used for source data should be designed from the standpoint of accurately measuring quarterly change. Little attention has been paid to whether the currently used seasonal adjustment procedures are suitable from this standpoint. In addition, BEA should consider whether more use of concurrent seasonal adjustment, with or without ARIMA, would improve reliability.

Finally, because the third month of a quarter receives little weight in the estimate of change for that quarter, there may be instances in which efforts to reduce revisions in the quarterly GDP estimates should focus on improving the final monthly source data rather than the preliminary monthly source data. As shown earlier, for a survey with three successive monthly estimates, two-thirds of the advance quarterly change is based on three monthly final estimates, while only one-ninth is based on the initial monthly estimate for the last month of the quarter.

Part 3. Effects of Annual Revisions on Quarterly GDP Estimates

In his study on the reliability of the quarterly GDP estimates, Young studied the reliability of the three current quarterly GDP estimates. In this part of the paper, the revisions in the quarterly estimates published in the regular annual revisions are studied.

⁸ See Albert A. Hirsch and Michael A. Mann, An Analysis of the Use of Time-Series Models to Improve Estimates of International Transactions, Bureau of Economic Analysis Working Paper 7 (Washington, DC: Bureau of Economic Analysis, April 1993).

⁹ The irregular component is the residual after the systematic components--the seasonal and trend-cycle--are determined by the seasonal adjustment method.

Dispersion

Table 4 shows the dispersion for quarterly changes in current- and constant-dollar GDP and its major components for 1981-90, the years for which comparisons using annual revisions could be made.¹⁰ These measures show that the incorporation of additional or more accurate source data in the first and third annual revisions substantially improves the reliability in comparisons with the third, or "final" current estimates. The dispersion for current-dollar GDP from the final to the first annual revisions declines from 1.34 percentage points to 0.84 percentage point; for constant-dollar GDP, the decline is from 1.45 percentage points to 0.83 percentage point. From the second to the third annual revisions, the dispersion for current-dollar GDP declines from 0.85 percentage point to 0.65 percentage point; for constant-dollar GDP the decline is from 1.08 percentage points to 0.85 percentage point. However, from the first to the second annual revisions for the current-dollar estimates, the dispersion is virtually unchanged, and for the constant-dollar estimates, there is a significant increase in the dispersion from 0.83 percentage point to 1.08 percentage points.

For the major components of GDP, except for nonresidential structures and producers' durable equipment, the first annual revision estimates recorded a smaller measure of dispersion than did the final estimates. Between the first and second annual revisions, the reliability of all major components improved. Between the second and third annual revisions, the reliability of all major components except for durable goods personal consumption expenditures (PCE) and exports of goods and services improved.

Table 4 also permits one to compare the size of the dispersion measure for GDP with that for its major components. The dispersion for the components was larger than that for GDP. The components with the smallest dispersion were PCE nondurable goods, PCE services, and State and local government purchases. The components with the largest dispersion, more than 6 times as large as that for total GDP, were gross private domestic investment, nonresidential structures, and services, and Federal Government purchases. Of these components, the largest dispersion was Federal Government purchases. As noted in the review of Young's study, the size for this component reflected a change in the treatment of the Commodity Credit Corporation that was introduced in the 1991 comprehensive revision.

¹⁰ To conduct this study, it was necessary to reduce the period covered in Young's study because there were no annual revisions in 1980 and 1991, years of benchmark revisions. See footnote 1 of table 4 for additional details.

Findings

Young found that the revisions in the advance current estimates were about the same size as those in the other current estimates. Thus, he questioned the need to continue the preparation of the preliminary and/or final quarterly GDP estimates. The comparisons in table 4 show, as expected, that generally there was continued improvement in the reliability of the estimates in each successive annual revision; therefore, the need for the annual revisions is not called into question. However, the relative size of the improved reliability from the final quarterly estimates to the quarterly estimates from the first annual revisions was larger than expected. The large improvement, which is evident in both the current- and constant-dollar GDP estimates, was unexpected because very few of the annual surveys that are used to prepare the revised GDP estimates become available in time to be incorporated into the first annual revisions. For example, the Census Bureau's annual surveys of retail trade and of manufacturers, which are used for the estimates of PCE goods and of producers' durable equipment, are not available until the second annual revision. If the improved reliability does not result primarily from the incorporation of such new source data, then the improvement might result from two other sources. One source would be the replacement at the time of the annual revisions of seasonal adjustment factors that were derived from the seasonal pattern of preceding years with factors that also reflect the pattern for the most recent year. The second source would be the incorporation of corrections to monthly or quarterly source data series that become available too late to be incorporated into the "final" current estimates.

If research of the detailed components shows that these are major sources of the large revisions in the first annual revisions, then BEA should consider revising previous quarters between annual revisions. For example, when the advance estimate for the fourth quarter of a given year is released in January, newly available corrections and revisions to source data series and updated seasonal factors could be incorporated into revised first-, second-, and third-quarter estimates. This change would allow for a more effective use of concurrent seasonal adjustment -- that is, using these updated factors for all months or quarters of the year. (For many of the source data series for which the issuing agency uses concurrent seasonal adjustment, the new factors are available but are only used for the most recent two or three months.) Changing BEA's revision practice as just described would provide more reliable quarterly estimates earlier than under the present annual revision schedule. If this change were merged with Young's suggestion to eliminate one or more of the current quarterly estimates, BEA might, for example, drop the final current estimates and revise the previous quarters of the year when the advance estimates are released.

Part 4. Revision Practices in GDP Source Data

Accuracy of the GDP estimates

The term "accuracy" refers to the total measurement error. On the assumption that each successive estimate is more accurate than previous ones, revisions can be viewed as measuring part of the total error in earlier estimates. The rest of the error in these estimates, which is unknown, becomes the total error in the latest estimates. The error in the latest estimates results primarily from the following sources: (1) Errors in the most recently available underlying monthly or quarterly, annual, or periodic census source data; (2) errors in the adjustments made by BEA to convert source data to the definitions and conventions used in compiling GDP; (3) errors in BEA's judgmental estimates for components for which there are no source data; and (4) errors because the most recently available source data have not yet been incorporated into the estimates. It is important to note that these types of measurement error have two different effects on the quarterly GDP estimates. The first type of error affects the levels of the estimates of GDP and its components; the second type affects the changes in the estimates. In this paper, the focus is on the second type of errors.

The implication of the presence of measurement error in the latest estimates limits the extent to which the BEA's measures of reliability can be used as measures of accuracy. The questionable relationship between reliability and accuracy is illustrated by the component that has a very high measure of reliability -- that is, very small revisions -- because the source data are never revised by the issuing agency even though the series maybe based on a very small sample. The next section of the paper discusses how three commonly used revision practices adversely affect the key monthly and quarterly source data series used to estimate GDP. For this purpose, revision practices are defined as those that affect only "not seasonally adjusted estimates."

