Statistical Policy
Working Paper 23

Seminar on New Directions in Statistical Methodology

Part 3 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 Reseerve 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

Session 1 - ECONOMIC CLASSIFICATION REVISIONS

Session 2 - DISCLOSURE LIMITATION METHODOLOGY

Session 3 - CUSTOMER SURVEYS

Session 4 - ADVANCES IN DATA EDITING

Session 5 - TIME SERIES REVISION POLICIES

Part 2

Session 6 - INCENTIVES IN SURVEYS

Thursday, May 26

Session 7 - COMPUTER ASSISTED SURVEY INFORMATION COLLECTION

Session 8 - LONGITUDINAL SURVEYS

Part 3

Session 9 - COGNITIVE TESTING AND SELF-ADMINISTERED QUESTIONNAIRES

Session 10 - STATISTICAL USES OF ADMINISTRATIVE RECORDS

Session 11 - SMALL AREA ESTIMATION

Session 12 - NONRESPONSE IN SURVEYS

TOWARD AN AGENDA FOR THE FUTURE

Part 3

Session 9

COGNITIVE TESTING AND SELF-ADMINISTERED QUESTIONNAIRES

Literacy Limitations and Solutions for Self-Administered Questionnaires

Judith T. Lessler and James M. O'Reilly
Battelle Memorial Institute

Introduction. The self-administered paper questionnaire is a standard method for asking questions on sensitive subjects. Yet as much as one-fifth of the adult population of the U.S. has levels of literacy which may make using the typical SAQ futile. This paper reviews the research on the efficacy of self-administered questioning on sensitive subjects. Then we will discuss how low literacy or other cognitive burdens can limit the effectiveness of self-administration which require reading.

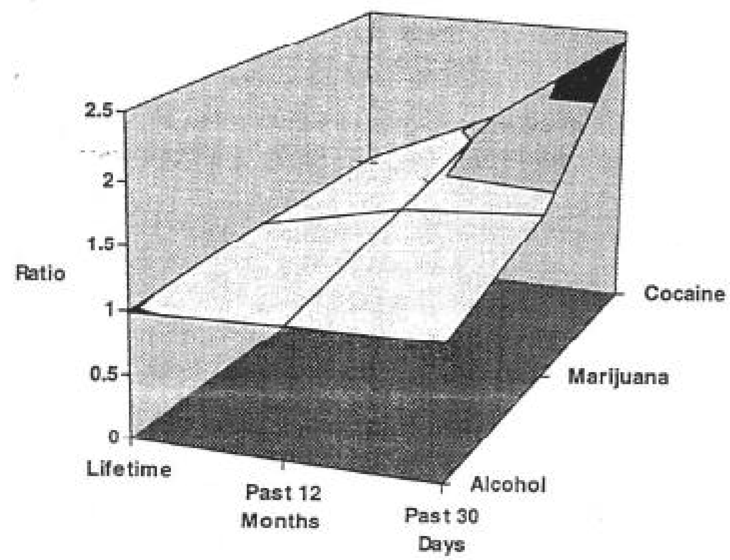
New computer technology, called Audio Computer Assisted Self-Interviewing (ACASI), now makes is possible to conduct self-administered interviews in which the text on the screen is accompanied by a high quality voice recording played over headphones. The first major field test of the technology was in the 500-person 1993 National Survey of Family Growth Pretest. We describe how ACASI was implemented in the study and the impact on abortion reporting and respondents' reactions to the technology.

Methods of Interviewing on Sensitive Subjects. Research has generally shown that more private methods of interviewing yield higher reports of sensitive behaviors (Bradburn, 1983; Miller, Turner, and Moses, 1990, Ch. 6; Catania, et al., 1990, and Schwarz et al., 1991). For example, Hay (1990) found differences in reported consumption of alcoholic beverages and cigarette use in a study of some 1500 students in grades 2 through 12 who were randomly assigned to receive either a SAQ or a personal interview. The differences were 74 versus 63 percent for over use of alcohol and 38 versus 30 percent for use of cigarettes. Turner, Lessler, and Devore (1992) in a large-scale field experiment, in which 3,200 respondents were randomly assigned to either an interviewer or self-administered questionnaire found that the difference between the two modes of data collection increased as the sensitivity of the behavior increased. Exhibit A shows the ratio between the proportion of SAQ respondents reporting a given behavior to the proportion of respondents reporting that behavior when the interviewer administered the questions. The exhibit displays the results for three time periods (lifetime, last 12 months, and last 30 days) and three types of drugs (alcohol, cocaine, and marijuana).

COPAFS2.DOC

Exhibit A. Ratio of Prevalence Estimates from SAQ and Interviewer-Administered Items

Drug Type	Lifetime	Past 12 Months	Past 30 Days
Alcohol	**0.99**	**1.04**	**1.06**
Marijuana	**1.05**	**1.30**	**1.38**
Cocaine	**1.06**	**1.58**	**2.40**



Examining this table, we note that as the sensitivity of the characteristic increases from alcohol to marijuana to cocaine, there is a concomitant increase in the superiority of the self-administered format relative to the interviewer-administered questions.