Revision practices that adversely affect GDP

For some surveys, the revision practice consists of a series of regular annual revisions and, if necessary, periodic benchmark revisions. For annual revisions, the monthly or quarterly series are adjusted to reflect annual survey data, which are based on larger samples, or to reflect annual census (universe) data. For periodic benchmark revisions, first the annual survey data and then the monthly or quarterly series are adjusted for all periods since the last benchmark. Examples of GDP source data prepared using this type of revision practice, the issuing agency, and the GDP components affected are as follows: Retail and wholesale trade sales and inventories from the Bureau of the Census (BOC) -- personal consumption expenditures (PCE) and change in business inventories (CBI); farm output and income from the Department of Agriculture--CBI and farm income; manufacturers' shipments and inventories from BOC--producers' durable

equipment and CBI; and establishment employment from the Bureau of Labor Statistics (BLS)--wages and salaries and PCE.

Although the revised monthly and quarterly estimates for source data series prepared using this practice are considered very accurate, they are nevertheless subject to measurement errors that are likely to introduce errors into the latest estimates of quarterly GDP changes. These errors arise because the agencies generally proportionately allocate across months (or quarters) differences (1) between the monthly or quarterly surveys and the annual surveys and (2) between the annual surveys and the census results where proportionate allocation is not appropriate. For example, the most common sources of differences are errors in the initial survey estimates, both monthly and quarterly and annually, due to delayed recognition of births and deaths, to classification errors, or to misreported data. For births and deaths, proportionate allocation of errors introduces new errors into the revised monthly or quarterly series because this type of allocation fails to recognize that changes in the business population are significantly affected by the business cycle and are not likely to have occurred proportionately over the period covered by the revision. For classification and misreporting errors, the errors also are not likely to have occurred proportionately throughout the revision period.

Another common revision practice used for monthly and quarterly series, which can be viewed as a variant of the practice described above, also is likely to introduce errors into the latest estimates of quarterly GDP changes. Under this practice, when periodic benchmark data are introduced, they are used to adjust only the monthly or quarterly and annual estimates for the year for which such data are available and to serve as the basis for the samples to be used for subsequent periods. Data for previous periods are not revised even though the data for these periods may include errors. Examples of GDP source data that are prepared using this practice, the issuing agency, and the GDP components affected are as follows: Foreign direct investment income in the United States from BEA -- corporate profits; State and local government value of construction put-in-place from BOC -- government purchases; and the Consumer Price Index from BLS -- PCE.

A third common revision practice occurs when the agency revises a series to remove selected measurement errors but not necessarily to remove the largest errors, which therefore remain in the revised estimates. The following monthly GDP source data are examples of this type of series: monthly merchandise trade from BOC -- net exports; new residential construction put in place from BOC -- residential fixed investment; and average hourly earnings from BLS -- wages and salaries.

The monthly merchandise trade data are regularly revised to record the export and import transactions based on Customs documents in the correct month but not to record reliable estimates of transactions for which Customs documents are not required to be filed

because of the value of the transaction is below some cutoff or "low value." These unreported low value transactions, which currently account for about 2.5 percent of exports and 4 percent of imports, are estimated using factors based on the amount of such transactions when the exemptions were granted. The adjustments are extrapolated by country, but not by commodity, using changes in reported transactions. This procedure is likely to introduce errors into the monthly changes in exports and imports.

The new residential construction put-in-place series, which is based largely on a sample of housing starts, is regularly revised to reflect additional monthly reports but not to reflect more accurate annual survey data. For the new single-family residential construction component of this series, the value of housing starts are "phased" in over many months based on a fixed monthly pattern of construction activity. This practice introduces errors into the monthly changes in the put-in-place series because the pattern is not updated on a regular basis (the present pattern was estimated for 1976) and there are no data to adjust the pattern for developments such as unusually bad weather. The errors created by this practice can be illustrated using the recent California earthquake. This disaster not only delayed starts, which the series does reflect, but also delayed activity on previously started houses, which the series does not reflect. For new multi-family residential construction, BOC conducts a monthly survey of construction put in place based on a sample of housing starts; this series is not benchmarked.

The final example of a "selective" revision practice is the average hourly earnings series, which is estimated from the BLS monthly establishment survey. Although the employment data collected in that survey are benchmarked annually, the earnings data are corrected only to reflect the revised employment data, which are used to weight the detailed industry earnings estimates to arrive at "all industry" totals. It is likely that a benchmarked hourly earnings series not only would change the levels of the series, but also the monthly changes.

Improving revision practices

This discussion of revision practices identifies some of the types of measurement errors that remain in BEA's latest estimates of changes in quarterly GDP. Although reducing these errors for some series would require the collection of new data, it should be possible for the issuing agencies to reduce certain types of errors with little or no additional resources. For example, errors caused by proportionate allocation of the impact of delayed incorporation of births and deaths could be significantly reduced. Instead of waiting many years until a regular benchmark revision, agencies could continuously track births and deaths and adjust their series annually, even with a one-year lag. This procedure would improve the accuracy of the annual estimates earlier and largely eliminate the proportionate allocation of the errors at the time the agency prepares

their benchmark revision of the survey. (Discussions of such changes are currently underway between BEA and BOC for their annual surveys.)

For series where benchmark revisions are not carried backwards in time, agencies should publish the amount of the sample "drift" since the last benchmark and adjust the historical series. If it is not possible for the agency to make these adjustment, then BEA and other users could make their own adjustments. The latter solution is less desirable because the agencies usually have information with which to make the adjustments that is not available to users, and when different users make the adjustments, they are likely to develop different adjustments.

Table 1.--Reliability of Current Estimates of Quarterly Changes in Real GDP, 1978-91
 [Percentages providing correct indication]

	All quarters			Omitting quarters with changes/differences of 1 percent or less	
	Direction of change	Larger/smaller change than in previous quarter	Change between +1 percent and +4 percent	Direction of change	Larger/smaller change than in previous quarter
	(56)	(55)	(56)	(48)	(43)
Advance.....	88	78	75	92	86
Preliminary.....	89	75	70	94	81
Final.....	89	76	66	94	84

NOTE.--The number of comparisons is shown in parentheses.

Table 2.--Dispersion in Revisions in the Quarterly Changes
in GDP and Its Components
[Percentage points/1/]

	Current dollars		Constant dollars	
	1978-82	1983-91	1978-82	1983-91
Gross domestic product:				
Advance.....	1.93	1.17	1.64	1.25
Preliminary.....	1.82	1.14	1.72	1.27
Final.....	1.82	1.15	1.75	1.33
Personal consumption expenditures:				
Advance.....	1.71	1.40	1.88	1.37
Preliminary.....	1.70	1.41	2.00	1.27
Final.....	1.80	1.35	2.12	1.30
Durable goods:				
Advance.....	5.72	4.20	5.00	3.96
Preliminary.....	5.15	3.88	5.11	3.63
Final.....	5.42	3.97	5.05	3.98
Nondurable goods:				
Advance.....	2.31	1.74	1.75	2.26
Preliminary.....	2.51	1.45	2.37	2.10
Final.....	2.50	1.37	2.39	2.03
Services:				
Advance.....	1.78	1.37	1.38	1.39
Preliminary.....	1.90	1.51	1.50	1.36
Final.....	1.96	1.59	1.56	1.42
Gross private domestic investment:				
Advance.....	13.20	9.38	10.64	9.53
Preliminary.....	12.67	8.62	10.24	9.30
Final.....	12.11	8.68	10.75	9.32
Fixed investment:				
Advance.....	7.01	3.03	5.59	3.74
Preliminary.....	4.96	2.43	4.08	3.29
Final.....	4.45	2.77	3.82	3.64
Nonresidential:				
Advance.....	8.24	3.67	6.36	4.42
Preliminary.....	5.63	3.19	4.15	4.07
Final.....	5.11	3.20	3.62	4.56
Structures:				
Advance.....	13.01	6.39	9.01	5.33
Preliminary.....	9.90	4.54	6.18	4.13
Final.....	9.47	4.92	6.10	4.66

**Table 2.--Dispersion in Revisions in the Quarterly Changes
in GDP and Its Components--Continued**
[Percentage points/1/]

	Current dollars		Constant dollars	
	1978-82	1983-91	1978-82	1983-91
Producers' durable equipment:				
Advance.....	7.09	4.02	6.65	5.21
Preliminary.....	5.17	3.87	4.85	5.39
Final.....	4.20	3.99	4.42	5.77
Residential:				
Advance.....	7.17	4.84	6.91	5.27
Preliminary.....	8.56	4.91	8.67	5.11
Final.....	7.63	4.98	7.89	5.22
Change in business inventories...
Net exports of goods and services:				
Exports:				
Advance.....	8.90	5.49	7.52	5.33
Preliminary.....	8.80	4.72	7.87	4.85
Final.....	8.02	5.19	7.07	5.67
Imports:				
Advance.....	5.48	8.12	7.21	8.92
Preliminary.....	4.98	7.24	5.64	9.29
Final.....	4.71	7.55	5.71	9.61
Government purchases:				
Advance.....	4.25	3.93	3.46	4.83
Preliminary.....	4.37	4.05	3.62	4.79
Final.....	4.34	4.05	3.75	4.89
Federal:				
Advance.....	11.40	9.09	10.36	10.70
Preliminary.....	12.29	9.11	10.48	10.49
Final.....	12.81	8.92	10.99	10.58
State and local:				
Advance.....	2.51	1.53	2.15	1.41
Preliminary.....	2.61	1.63	2.17	1.62
Final.....	2.34	1.65	2.20	1.60

1. Calculated from quarterly percentage changes at seasonally adjusted annual rates.

Table 3.--Annual Averages of Dispersion and Bias in Revisions in the Quarterly Changes in GDP
[Percentage points/1/]

Year	Dispersion				Bias							
	Advance to Preliminary	Preliminary to Final	Advance to Final	Advance to Preliminary	Preliminary to Final	Advance to Final	Advance to Preliminary	Final				
									To third annual revision estimate			
Current-dollar estimates												
1978...	0.5	0.6	1.0	2.1	1.9	1.9	-0.5	-0.2	-0.7	-2.0	-1.6	-1.4
1979...	.5	.3	.3	1.2	1.1	1.0	-.4	.1	-.3	0	.4	.3
1980...	1.0	.5	.8	.7	1.1	1.3	.3	-.1	.2	-.3	-.6	-.5
1981...	2.0	.4	2.3	3.4	2.0	1.9	-2.0	-.2	-2.3	-3.1	-1.0	-.8
1982...	.7	.7	.3	1.8	1.3	1.7	.7	-.5	.1	.7	0	.5
1983...	.5	.2	.7	2.4	2.0	1.8	.3	0	.3	.5	.2	.2
1984...	.7	.4	1.0	1.6	1.8	1.7	-.4	-.3	-.6	.1	.5	.7
1985...	.4	.4	.8	.8	1.2	1.6	.4	.4	.8	-.8	-1.2	-1.6
1986...	.5	.2	.4	1.3	1.5	1.6	0	.2	.2	-.4	-.3	-.5
1987...	.6	.4	.8	1.2	.6	.8	-.6	0	-.6	-1.2	-.6	-.6
1988...	.7	.2	.9	1.4	.7	.5	-.7	-.2	-.9	-1.4	-.7	-.5
1989...	.6	.4	.8	1.3	1.3	1.3	0	.3	.3	.6	.7	.4
1990...	.2	.4	.6	NA	NA	NA	.2	.4	.6	NA	NA	NA
1991...	.4	.3	.4	NA	NA	NA	0	.2	.2	NA	NA	NA
1992...	.9	.2	.9	NA	NA	NA	-.9	-.1	-.9	NA	NA	NA

Table 3.--Annual Averages of Dispersion and Bias in Revisions in the Quarterly Changes in GDP--Continued
[Percentage points/1/]

Year	Dispersion				Bias							
	Advance to Preliminary	Preliminary to Final	Advance to Final	Advance To third annual revision estimate	Advance to Preliminary	Preliminary to Final	Advance to Final	Advance Preliminary	Final			
										To third annual revision estimate		
Constant-dollar estimates												
1978...	.2	.5	.7	2.4	2.3	2.0	-2	-.2	-4	-1.4	-1.2	-1.0
1979...	.8	.2	.6	1.0	.8	.8	-.6	.1	-.5	-.5	.1	0
1980...	.5	.7	.7	.5	.5	.8	.3	-.3	0	-.2	-.5	-.2
1981...	1.0	.3	1.2	2.0	1.3	1.1	-1.0	-.2	-1.2	-1.8	-.8	-.6
1982...	.6	.6	.2	2.4	1.8	2.4	.3	-.5	-.2	1.0	.8	1.2
1983...	.4	.4	.7	1.5	1.4	1.1	0	-.1	-.1	-.3	-.3	-.2
1984...	.7	.4	1.0	1.5	1.8	1.8	-.3	-.1	-.4	-.2	.5	.6
1985...	.7	.5	.7	1.5	1.7	2.1	.2	.5	.6	-1.5	-1.7	-2.1
1986...	.4	.4	.2	2.1	2.1	2.3	0	.1	.1	0	0	-.1
1987...	.3	.3	.3	1.2	1.1	1.3	-.1	-.1	-.2	-1.2	-1.0	-.9
1988...	.5	.1	.6	.9	1.2	1.3	-.5	0	-.5	-.9	-.4	-.4
1989...	.6	.2	.7	1.7	1.6	1.6	0	0	0	1.1	1.1	1.0
1990...	.3	.3	.6	NA	NA	NA	.3	.3	.6	NA	NA	NA
1991...	.5	.3	.5	NA	NA	NA	.1	.3	.4	NA	NA	NA
1992...	.7	.3	.6	NA	NA	NA	-.7	.1	-.6	NA	NA	NA

NA-Not Available
1. Calculated from quarterly percentage changes at seasonally adjusted annual rates.