Cognitive Aspects of Conventual SAQs. Thus, self-administered questionnaires can have a positive impact on data quality because of the increased privacy. In addition, SAQs allow respondents to control the pace of the interview, and no additional variance is introduced by the interviewers.[1] However, conventional SAQs do have drawbacks. First and foremost, they require that the respondent can read.[2] In addition, the respondents must complete a number of the questionnaire administration tasks such as finding and reading instructions, implementing skip patterns, and marking answers. They are prone to the same types of errors that are seen in interviewer administered questionnaires—missing, out-of-range, and inconsistent answers. Lessler and Holt (1987) found that some respondents who could read the questions had difficulty understanding the conventions concerning recording of answers and movement through forms.

Cognitive testing of self-administered has noted problems in each of these areas:

Reading problems:

- Complete inability to read the questions
- Failure to understand specific terms or phrases

---

1 The presence of high levels of interviewer variance in the decennial census was one of the motivations for adopting a mailout-mailback self administered method for the census beginning in 1960.

2 The National Adult Literacy Survey (NALS) was conducted in 1992 using a nationally represented sample of 13,600 persons aged 16 and older. Literacy was measured in terms of five proficiency levels on three scales—prose, document and quantitative. The survey found that the percentage of adults in the lowest level of proficiency was 21 percent for prose literacy, 23% in document literacy, and 22% in quantitative literacy.

The lowest level of prose literacy is described as "Most of the tasks in this level require the reader to read relatively short text to locate a single piece of information which is identical or synonymous with the information given in the question or directive" (pg. 74). For document literacy the lowest level means: "Tasks in this level tend to require the reader either to locate a piece of information based on a literal match or to enter information from personal knowledge onto a document. Little, if any distracting information is present" (pg. 85). Quantitative literacy at the lowest level means: "Tasks in this level require readers to perform single, relatively simple arithmetic operations, such as addition. The numbers to be used are provided and the arithmetic operation to be performed is specified" (pg. 94).

National Center for Education Statistics, 1993. Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey. Washington, DC: U.S. Department of Education.

- Use of a time consuming two-step strategy for reading questions
  First the questions are read to decode the words and then they are read a second time to get the meaning of the sentences
  Complete inability to read the questions
- Not reading all of the questions or response categories in order to reduce the reading task

Questionnaire administration problems:

- Ignoring or neglecting to read instructions
- Difficulty finding the instructions
- Physical difficulties with marking answers that require filling circles for mark-sense forms or writing in small spaces
- Difficulty understanding or failure to follow skip instruc
- Missing questions
- Writing in illegible or out-of-range responses
- Failure to follow marking instructions
- Idiosyncratic response or marking conventions

The result of these difficulties is that researchers using SAQs typically simplify the questionnaires and avoid contingent questioning. Contingent questioning is avoided for two reasons. One is to reduce the chance that data is lost because of the errors that respondents make through incorrect implementation of skip patterns and the second is to increase the privacy of responses. For example, the National Household Survey on Drug Abuse uses SAQs to ask questions on use of alcohol, misuse of prescription drugs, and use of illegal drugs. In that survey, interviews are conducted in the respondent's home. Some of the questions are interviewer administered. On the more sensitive topics respondents are instructed how to complete the SAQs. Respondents are required to mark an answer for every question in these sections in order to (1) increase their privacy, (2) prevent errors in implementing skip instructions, and (3) eliminate the tendency for respondents to mark no on gate questions on use of a particular substance in order to reduce the response burden of answering detailed questions about the drug. It is believed that if respondents are allowed to skip answers, they will realize that interviewers are able to distinguish those who were and were not drug users and, as a consequence, be less truthful.

ACASI Technology. Audio computer assisted self-interviewing (ACASI) has been developed to overcome some of the difficulties associated with the response to self-administered questionnaires. When a computer-assisted self-administered interview (CASI) is used, the computer can take care of the "housekeeping" or administrative tasks for the respondent. By adding simultaneous audio renditions of each question and instruction aloud, ACASI can remove the literacy barriers to self-administration.

In CASI respondents read the questions as they appear on the screen and enter their answers with the keyboard (or some other device). In Audio-CASI, an audio box is attached to the computer and respondents put on headphones and listen to the answers as they are displayed on the screen. Respondents have the option of turning off the screen so that people coming into the room cannot read the questions, or turning off the sound if they can read faster than the questions are spoken, or keeping both the sound and video on as they answer the questions. Respondents can interrupt the question while it is in process.

Comparisons of CASI with personal interviews have noted findings similar to those cited above for the comparison of SAQs to interviewer administered questionnaires. Waterton and Duffy (1984) compared reports of alcohol consumption under CASI and personal interviews. Overall, reports of alcohol consumption were 30 percent higher under the CASI procedure, and reports of liquor consumption were 58 percent higher. This may understate the potential gains because in this study respondents were first asked if they had consumed any alcoholic beverages in the past seven days by an interviewer. Only those respondents who indicated that they had done so received the CASI interview.[3] Several recent studies comparing CASI to personal interviews in clinical settings have also noted the superiority of this method. Locke (1992) found significant differences between the reporting of HIV-risk behaviors when CASI was used to administered questions to donors at an American Red Cross donor center (4.4 percent versus 0.3 percent in the traditional interview procedure). Robinson and West (1992) compared reporting of symptoms in a genito-urinary clinic using CASI, SAQs, and physician interviews. They found that more symptoms were reported by computer than by paper, and both found more symptoms in physician interviews. Levine, Ancill, and Roberts (1992) found that patients who had been admitted to a hospital after harming themselves were more likely to report suicidal ideation in a computer interview than to a physician. The CASI version of the Diagnostic Interview Schedule (DIS) yielded diagnostic information consistent with the traditional interviewer administered DIS, and patients considered the computer contact to be less embarrassing (Erdman et al., 1992). A computer interview with sex offenders yielded large numbers of previously undetected crimes (Weinrott and Saylor, 1991), and a comparison of clinician and computer interviews directed at identifying obsessive compulsive disorders found that the two methods were equally good at distinguishing those with the disorder and that patients showed no preference of the clinician interviews (Rosenfield et al., 1992).

Formal comparisons of Audio-CASI versus other modes are just now being conducted. O'Reilly, et al. (in press) compared paper SAQs, CASI, and Audio-CASI in a small scale experiment designed to assess the potential for the technology. Subjects answered questions on drug use, sexual behaviors, and income. A greco-

---

3 In the literature, this study is often reported as a CAPI study. It was actually a CASI study in which computers were taken into the homes of respondents and asked to enter their responses on selected questions while the interviewer stood in a part of the room that did not permit observation of the respondent's answers.

latin square design was used to assign subjects to one of three interviewing modes for each topic producing an experiment that was fully balanced across mode and content. For eight of nine rating scales comparing these modes, respondents reported a preference for one of the two CASI methods. Sample sizes were quite small in this study (n=40); however, O'Reilly, et al. found that the CASI methods tended to produce significantly more reports of marijuana and cocaine use. Few differences in sexual behaviors were found.

Respondents were asked which method they thought was better for nine facets:

(1) Liked best
(2) Best for asking sensitive questions
(3) Easiest to change answers
(4) Most interest
(5) Easiest to use
(6) Best for getting honest answers
(7) Best for privacy after interview
(8) Best for privacy interview, and
(9) Overall preference

For all but number 3, the two CASI methods, audio and video-only, were rated significantly better. ACASI was rated consistently higher than video-only CASI. However the difference was significant for three items: overall preference, interest, and ease of use.

National Survey of Family Growth (NSFG) Pretest. Under funding from the National Center for Health Statistics (NCHS), scientists at NCHS, Battelle and the Research Triangle Institute (RTI) collaborated in a formal field experiment that compared abortion reporting under three different interviewing conditions. Respondents were randomly assigned to receive either an in-home CAPI interview only, an in-home CAPI interview followed by a Audio-CASI interview that asked additional questions about abortions, or an interview at a neutral site away from the respondent's home. The respondents in the Audio-CASI treatment were first asked to report their abortions to the interviewer during a section of the CAPI interview that asked about the outcome of each pregnancy that they ever had. The question asked:

Now I'd like to ask some questions about your Nth pregnancy.

Please look at Card B-1. Thinking about your Nth pregnancy, in which of the ways shown on Card B-1 did the pregnancy end?

(READ LIST. CODE ALL THAT APPLY.)

Miscarriage?	(Occurs naturally, during the first 6 months of pregnancy)
Stillbirth?	(Baby born dead after 7 or more months of pregnancy)
Abortion?	(Induced during the first 6 months of pregnancy; include D&C, vacuum extraction, suction, and saline injections)
Ectopic Pregnancy?	(Occurs outside the uterus or womb)

Livebirth by Cesarean section?
Livebirth by vaginal delivery? (Includes delivery through natural or induced labor)

At the end of the interview, respondents were trained in the Audio-CASI procedures and were asked additional questions on abortion.

Prior to the field experiment, the Audio-CASI interview was tested in the cognitive laboratory. Respondents were brought into the laboratory and were first asked to completely answer a series of questions from the NSFG. Following this, the use of the Audio-CASI implementation was explained to the respondent, and she answered questions using the computer herself. Since the questions on abortion were considered to be highly sensitive, we did not ask respondents to think-aloud during laboratory testing. Instead, the interviewer stood across the room from the respondent and was asked to describe what the respondent was doing as the respondent listened and answered questions. Thus, the respondent reported things like:

- She is reading the first question[4]
- I am putting in my answer
- I made a mistake, and I am backing up
- She is reading the next question
- I interrupted her since I already read the question
- I do not know what to do now

The first round of testing revealed that some respondents needed help learning how to enter their answers. Thus, a training interview was constructed that contained questions that were not on the interview and were not sensitive. The interviewer went over these questions with the respondent who then completed the rest of the interview on her own. The field experiment included a comparison of ACASI, in-home CAPI, and out-of-home CAPI. It was hypothesized that the willingness of women to report sensitive informaton would be increased if they were interviewed outside of their homes because in prior rounds of the survey, respondents had indicated that one of their concerns was that family members would overhear their responses.[5]