Table 4.--Dispersion in Revisions in the Quarterly Changes
in GDP and Its Components, 1981-90/1/
[Percentage points/2/]

	Current dollars	Constant dollars
Gross domestic product:		
Advance.....	1.38	1.32
Preliminary.....	1.31	1.36
Final.....	1.34	1.45
First Annual.....	.84	.83
Second Annual.....	.85	1.08
Third Annual.....	.65	.85
Personal consumption expenditures:		
Advance.....	1.48	1.47
Preliminary.....	1.50	1.39
Final.....	1.51	1.49
First Annual.....	1.32	1.19
Second Annual.....	.82	.71
Third Annual.....	.50	.47
Durable goods:		
Advance.....	4.49	4.22
Preliminary.....	4.15	4.16
Final.....	4.46	4.55
First Annual.....	3.27	3.41
Second Annual.....	1.79	1.91
Third Annual.....	1.84	1.94
Nondurable goods:		
Advance.....	1.85	2.15
Preliminary.....	1.61	2.09
Final.....	1.56	2.04
First Annual.....	1.36	1.38
Second Annual.....	1.10	1.03
Third Annual.....	.63	.63
Services:		
Advance.....	1.46	1.34
Preliminary.....	1.62	1.35
Final.....	1.74	1.45
First Annual.....	1.62	1.23
Second Annual.....	1.11	.97
Third Annual.....	.70	.58

Table 4.--Dispersion in Revisions in the Quarterly Changes
in GDP and Its Components, 1981-90/1/--Continued
[Percentage points/2/]

	Current dollars	Constant dollars
Gross private domestic investment:		
Advance.....	11.38	10.48
Preliminary.....	10.37	9.93
Final.....	10.48	10.13
First Annual.....	7.10	7.21
Second Annual.....	5.28	5.12
Third Annual.....	4.88	5.22
Fixed investment:		
Advance.....	3.97	4.22
Preliminary.....	3.28	3.83
Final.....	3.38	3.99
First Annual.....	3.09	3.27
Second Annual.....	2.59	2.93
Third Annual.....	2.09	1.87
Nonresidential:		
Advance.....	4.61	4.84
Preliminary.....	3.99	4.56
Final.....	3.73	4.71
First Annual.....	4.21	4.50
Second Annual.....	3.99	4.65
Third Annual.....	2.37	2.06
Structures:		
Advance.....	7.98	6.19
Preliminary.....	6.17	4.73
Final.....	6.48	5.16
First Annual.....	6.66	5.78
Second Annual.....	6.49	5.87
Third Annual.....	5.07	4.31

Table 4.--Dispersion in Revisions in the Quarterly Changes
in GDP and Its Components, 1981-90/1/--Continued
[Percentage points/2/]

	Current dollars	Constant dollars
Producers' durable equipment:		
Advance.....	4.59	5.56
Preliminary.....	4.10	5.61
Final.....	3.85	5.66
First Annual.....	4.59	6.30
Second Annual.....	3.79	5.31
Third Annual.....	1.40	2.84
Residential:		
Advance.....	6.30	6.61
Preliminary.....	6.48	6.58
Final.....	6.35	6.46
First Annual.....	4.69	5.72
Second Annual.....	3.81	4.22
Third Annual.....	3.54	3.62
Change in business inventories.....
Net exports of goods and services:		
Exports:		
Advance.....	6.13	5.80
Preliminary.....	5.57	5.29
Final.....	5.63	5.61
First Annual.....	4.41	4.25
Second Annual.....	3.09	3.52
Third Annual.....	3.73	4.20
Imports:		
Advance.....	7.99	8.74
Preliminary.....	7.85	9.38
Final.....	7.95	9.46
First Annual.....	4.76	5.50
Second Annual.....	3.44	3.94
Third Annual.....	3.25	4.80

Table 4.--Dispersion in Revisions in the Quarterly Changes
in GDP and Its Components, 1981-90/1/--Continued
[Percentage points/2/]

	Current dollars	Constant dollars
Government purchases:		
Advance.....	4.22	4.74
Preliminary.....	4.46	4.87
Final.....	4.51	5.01
First Annual.....	4.19	4.49
Second Annual.....	3.52	3.75
Third Annual.....	2.91	3.21
Federal:		
Advance.....	10.01	11.30
Preliminary.....	10.54	11.48
Final.....	10.61	11.78
First Annual.....	10.45	10.75
Second Annual.....	8.75	9.55
Third Annual.....	7.14	7.55
State and local:		
Advance.....	1.66	1.56
Preliminary.....	1.81	1.76
Final.....	1.83	1.71
First Annual.....	1.49	1.43
Second Annual.....	1.18	1.17
Third Annual.....	.85	.71

1. As previously indicated in the text, for the comparisons shown in this table, it was necessary to use the period 1981-90. In addition, because the annual revisions in 1985 and 1991 were replaced by comprehensive revisions, the comparisons exclude years with no annual revision. The first annual revision comparisons exclude 1984 and 1990, the second annual comparisons exclude 1983 and 1989, and the third annual comparisons exclude 1982 and 1988.

2. Calculated from quarterly percentage changes at seasonally adjusted annual rates.

RAISING THE NATION'S UNEMPLOYMENT RATE

John E. Bregger
U.S. Bureau of Labor Statistics

I. Introduction

Data released on February 4, 1994, reflected major revisions in the questionnaire and collection methodology that were introduced into the Current Population Survey (CPS), following a planning and developmental process extending over the previous 8-year period. Looking back over these 8 years, was the process worth the effort and cost? Did the data improvements exceed the losses engendered by the breaks in time series? Were public understanding and appreciation of the data negatively affected by the changes? These and many other questions will hopefully be answered in this paper.

II. What happened?

In a nutshell, the nation's overall unemployment rate was found to be somewhat higher in early 1994 than it was in late 1993, resulting directly from a wholesale, stem to stern, series of changes to the survey questionnaire and the total conversion to computer-assisted interviewing. In addition to the rate of unemployment, a number of other important data series were affected by the CPS revisions, including the estimation of discouragement and of persons working part time involuntarily.

Based on tests of the new system, the effect on the overall annual average unemployment rate for 1993 was estimated to be about half a percentage point. That is, utilizing data gathered from a totally separate, parallel, survey, the newly redesigned questions, asked by interviewers using laptop computers or calling from a centralized interviewing facility, identified more people unemployed than under the then current procedures. And if that wasn't dramatic enough, changes in definition of discouragement caused the figure to tumble some 60 percent, and measurement refinements lowered economic part-time employment by 20-25 percent. In other words, of the three most important measures of labor market slack, the one most people point to and talk about -- the unemployment rate -- has now been raised (on a statistically significant basis), while the other two have been lowered markedly. Early results (January-April 1994) from the implementation of the new survey and procedures confirmed the direction of the expected changes.

III. History

How (or why) did this come about? The answer to this question comes in several parts, involving secular changes in the economy, a Presidential commission, the advent of new technologies and surveying knowledges, and, of course, careful planning. These will all be briefly described in a whirlwind tour of the historical backdrop for changing the Current Population Survey.

First, the secular changes. The last time that the questionnaire had been changed to any degree was in 1967, resulting from a period of research in the aftermath of the President's Committee on Employment and Unemployment Statistics (the Gordon Committee). In the subsequent years, many societal changes have taken place, including the more prominent role of women, especially mothers, in the labor force; the continuing shift from a goods- to a service-producing economy; changes in the way business operates, such as opening and closing hours; and, somewhat related to the other factors just cited, shifts in the nature of employment, including more part-time work and less permanent attachment of employees to their employers.