Abortion Reporting. Prior rounds of the NSFG identified significant underreporting of abortion (Jones and Forrest, 1992). Exhibit B compares the results from the Audio-CASI question on whether or not the woman had ever had an abortion and both the pilot questions and pregnancy outcome questions in Section B. There was one refusal to the Audio-CASI question on whether the woman had ever had an abortion in her

---

4 Respondents refereed to the computer as "she" because the recordings were done by a woman.

5 An incentive experiment was also included. The out of home respondents were paid $40.00 and the in home respondents received either no incentive or a $20.00 incentive.

lifetime so that there are 177 rather than 178 respondents in this second set of Audio-CASI tables. We note that six additional women reported having had an abortion at some time in their life in the Audio-CASI interview. The six additional women who reported an abortion represents a 14 percent increase in the number of women reporting ever having had an abortion.

Exhibit B. Relationship of Abortion Reporting in the Pregnancy Outcome Section and to Abortion Reporting in the ACASI Interview National Survey of Family Growth -- Cycle V Pretest

	Abortion reported as a birth -- outcome		
ACASI: Ever had an abortion	Yes	No	
Yes	42	6	48
No	0	129	129
	42	135	177

Exhibit C shows detailed information on abortion reporting by site of interview, incentive, and type of interview. Two series of numbers are shown for the ACASI respondents—the number of abortions that they reported in answer to interviewer questions in Section B of the interview and the number reported in subsequent ACASI interview. Finally, in Exhibit D we show the results on the number of abortions reported in Section B and the Audio-CASI interview for those 178 respondents who completed the Audio-CASI interview. Note, in this table the following:

> Women who had reported an abortion in Section B reported additional abortions in the Audio-CASI interview.
>
> All of the differences in numbers of abortions reported are above the main diagonal indicating that the different numbers of abortions reported in the Audio-CASI are probably not due to random error.

Exhibit C: Distribution of the Number of Abortions Reported in Section By Treatment and Incentive

			In Home						A-CASI						Neutral	
			$0		$20		Total		$0		$20		Total		$40/Total	
# of abortions	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
0	380	77.1	83	86.5	56	77.8	139	82.7	78	79.6	58	72.5	136	76.4	105	71.4
1	79	16.0	10	10.4	12	16.7	22	13.1	16	16.3	11	13.8	27	15.2	30	20.4
2	24	4.9	1	1.0	4	5.6	5	3.0	3	3.1	9	11.3	12	6.7	7	4.8
3	9	1.8	2	2.1	0	0.0	2	1.2	1	1.0	1	1.3	2	1.1	5	3.4
4	1	0.2	0	0.0	0	0.0	0	0.0	0	0.0	1	1.3	1	0.6	0	0.0
Total	493	100	96	100	72	100	168	100	98	100	80	100	178	100	147	100
One or more abortions reported in B			13	13.5	16	22.2	29	17.3	20	20.4	22	27.5	42	23.6	42	28.6
One or more abortions reported in A-CASI									24	24.5	24	30.4*	49	27.1*		

* The proportions of A-CASI $20 incentives is based on 79 cases and total for A-CASI is based on 177 cases.

COPAFS2.DOC

Exhibit D: Number of Abortions Reported in Section B and in ACASI

	ACASI Section						
Section B	0	1	2	3	4	5	DK
0	129	4	1	0	0	0	1
1	0	24	2	1	0	0	0
2	0	0	11	1	0	0	0
3	0	0	0	1	0	1	0
4	0	0	0	0	1	0	0

We also fit a series of logistic regression models to determine if there were significant differences due to interviewing conditions. As independent variables, we included the type of interview (CAPI only, Audio-CASI, or neutral site), incentive for in-home interviews (none or $20), race/ethnicity (Hispanic, black, non-Hispanic-non-black), marital status (married, not married), income (unknown, greater than $20,000, or other), and age. We used a stepwise selection procedure in which an independent variable that was significant at the 0.15 level was added to the model. Exhibit E summarizes the results.

Exhibit E. Analysis of the Impact of Characteristics of Women and Interview Conditions on Abortion Reporting
National Survey of Family Growth -- Cycle V Pretest

	Parameter estimate	Standard error	Probability (Chi-square)	Odds Ratio
Intercept	-2.52	0.49	0.0001	1.081
Incentive - 20	0.38	0.27	0.1348	1.488
Married	-0.34	0.23	0.1428	0.714
Age	0.03	0.01	0.0264	1.033
Audio-CASI	0.54	0.27	0.0419	1.723
Neutral site	0.83	0.31	0.00672	2.294

Based on these results, we concluded that both the neutral site and the Audio-CASI increases the number of women who report that they ever had an abortions.[6]

Respondent Attitudes. We also asked respondents who received the Audio-CASI interview their attitudes toward the alternative methods of reporting abortion. Exhibit F presents the results.

6 We also examined the reporting of the number of abortions and found that given a woman had reported an abortion, there were not significant differences in the number of abortions reported.

Exhibit F. Respondents' Attitudes Toward Methods of Reporting Abortion Among Women who Received the Audio-CASI Interview National Survey of Family Growth -- Cycle V Pretest

Response	Percent respondents
How do you rate telling the interviewers your answers to questions on abortion?	
poor	15.2
fair	20.3
good	30.5
very good	17.5
excellent	16.4
How do you rate using the computer and earphones to answer questions on abortion?	
poor	2.8
fair	8.5
good	17.5
very good	26.0
excellent	45.2
Which method of answering questions on abortion is the most private?	
earphones and computer	62.7
no difference	32.2
telling the interviewer	4.5
Don't know	0.6
Which method do you recommend for the main study?	
Interviewer	16.9
Computer	58.2
Do not ask about abortion	2.8
Does not matter	22.0

In general, these women recommended the Audio-CASI procedure for abortion reporting.

Description and Demonstration of the Audio-CASI System. The ACASI system used has the following features:

- Implements a full range of audio functions so that audio self-interviewing can offer as many capabilities as interviewer-administered systems
- Runs on a powerful, existing CAI development platform
- Uses MS-DOS operating system

From an implementation and operational point of view, the key requirement is the second—that the audio system be built as an extension of an existing CAI development system. It is not difficult to build PC systems which can generate sound through digital audio devices, display questions, and record answers. However, to have a system which can conduct a complex questionnaire with integrated audio is much more difficult. One especially important requirement is the ability to allow the user to backup easily, correct a previous entry, and be directed forward following a route appropriate for the latest set of responses. If the underlying CAI platform is not robust and widely used in complex applications, then the stability of the ACASI application during interviews is likely to be problematic.

The system used the Blaise CAI system as its base. The Blaise system is a product of the state statistical agency of the Netherlands—Statistics Netherlands and is widely used across Europe by government statistical agencies for computer-assisted personal interviewing, telephone interviewing, data entry and date editing. Questionnaires are programmed in the Blaise CAI language by defining the questions, their answer choices, and the logic of the questioning, including tailored text fills and consistency checks. Blaise then compiles the questionnaire code into a executable DOS application, automatically handling the question administration, screen and keyboard control, range and consistency checks, data management and navigation through the questions.

The audio capability is implemented through a background DOS process which the Blaise instrument triggers as each question is displayed on the screen. This process interprets commands specifying the recorded digital audio files to play in order to duplicate in audio just what is displayed on the screen.

The hardware for the audio system is a small, one-pound external analog-digital box that is connected to the notebook PC by two cables and headphones. The audio quality in the system is quite high. The items are a digitally recorded human voice—not synthesized. The system is both very flexible and fast. It has the capability of rendering questions with variable components. For example, in the NSFG ACASI instrument, when a women said she had had an abortion, she was asked when it happened and how many weeks pregnant she was at the time. Then the following series of questions were asked:

> Based on this, this pregnancy began around [MONTH AND YEAR OF CONCEPTION].
>
> So we can understand how well birth control methods work, I would like for you to tell me what methods of birth control you were using—if you used any—during the three months before this pregnancy began.
>
> Were you using any method of birth control in [MONTH AND YEAR PRECEDING CONCEPTION]?
>
> As I read the methods, please press 1 for YES if you used that method in [MONTH AND YEAR PRECEDING CONCEPTION].

Then the woman is asked for each of the three months preceding conception, whether she used each of the birth control methods she had reported ever having used earlier in the CAPI section. In this question series, the ACASI system must be able to generate audio questions which can vary on the month and year and among 19 possible contraceptive methods. The system was able to instantly concatenate and play the appropriate audio files to duplicate the screen text properly.

DEMONSTRATION OF THE A-CASI SYSTEM WILL OCCUR HERE

COPAFS2.DOC

REFERENCES

Bradburn, Norman M. (1983) "Response Effects" in P. Rossi, J. Wright, and A. Anderson (eds.) Handbook of Survey Research, New York: Academic Press.

Catania, J.A., Gibson, D.R., Chitwood, D.D., and Coates, T.J. (1990). *Methodological Problems in AIDS Behavioral Research: Influences on Measurement Error and Participation Bias in Studies of Sexual Behavior*. Psychological bulletin (1990), vol. 108, No. 3, 339-362 by American Psychological Associates, Inc.

Erdman, H.P., Klein, M.H., Greist, J.H., Skare, S.S., Husted, J.J., Robins, L.N., Helzer, J.E., Goldring, E., Hamburger, M., & Miller, T.P. (1992). *A Comparison of two computer-administered versions of NIMH Diagnostic Interview Schedule*. Journal of Psychiatric Research, 26(1), 85-95.

Forsyth, Barbara, Lessler, J.T., and Hubbard, Michael, (1992). *Cognitive Evaluation of the Questionnaire*, Survey measurement of Drug Use: Methodological Studies, National Institution Drug Abuse.

Hay, David A. (1990) "Does the Method Matter on Sensitive Survey Topics," Survey Methodology, Vol. 16, No. 1.