The next Presidential commission to study the statistics, the National Commission on Employment and Unemployment Statistics (the Levitan Commission), issued its report, including a number of recommendations in the labor force area, on Labor Day 1979. For our purposes, its most significant recommendation was for major conceptual changes in the way we measure labor market discouragement. And, while this particular recommendation was accepted for implementation by Secretary of Labor Donovan two years later, it was not implemented, owing to a lack of available funding in the early 1980s for a parallel test panel of households for testing potential questionnaire changes.

At the same time that these developments were going on, there have been many innovations in the way data are collected, innovations that could be expected to improve the quality of data. Foremost among these have been the recognition of the relevance of the theories and methods of cognitive psychology in designing survey instruments and the use of the computer in the interviewing process. With respect to cognitive psychology, under the auspices of the Committee on National Statistics, National Research Council, psychologists, other behavioral scientists, and survey methodologists had come together in 1984 to discuss the contributions that each discipline could make to survey design and, in so doing, helped to launch the cognitive aspects of survey methodology movement. One of the legacies of that advanced seminar is a four-component cognitive model of the question-response process -- comprehension,

retrieval, judgment, and response -- that has provided a very useful framework for designing and evaluating survey questions.

Use of the computer for data collection has been around for some time, but perhaps not for such a large undertaking as a monthly sample survey of 60,000 households, and certainly not for use in personal decentralized interviewing. Testing had suggested that, not only was a large-scale application doable, but, more importantly, it offered incredible gains in a variety of ways. Among these were fewer constraints on the number or variations of questions that could be included in the instrument; greater accuracy of data collection, in that interviewers were more likely to ask questions as worded (some had been anticipating what question they would be asking next); and accuracy of data transcription and transmission. When coupled with the desire to change and add questions to improve overall accuracy of identifying labor force status, the potential for improvement was ever so much greater, because the computer could permit intricate skipping and the storage of earlier information for later use that no interviewer could carry out in a pencil and paper environment.

The planning process for carrying out all of this commenced with a series of conferences involving the senior staffs of the Bureau of Labor Statistics and the Census. The two agencies held a series of meetings beginning in April 1986 and two years later had a detailed plan to redesign the CPS, essentially in its entirety, with the questionnaire-related changes being the centerpiece. Budget submissions, with extensive, year by year, spending plans, were sent forward to OMB in time for the 1990 budget cycle. And, with favorable indications of approval forthcoming, work actually began in late 1988 toward a comprehensive survey redesign, with 1994 being the principal target date.

With respect to revisions to the questionnaire, a number of BLS-Census work groups were set up to develop a new questionnaire. The questionnaire was to be designed under the following guidelines: 1) It would not be constrained by the one-page limitation then in place; 2) it would take advantage of all aspects of automated data collection; 3) it would build upon recommendations of the 1978-79 Presidential Commission (and, to a lesser degree, the 1961-62 Commission); and 4) it would utilize to the maximum extent possible the knowledges available from cognitive science. Behavioral science laboratories were established in both agencies that brought in volunteers from the outside to react to various questions and question sequences. Questions explored included: What are the meaning of terms such as "work," "last week," "layoff," and "private company?" What method of collecting information on

the "actual number of hours worked last week" produces the most accurate data? How can response options be revised to simplify reporting and improve the categorization process -- and, in doing so, reduce measurement error? How could sensitive questions, such as on earnings, be revised to minimize nonresponse and improve reporting accuracy? How might the process of verifying information from a prior month's interview, rather than asking for the same information every month affect the quality and accuracy of the data? By the time we were through, we had managed to come up with satisfactory answers, for ourselves at least, to most of these as well as many other questions.

As a new questionnaire began to take shape, field testing became the next order of business. By this time (1990), two alternative versions of a potential new set of questions were in hand, along with the CPS questions then in use. The goal was to determine the best overall question wording from the three. To do this, the two Bureaus conducted a computer-assisted telephone interview (CATI), random digit dialing (RDD) test at the Census Bureau's centralized interviewing center in Hagerstown, Maryland. The first phase of this CATI/RDD test extended from July 1990 to January 1991, involving approximately 72,000 persons. Its purpose was to compare the then current version of the CPS questionnaire with the two test versions.

The principal product of the first phase was the selection of a single alternative questionnaire, close to the official version now in place, with appropriate additions and improvements that were deemed necessary due to the results of the testing. A second test phase was conducted between July and October of 1991 with approximately 30,000 persons, again via CATI/RDD; with very limited changes, this became the final version to be used in an 18-month parallel survey. During both the phase-one and phase-two testing, as well as the parallel survey phase, researchers employed a variety of methodologies to evaluate alternative question formats. These included respondent debriefings (via follow-up probe questions and vignettes), interviewer debriefings (via focus groups and debriefing questionnaires), response-distribution analysis, item nonresponse analysis, and behavior coding.

IV. The parallel survey

As researchers have long understood and as was once again verified in the CATI/RDD testing, if one or more important questions are changed (even slightly) in a continuing survey setting, we can expect different results. In other words, changing several questions in the CPS could be expected to have an effect on major measures, such as the rate of unemployment. Since the total number of potential questions increased from about 45 to 128 (no one, of course,

ever is asked more than a few of these!) and the wording of almost every question was changed, there was a virtual guarantee that we should expect differences on most of the statistical measures emanating from the survey. Thus, it was more than prudent to plan for a parallel, or overlap, survey for an adequate period of time in order to get some handle on the pact of these changes.

Ideally, we would have liked to have had a parallel sample extending for at least 2-1/2 years, with the same number of households as the ongoing CPS. This would have guaranteed a fully seated set of sample data for a full year, in terms of the 4-8-4 rotation group pattern. But, because our funding was limited, we had to settle for a 12,000 household sample covering the 18-month period, July 1992 through December 1993. Termed the CATI-CAPI Overlap Survey (CCO) internally and the Parallel Survey (PS) externally, this survey introduced the laptop computer (the CAPI portion of the CCO) into large-scale data collection.

One of our initial concerns was how well interviewers would adapt to using laptop computers and whether respondents would react favorably as well. We need not have worried: Both groups seemed to be happier. Interviewers, while concerned that questions did not pop up on their screens fast enough, appreciated the accuracy of the computer and thought that using it made them appear more professional. Respondents who were interviewed in person appeared to be more interested in the survey -- some, for example, invited interviewers to "plug in" -- and paid closer attention to the questions.

Ideally, with changes to the ultimate CPS coming from two directions -- the questionnaire itself and computer-assisted interviewing -- it would have been desirable to isolate the data effects on differences of these changes (questionnaire and interview mode). Alas, this was not possible. As a consequence, the significant differences between the on-going CPS and the PS that were identified can only be ascribed to the overall change in the survey and not specifically to the questionnaire or collection mode. Thus, we have been unable to discern, for example, what the specific effects have been on, say, the overall unemployment rate from changes in the questionnaire wording and question sequencing.

As soon as early PS figures started becoming available to BLS and Census researchers, it became obvious that we were indeed seeing marked changes in important statistical measures. The overall unemployment rate was higher, particularly among women and older workers but essentially across all worker groups. The employed differences were especially interesting, because more women but fewer men were found to have jobs.