Jones, Elise F. and Jacqueline Darroch Forrest (1992) "Underreporting of Abortion in Surveys of US Women: 1976 to 1988," Demography, Vol.29, No.1.

Lessler, Judith and Mimi Holt (1987) "Using Response Protocols to Identify Problems in the US. Census Long Form," Proceedings of the Survey Research Methods Section, American Statistical Association.

Levine, S.; Ancill, R.J.; Robert, A.P. (1989). *Assessment of Suicide Risk by Computer-Delivered Self-Rating Questionnaire: Preliminary Findings*. Acta Psychiatric Scandinavia (1989) Sep Vol 80(3) 216-220.

Locke, S.E., Kowaloff, H.B., Hoff, R.G., Safran, C., Popovsky, M.A., Cotton, D.J., Finkelstein, D.M., Page, P.L., & Slack, W.V. (1992). *Computer-Based Interview for Screening Blood-Donors for Risk of HIV Transmission*. Journal of the American Medical Association, 268 (10), 1301-1305.

Miller, H.G., Turner, C.F., and Moses, L.E., eds (1990). Chapter 6: Methodological Issues in AIDS Surveys. In AIDS: The Second Decade. Washington, DC: National Academy Press.

National Center for Education Statistics, 1993. Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey. Washington, DC: U.S. Department of Education.

O'Reilly, J.M, Hubbard, M., Lessler, J.T., Biemer, P. & Turner, C.F. "Audio and Video Computer-Assisted Self-Interviewing: Preliminary Tests of New Technologies for Data Collection." 1994. Journal of Official Statistics (In Press).

Schwarz, N., Strack, F., Hippler, H., and Bishop, G. (1991). *The Impact of Administration Mode on Response Effect in Survey Measurement*. Applied Cognitive Psychology, 5, 193-212.

Turner, C.F., Lessler, J.T., and Devore, J.W. (1992). *Effects of Mode of Administration and Wording on Reporting of Drug Use*. In Survey Measurement of Drug Use: Methodological Studies, eds. C.F. Turner, J.T. Lessler, and J.G. Gfroerer, DHHS Publication (ADM) 92-1929, Washington, DC: Government Printing Office.

Waterton, J.J., and Duffy, J.C. (1984) "A Comparison of Computer Interviewing Techniques and Traditional Methods in the Collection of Self-Report Alcohol Consumption Data in a Field Survey." International Statistical Review 52:173-182.

Weinrott, Mark R.; Saylor, Maureen (1991). *Self-Report of Crimes Committed by Sex Offenders.* Journal of Interpersonal Violence (1991) Sep Vol 6(3) 286-300.

THE LANGUAGE OF SELF-ADMINISTERED QUESTIONNAIRES AS SEEN THROUGH THE EYES OF RESPONDENTS[1]

by

Cleo R. Jenkins
U.S. Bureau of the Census
and
Don A. Dillman
U.S. Bureau of the Census and Washington State University

[1]This paper draws heavily from a previous paper of ours (see Jenkins and Dillman, 1993). Cleo R. Jenkins is a Survey Statistician, Center for Survey Methods Research, U.S. Bureau of the Census, and Don A. Dillman is Senior Survey Methodologist, Office of the Director, U.S. Bureau of the Census, and Professor and Director of the Social and Economic Sciences Research Center at Washington State University. The views expressed in this paper are those of the authors and are not necessarily those of the U.S. Bureau of the Census or Washington State University. Questions concerning this paper should be addressed to Cleo Jenkins, U.S. Bureau of the Census, 433 Washington Plaza, Building 1, Washington DC, 20233-4700 (301-763-7331).

Acknowledgments: The authors would like to thank Elizabeth Martin and Theresa J. DeMaio of the Center for Survey Methods Research, U.S. Bureau of the Census, for reviewing this paper and providing insightful comments.

The Language of Self-Administered Questionnaires As Seen Through the Eyes of Respondents

by
Cleo R. Jenkins and Don A. Dillman

"... we must recall that language includes much more than oral and written speech. Gestures, pictures, monuments, visual images, finger movements--anything consciously employed as a sign is, logically, language."

John Dewey in Newell, A and Simon, H.A.
Human Problem Solving, *1972, pp. 65-66*

I. INTRODUCTION

Much survey research has been directed at studying interviewer-administered questionnaires. As a result, we have learned a great deal about question wording and sequencing effects in surveys and about the effect of memory on data quality (e.g., Jobe et al, 1993; Jobe et al., 1990; Lessler, 1989; Converse and Presser, 1986; Belson, W., 1981). Although it is equally important to understand these sources of error in self-administered questionnaires, it is not sufficient. The graphical presentation of information is every bit as important because it too has something to say to the respondent.

In Tourangeau's model (1984), as well as other models of the survey interview process, the first step is specified as "comprehending the question." Depending on the model, different steps follow, but generally, "retrieval of the relevant facts, judgment, and finally, response" are mentioned.

Although "comprehending the question" is the first step in an interviewer-administered survey, the task is different in a self-administered survey. In a self-administered survey, respondents must first "perceive the information" before they can comprehend it. Once they perceive it, they must "comprehend the layout of the information" as well as "the wording of the information." Furthermore, respondents must comprehend much more than just the wording of the survey questions and response categories. In a self-administered survey, respondents are often given introductory material and instructions. Also, they must comprehend directions that are meant to guide them through the form.

In an interviewer-administered questionnaire, the interviewer plays a critical role in what information the respondent perceives. In a self-administered format, the entire onus of perception is on the respondent, and we have not developed procedures for controlling errors that might arise as a result of their not perceiving information as we intend. In fact, we have not studied this much at all.

In addition, we need to pay attention to what motivates respondents to answer surveys. Cialdini (1988) has argued that people decide whether to perform a requested task on the basis of the inherent attractiveness of that task <u>and</u> other social or psychological influences, including

- reciprocation (the tendency to favor requests from those who have previously given something to you),
- commitment and consistency (the tendency to behave in a similar

way in situations that resemble one another),

- social proof (the tendency to behave in ways similar to those like us),
- liking (the tendency to comply with attractive requests),
- authority (the tendency to comply with requests given by those in positions of power), and
- scarcity (the tendency for rare opportunities to be more highly valued).

Groves and others (1992) provide examples of how each of these can be utilized to encourage survey participation. Although most of the examples refer to interviewer behavior or the implementation process, some can be applied to questionnaire design. For example, the fact that people tend to comply with attractive requests suggests that respondents will be more likely to answer an attractive questionnaire than an unattractive one.

Groves and others also argue that the helping tendencies of people can be utilized to encourage response. Three emotional states have been found to be associated with decisions to help another: anger, happiness, and sadness. One would expect that people are likely to become angry and therefore less likely to respond to a mail survey when the questions or the instructions are not easily understood.

Finally, the literature on opinion change (Petty and Capioppo, 1986) suggests that when a topic is of high personal relevance, subjects will change their opinion based on an in-depth review of a message. However, when the topic is not important to the subject, they will rely on a heuristic review, such as the credibility of the source. This literature suggests that if a questionnaire is not really important to a respondent, then we probably aren't going to persuade them to complete it by presenting them with an in-depth, highly logical, persuasive discussion of why they should complete it. Instead, we should rely on other means.

II. GRAPHICAL DESIGN PRINCIPLES

In the remainder of this paper, we present questionnaire design principles that struggle with how to best present information to (1) motivate respondents to attempt the tasks presented to them and to (2) aid respondents to accurately answer the questionnaire once they are motivated to do so. Most of the principles have not been tested carefully on controlled designs, although we try to cite those occasions in which they have been. Principle 20 (structuring and organizing a questionnaire) is an example of a principle drawn from experimental evidence with the Decennial Census questionnaires. Most of the other principles are drawn from the results of cognitive interviews with both the Schools and Staffing Survey and the Census of Construction Industries and from the redesign of the Survey of College Graduates. The response effects of the redesigned examples remain generally untested at this time. Therefore, the principles should be viewed as reasonable hypotheses for improving response, lowering item non-response, and improving accuracy. A major reason for writing this paper is to encourage experimental research on these issues.

1. **Present information in a format that respondents are accustomed to reading.**

We consider this the most important principle and one that is constantly and inadvertently violated. Jenkins et al. (1992a) present the results of cognitive interviews with the Schools and Staffing Survey. Example 1 shows the cover page of the School Questionnaire from the Schools and Staffing Survey. Jenkins et al. conclude that the readers, persistent as they were, usually did a pretty good job of following this page until they reached the end of the first column. These respondents read through the title information, then the first two paragraphs on the left-hand side of the page. Because these paragraphs refer to the label, they turned the questionnaire sideways to look at the label. When done, they returned to where they had left off on the left-hand column, and continued to read down the column. Instead of continuing to the top of the second column, however, generally they turned the page.

Example 2 presents a diagrammatical representation of the cover page's reading structure. It reveals the eye's necessary movement across the page. As can be seen, the current format requires respondents to make some pretty large unexpected leaps across the page, unexpected in the sense that a person anticipates reading a line of information from left to right, starting at the top of the page and moving down it.

It is not surprising, therefore, that the skimmers didn't bother to read this page at all. Generally, they glanced at it and turned the page. Because the skimmers never read the school named on the label, they often reported for the wrong school. In fact, this error was so great in 1991 that data for 10 states needed to be suppressed at first.

We present a redesigned version of the cover page using a natural reading format in conjunction with the next principle.

2. **Present only the most relevant information using graphical design features and composition.**

Another problem with the School Questionnaire's cover page is that it presents too much information. Skimmers quickly dismissed this information, probably because nothing was made particularly salient to them and they were not willing to look for that which was important. This suggests that the most important information needs to be made easily perceptible.

Example 3 shows a "user-friendly" cover page adapted from Jenkins and Ciochetto (1993). This cover page presents only the information the respondent needs to begin completing the questionnaire and it does so using a natural reading format and graphical design features. Jenkins and Ciochetto deliberately used a box that contains an unshaded area within a shaded one to showcase the very important instruction that was overlooked on the original questionnaire.