There were also other dramatic changes. As expected, the new, more restrictive, measurement of discouraged workers resulted in some 60 percent fewer persons being counted in that category. To be classified as discouraged under the revised scheme, persons who wanted a job but had not looked for work in the prior 4 weeks had to have searched for work during the prior 12 months and not currently looking for work because of discouragement over the job market, while also being available to take a job during the reference week. Similarly, as a result of better question specificity, there was a 20-25 percent reduction in the number of persons working part time for economic reasons (that is, working less than 35 hours during the reference week because of poor business conditions or because of an inability to find full-time work). To be so classified, a person who usually works part time must now indicate that s/he wants a full-time job and was available to take one during the reference week.

V. Communications with official Washington

As soon as the researchers were able to verify that they had accurate data from the PS and thus could estimate the differences that the new questionnaire and collection methodology were yielding, taking appropriate measures of statistical significance into account, it was time to start communicating "up the line." We were, quite naturally, concerned about what kind of reactions there would be to a significantly higher rate of unemployment. There was, after all, a new administration in office that perhaps did not need to be saddled with yet another major issue. It had enough on its plate already.

The researchers had put together a formidable package of tables and analysis, with explanations for the many diverse changes observed over the comparison periods. Initially, 6-month comparisons of the PS with official CPS, covering the period September 1992-February 1993, were utilized, and these were the first figures to be viewed at higher levels, first of all with the heads of the two agencies. Soon thereafter, the first annual average data became available, representing the period September 1992-August 1993. Using these figures, a memorandum detailing the changes that we were expecting to introduce in January 1994 and the expected data effects was sent to Secretary Reich in late October, and this memorandum was forwarded on by him to the President. An hour-long meeting was held with Secretary Reich and top Department of Labor staff on November 1, and this was followed in short order by other high-level briefing sessions with other members of the Administration and the Federal Reserve Board.

Reaction was reasonably swift. All of the changes that were contemplated for implementation in January 1994 were fully acceptable, despite a concern that, with a higher rate of unemployment, the public might fail to recognize that the economy was still gaining steam. Indeed, it was this concern that led directly to a request to sustain the collection of parallel survey data using the paper and pencil methodology beyond the year and a half that had been planned and funded. Monies were found and commitments made to sustain the parallel survey beginning in January 1994, this time with the old questionnaire and procedures. That way, after an initial period where respondents and interviewers might be affected by the previous test, we would have a continuing measurement of the differences that were identified for 1993 as the data on the new basis became available.

VI. Communications with the outside world

A detailed planning document that had been produced and constantly updated had long identified November 16, 1993, as the date of the first public announcement of the plans for introducing changes into the Current Population Survey and detailing what the expected data effects were. Armed with briefing packets and a plethora of other useful information, Commission Katharine Abraham and members of the BLS and Census Bureau staffs presented an extensive array of information to the national economic media. Articles appearing throughout the country the next morning, as well as the more immediate wire service stories, suggested that the press well understood what was transpiring. In particular, the notion of "gender bias," which had emerged from data findings, was significantly played up. With few exceptions, they got it right. Not all did, however, as suggested by the headline, "U.S. won't ask women if they cleaned house."

BLS didn't stop with a one-day media session in Washington, D.C. There was a full-day session with technical users on the next day (November 17), also in Washington, which some 150 persons attended. Interest was running high. In December and January, combined data user-media sessions were held in 13 other large cities throughout the entire country, including New York, Chicago, and Los Angeles. Many people turned out to learn what was expected to happen and how their local unemployment rates might be affected.

By this time, we had unemployment rate comparisons for the Census regions and divisions, as well as some data for seven large states. Our uncertainty was quite high as to the reliability of our sub-national comparisons, and this was carefully communicated. Fortunately, our concerns did not fall on deaf ears, and most people, including the local

media, did not play up some of the wide differentials, such as rates that were slightly more than a point higher in the Middle Atlantic Division (covering New York, New Jersey, and Pennsylvania). This was all the more fortunate when, in actual fact, this particular region did not show large jumps in early 1994. In other words, our concerns about the reliability of sub-national data were well-founded, especially since the PS, unlike the CPS, was not a State-based design and the sample size was so much smaller.

VII. The final outcome and lessons learned

By the time the data for January 1994 were released on February 4th, it seemed that everyone -- government officials at all levels, the media, financial analysts, and the public at large -- was well aware that big changes were to be expected. And thus there were seemingly no surprises. The overall rate of unemployment for January was 0.3 percentage point higher than the December 1993 figure, quite reasonable, given expectations that we could expect as much as an 0.6 increase on an annual average basis (all other things remaining equal), 0.5 from the questionnaire and methodological changes and 0.1 from the introduction of the 1990 census-based population figures (adjusted for the estimated undercount) into the estimation procedures.

Did these data results from January, which were followed by an 0.2 percentage point drop in February, imply that our expectations based on parallel survey results -- the population effect was "guaranteed" -- were too large? Or, did January and then February really show large improvements in the underlying rate of unemployment? Even with two more months in, I think that we are still waiting to answer these two questions with more data. (Isn't that always the case?) It would appear, however, that our expectations for 1994 results for the official figures are essentially accurate, that is, the new questions and methodology suggests that the old questions did a good job of measuring mainstream labor market behavior, but not as well for more marginal types of activities, such as might be typical for certain women, youth, and older persons, for whom more jobseeking and more jobholding were found. Now, these missed activities tend to be of a seasonal nature and thus more likely to occur in certain months of the year. January, February, and March are months for which this sort of seasonality is fairly low; it can be expected to be much higher in months like May, June, and July. This implies that we can therefore anticipate higher levels of activity, particularly jobseeking activity, in the spring and summer months. And our seasonal adjustments, which are for the moment necessarily based on experience under the former procedures through the end of 1993, are somewhat "off." So, the answer to the second question regarding the January and February (plus March and April) results would appear to be

that, yes, we were seeing some improvement in the economy, but perhaps not quite as good as implied by these figures.

These early results also suggest that it will be quite some time before we have a full, clear realization as to all of the data effects that have been brought about with the new CPS. Ideally, for example, we should have had a longer lead-time than a year and a half with the parallel panel, so that it could have settled in better and given us more direct comparisons with the official CPS figures. Budget exigencies rarely resolve the hindsight "shouldas." New seasonal patterns are not fully discerned for at least 5 years, and we therefore may have to wait that long to expect to attain a degree of accuracy in month-to-month movements in employment and unemployment that we are fully comfortable with. The seasonal adjustment process will improve over time as data based on the new procedures are gradually taken into account. It is also possible that other improvements could be made to expedite the process.

The seasonality issue just discussed and the potential breaks in series for a number of measures, particularly those of labor market slack, inevitably raise the question as to the whether the process should have been embarked on at all. From my own viewpoint, the answer is clear: Breaks in series and comparatively short losses in time-series comparisons, while never desirable with any degree of frequency, are vitally necessary to ensure that we are accurately measuring what is occurring in our economy. We must recognize that there is always a cost to bring about improvements in data collection of economic phenomena. If we focus only on data consistency and therefore take our eyes off the prize of data improvement in a constantly changing society, we will never even attempt to undertake improvements in the measurement and collection of statistical surveys in the first place. Once undertaken, it is imperative that we go all the way, that is, make all of the improvements that are discernable and viable and then carefully measure their impact through a separate parallel survey. That is precisely what we have done with the CPS, and I firmly believe that the payoff was well worth the short-term losses that we are experiencing.

Perhaps the most significant lesson we learned from all of this was one that was a major winner: full, extensive communication. By careful interaction with, firstly, our internal customers -- i.e., the Administration and the Congress -- and then our external customers -- the media and the public -- there were few, if any, surprises. Friday, February 4th, turned out to be a business as usual, ho-hum day. Everyone knew or thought they knew what was going to happen, their expectations were more or less met, and thus not a lot of news was good news. My recommendation for any statistical agency undertaking major changes in surveys or

data series, therefore, is to err heavily on the side of both extensive and continuing communication with every possible group -- not just the media, not just here in Washington, but with everyone everywhere.

Carrying out the questionnaire-related redesign of the Current Population Survey cost the taxpayers an estimated \$30 million. Was it a worthwhile expense? Coming from a highly biased person, one can take my answer with a grain or two of salt, which is a resounding yes! In return, the Nation is getting better, more accurate figures on the labor force activities of the population. It is getting new kinds of important statistics, such as monthly data on labor market discouragement (on a totally revised conceptual basis) and on multiple jobholding. And it is getting the assurance that the measurements of the labor force, employment, unemployment, and those not in the labor force have been carefully studied and researched. And that, I would argue, is an incredibly great bang for our bucks.

A potential lesson that I hope we will not forget is that the total job is not as yet completed. The data comparisons for 1993, based on parallel survey and official statistics, need to be studied much closer than we have been able to thus far. The new figures for 1994 and beyond will require careful analysis. Continuing research on bridging data estimation both back in time and forward in time should continue, with the intent of assisting time series users in their analytical endeavors. We should have learned well the benefits that behavioral science has given us in terms of all future data collection; thus, for example, survey supplements, whether ongoing ones such as income and work experience or new ones such as the upcoming inquiry into contingent work, should be subjected to careful cognitive testing. Lastly, we should carry forward what we have learned into other CPS-related areas -- such as instituting improvements into the "control card," in which demographic characteristics are identified, or in improving the coverage of minority groups in the data collection process.

Finally, the last Presidential commission to examine labor force statistics issued its report 15 years ago. It took that long to implement a couple of its important recommendations. It would not be too radical to suggest that another commission ought to be established in the not too distant future to assess the viability and adequacy of the 1994 changes and then to determine appropriate future directions as we embark on the 21st century.

References

- Chester E. Bowie, Lawrence S. Cahoon, and Elizabeth A. Martin, Overhauling the Current Population Survey: Evaluating changes in the estimates, Monthly Labor Review, September 1993.
- John E. Bregger, The Current Population Survey: a historical view and BLS' role, Monthly Labor Review, June 1984.
- John E. Bregger and Cathryn S. Dipppo, Overhauling the Current Population Survey: Why is it necessary to change?, Monthly Labor Review, September 1993.
- Thomas B. Jabine, Miron L. Straf, Judith M. Tanur, and Roger Tourangeau, Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines, 1984 (National Academy Press, Washington, D.C.).
- Elizabeth A. Martin, "Surveys as Social Indicators: Problems in Maintaining Trends, Chapter 16 to Handbook of Survey Research, 1983 (Academia Press, Inc.).
- National Commission on Employment and Unemployment Statistics, Counting the Labor Force, Labor Day, 1979 (U.S. Government Printing Office, Washington, D.C.)
- Anne E. Polivka, Comparisons of Labor Force Estimates from the Parallel Survey and the CPS During 1993: Major Labor Force Estimates, CPS Overlap Analysis Team Technical Report 1, March 18, 1994.
- Anne E. Polivka and Jennifer M. Rothgeb, Overhauling the Current Population Survey: Redesigning the Questionnaire, Monthly Labor Review, September 1993.
- President's Committee to Appraise Employment and Unemployment Statistics, Measuring Employment and Unemployment, 1962 (U.S. Government Printing Office, Washington, D.C.)
- Jennifer M. Rothgeb, Revisions to the CPS Questionnaire: 1994 Effects on Data Quality, CPS Overlap Analysis Team Technical Report 2, April 6, 1994.
- Jenny Thompson, Mode Effects Analysis of Labor Force Estimates, CPS Overlap Analysis Team Technical Report 3, April 14, 1994.

COMMENTS ON PARKER AND WEADOCK, TIME SERIES REVISIONS:
THE EFFECTS ON GROSS DOMESTIC PRODUCT

Murray F. Foss
American Enterprise Institute

I. Introduction

This paper by Robert Parker and Teresa Weadock is an interesting study that extends the series of evaluations of quarterly GDP estimates undertaken in BEA by Allan Young in 1974 (and most recently in the October 1993 Survey of Current Business) and initiated by George Jaszi a decade earlier.

In my comments I will discuss some differences between the current and constant dollar figures. I suggest other things the authors might have looked at, some of which would be quite easy. I then raise the question about which figures are the best ones for gauging "the true quarterly change," a point implied by the authors in their criticism of source data. And then I ask what lessons we should learn from all of this.

II. Lack of symmetry between current and constant dollar measures

The Parker-Weadock (henceforth PW) measures of reliability are typically presented in terms of current dollars and constant dollars. We are interested in both but the two sets of figures are really not symmetrical, and I wish the authors had discussed this asymmetry. While the monthly PPI data are revised once going back 4 months as a result of the incorporation of late returns and the correction of errors by respondents and by BLS in the initial reports, this is not true of the current monthly CPI. Yet if we look at the succession of dispersion measures in Table 2 from advance to preliminary to final, for GDP as a whole as well as for personal consumption the dispersion gets worse in the constant dollar series.

Aside from making very few revisions on a monthly basis it is not the practice of BLS to conduct a bigger survey after the calendar year is over--what might be an Annual Survey of Prices, analogous to, say, the Annual Survey of Manufactures. Every ten years BLS changes its market basket for the CPI to take account of changes in consumption patterns. New and different products are appearing on the market constantly, and these BLS treats in a variety of ways, depending on continuing probability sampling to pick up new products and types of outlets.

It would be good to know whether this deterioration in the reliability of successive constant dollar estimates is statistically significant. Is it simply a reflection of newer seasonal factors, which are revised by BLS each year going back several years, and to what extent does it reflect a benchmark (10-year) change?

III. Some other perspectives of reliability

PW, like Allan Young, examine the reliability of the quarterly estimates of GDP from an historical point of view. That is useful because it permits one to say something about possible long-term trends in reliability. But outlined below are other ways that I would like to see examined.

A. Business cycle perspective

The GDP statistics are the single most important indicator about what is going on in the national economy. But as we all know economists are often in the dark about whether an expansion has begun or whether the economy has slipped into a recession. So I would like to see how these reliability measures--both dispersion and bias--behaved around turning points. Looking at the historical record we can ask if there are any patterns, for example, in the four quarters up to and including the business cycle peak and in the first four quarters of downturns (which average not quite a year in length). Do these patterns differ from one another? Would they differ from the pattern in the first four quarters of an upturn? It would not be hard to find a rationale for any differences that might turn up; for example something concerning the quality of statistics within the firm over the business cycle but any patterns would be of interest in themselves.

B. Inflation

It would be interesting to examine the data for possible differences when the rate of inflation differed. Is there a difference between 1973-80, when inflation was very high and 1983-90, when inflation was much lower? It is more difficult to capture a change in real output when inflation is high than when it is low. When buyers resort to new sources of supply or when sellers change their discounts from list prices, the Producer Price Indexes may be slow to adapt even though the current dollar figures on sales reflect these changes immediately.

C. The current data

The first three estimates of a given quarter--the advance, preliminary and final--carry a lot of weight because these are the figures that affect decisions by business and government in the short run. So one could use this criterion: given the advance and the final, how often did the preliminary move in the direction of the final (third) change? For example, if the advance change is 1.9 and the final is 1.5, we can ask if the preliminary moved down from 1.9 or exceeded it. Small misses in direction would be ignored, following the authors' approach.

D. Calendar quarters separately

It would also be a simple matter to collate the measures of dispersion and bias by calendar quarters, to see if the fourth quarter differs from the other three quarters. The reasoning behind this is that in organizing their work accountants put most emphasis on the annual report, which comes out a few months after the end of the year. (I realize that fiscal years pose some problem). Accountants do things at the end of the year that they don't do during the year. For example, they may take a physical count of inventories at year-end but use shortcut methods to estimate inventories for months and quarters. Earlier errors that show up at the end of the year are corrected in the final quarter. If the errors are in one direction the fourth quarter correction will tend to be reduced if not eliminated by seasonal adjustments. After year-end, accountants may go back and revise earlier quarterly figures; this may be a regulatory requirement but I am not certain. If accountants in fact do a lot of estimating during the year such a practice could give rise to revisions between advance-preliminary, on the one hand, and final-first annual, on the other.

E. Final sales and inventory change

Estimating the quarterly change in business inventories is an inherently difficult task and remains so even with the many improvements made by the Census Bureau and BEA over the years. The change in inventory change is ordinarily a significant part of the average quarterly change in GDP. The inventory estimates are not shown explicitly by the authors because of the particular measures they employ for GDP and all other components. It would seem from table 4 that revisions in inventories are a significant source of total revisions. It would be a good idea to examine a common measure published by BEA, namely, total final sales, which excludes inventory change. Obviously the shifting of farm inventories between the Federal Government and private business creates a problem but it would not seem to be too difficult to make allowance for this.

As a matter of fact, the change in business inventories ought to be shown explicitly with its own reliability measures because it is so difficult and involves much judgment not only by BEA but also, I would guess, at the firm level. This suggests an additional reason why the authors should show successive revisions of nonfarm CBI: the monthly CBI's are subject to far greater variation than any of the flow components. Maybe exports and imports as now calculated would be close runnersup. PW and Allan Young point out that the final month in the quarter has a weight of only one-ninth and the second and first months weights of two-ninths and three-ninths, respectively. If the expenditure components were random numbers, the fact that they have such "small" weight would not be so important. The fact is that this month's seasonally adjusted retail sales must be very close to last month's. A one percent

difference is a big seasonally adjusted change. But that is not true of CBI. One month of inventory change can be positive, the next month, negative. Such a pattern is possible because sales can be higher or lower than expected by the firm and, with production plans based on expected sales, inventories will be correspondingly lower or higher. The same is true of incoming supplies to the purchasing firm--a consequence of capacity limitations, strikes, natural disasters, etc. This is not to deny that during the expansion firms tend to build stocks and during the contraction they cut them. But I urge the authors to do the dispersion measures of nonfarm CBI (and the GDP) in constant dollars. This would be a good test.

IV. What is the "true" quarterly change?

Given the way quarterly data are revised to make them compatible with subsequent annual figures and benchmark annual totals from the quinquennial censuses, how can we be sure that the very final quarterly pattern that emerges is superior to all previously published quarterly data for a given year? Parker and Weadock criticize the Census Bureau for making proportional adjustments in originally published monthly and quarterly data. This is an old problem. For example, Morris Cohen raised the same issue at an Income and Wealth Conference 15 years ago. He said that the data were being oversmoothed and that cyclical fluctuations were being damped if not eliminated. The late Otto Eckstein agreed with this point of view but it remains a minority opinion. The dominant revision philosophy is above all to get the long-term trend correct. If that is so, cyclical fluctuations must be fitted into the trend (that is, benchmark) values for a given year.

The answer, of course, is to get more and better within-the-year data. There is no substitute for this. That was said 15 years ago and, I am sure, many times before that. People who make decisions in business and government have a big stake in the currently available quarterly numbers and after some 50 years deserve more improvements than the agencies have made. Economists studying the business cycle have an important interest in getting the record straight.

V. What lessons should we learn from these studies?

Parker and Weadock, like Allan Young, raise the possibility of dispensing with the second and third quarterly estimates for a given quarter. A single current estimate, namely, the advance, might conceivably save some money. I am not sure that the nation would be better off. As Allan Young points out, the detailed estimates might suffer. As for the total, there are so many people and firms engaged in this business today that several estimates would make their appearance to fill the void left by BEA. Unfortunately these estimates would differ from one another. I would guess that large organizations like the Federal Reserve would

make their own estimates. Estimates of GDP made by outsiders are not likely to be as good as those made by BEA. So I am inclined to stick to the present system. It is less bad than what might supplant it.

Parker and Weadock are impressed by the improvement in reliability from final current to the first July estimate. They should, of course, find out how much of the improvement is due to better seasonals. But they should not be surprised that firms send in better data after the year is over.

I think that both the source agencies and BEA should do more field work to find out why numbers submitted to the government change. The agencies ought to do more to find out how firms obtain their monthly and quarterly numbers, the extent to which they rely on within-the-firm estimates and end-of-year adjustments. Sales probably are not much of a problem. But other statistics like inventories and profits may be. Irving Rottenberg and I found that firms using LIFO accounting had great trouble estimating inventories on a monthly basis. The fact is that firms use shortcuts of all sorts; these are doubtless useful to the firm but may be subject to biases that are functions of the stage of the business cycle or the rate of inflation. I don't have in mind huge field surveys checking data quality. One can get many insights from small field trips. I believe that BEA at least does too little in this regard. And, to repeat, Census and BLS ought to do better in obtaining within the year figures to accompany more comprehensive annual and benchmark surveys.

**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 Measures 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 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 on Establishment Surveys (NTIS Document Sales, PB88-232921)
16. A Comparative Study of Reporting Units in Selected Employer Data Systems (NTIS Document Sales, PB90-205238)
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)
21. Indirect Estimators in Federal Programs (NTIS Document Sales, PB93-209294)
22. Report on Statistical Disclosure Limitation Methodology (NTIS Document Sales, PB94-165305)
23. Seminar on New Directions in Statistical Methodology (NTIS Document Sales, PB95-182978)

Copies of these working papers may be ordered from NTIS Document Sales, 5285 Port Royal Road, Springfield, VA 22161 (703)487-4650