Example 4 exhibits the straight forward reading structure of this page. No need for the eye to do anything out of the ordinary, which we are more and more convinced is critical to designing good self-administered questionnaires.

3. **Pique respondents' interests early in the questionnaire.**

A third problem with the School Questionnaire was that respondents found themselves being asked to passively read a lot of material: the cover page, the cover letter (which was placed on the inside cover of the questionnaire booklet), and instructions (see Example 5). Skimmers skipped over this material entirely and went directly to question (a) in the middle of page 4 to begin completing the questionnaire.

Based on this, the third principle is: pique respondents' interest early in the questionnaire. Don't begin the questionnaire with a lot of prose, begin by asking a question or two. We hypothesize that respondents are more likely to read information once they have become actively engaged in answering the questionnaire.

As shown in Example 6, Jenkins and Ciochetto (1993) suggest beginning the Student Records Questionnaire by asking respondents to record the current time followed by a screener question. It is only after they ask these questions that they present a condensed version of the cover page and letter information. Even when they present this information, they deliberately used a question-and-answer format to keep the respondent actively engaged.

4. **Dominantly feature questions over additional explanatory information.**

This principle is violated in the Census of Construction Industries Questionnaire. Example 7 presents the "dollar value of business" item from that questionnaire. As can be seen, this item, like all of the items on the questionnaire, begins with a brief capitalized heading in bold that is meant to quickly convey the nature of the item. DeMaio and Jenkins (1991) conclude that respondents often neglected to read beyond the heading because the heading provided them with just enough information to formulate their own question. And, of course, they formulated the wrong question.

Example 8 presents a revised version of the item. In this version, the item heading is replaced with a bold-faced, comprehensive question. Bold-faced type was used to convey the importance of the question. Also, it serves as a road map for questions like this that have a leading phrase followed by several parts that are interrupted with other information. The other information is put in light-faced type.

5. **Include in each question all of the relevant information necessary for respondents to answer it, rather than specifying information in a subsequent instruction.**

This principle is demonstrated using another item from the Census of Construction Industries Questionnaire, the "number of employees" item (see Example 9). Respondents would read the heading here, and sometimes the question, then they would turn their attention to the answer boxes at the right (DeMaio and Jenkins, 1991). At this point, their eyes were drawn immediately to the column headings rather than the header. The header reads "Number of employees of this establishment during the pay period including the 12th of--," and because "pay period" is not mentioned in either the heading or the question, several respondents mistakenly thought this item was referring to monthly or quarterly time periods.

A revised version of this item is presented in Example 10. Besides removing the heading, the pay period header was made the leading phrase in this version. This question now contains all of the information the respondent needs to answer it. And just to make sure respondents don't misunderstand, the pay periods are also repeated in each of the column headings.

6. Vertically align the questions and response categories.

As can be seen in the top view of Example 11, the Public School Questionnaire uses a question-on-the-left-answer-on-the-right format. Jenkins et al. (1992a) conclude that respondents often did not read instructions in this format. This is because they generally began to search for the answer once they read the question. As a result, their thoughts and consequently, their eyes were drawn away from the left-hand side of the page, where the instructions lay, to the right-hand side, where they knew the answer categories were.

The second view in Example 11 uses a vertical alignment. This places the instructions directly before the answer category, where respondents are more likely to perceive them. However, this may not solve the problem of respondents either overlooking or ignoring instructions. As already mentioned, respondents have a tendency to read only as much as they think is necessary to answer a question. Therefore, even if they perceive the instructions, they may still ignore them. If the instruction is relatively simple to begin with, a better solution is to incorporate it into the body of the question, as demonstrated in the last view.

7. If incorporating needed information into the question makes it too complicated to understand, then provide accompanying instructions at the place where they are needed.

If an instruction is long and/or complicated, incorporating it into the body of the question is likely to fail. As can be seen in Example 12, Item 2 of the Public School Questionnaire asks how many students were enrolled in the school on or about October 1 of this school year. Jenkins et al. (1992a) conclude that this item was difficult for respondents to read and understand because the flow of the question is interrupted by two parenthetical phrases and a lengthy two-sentence instruction. This leads us to conclude that one should never try to insert a stand-alone instruction between phrases of a continuous question.

Question 13a(2) in Example 8 illustrates the use of include and exclude statements that are too lengthy to incorporate into the body of the question. Here, the instructions are placed directly after the question. This places them as close to the question as possible without disrupting its flow. Still, further research is needed to determine the best method for coaxing respondents to read information that is not easily incorporated into the question.

8. Utilize single-task formats rather than multi-task formats.

Item 30 from the Public School Questionnaire (shown in Example 13) asks respondents to cross classify their employees by full- or part-time status and assignment. Jenkins et al. (1992a) conclude that quite a few respondents seemed able to process only one aspect of this item--the categorization of employees by job description--and were unable to simultaneously deal with the additional request to report these employees by both full- and part-time status.

In Example 14, the multi-tasked format is replaced with a simpler one. This is accomplished by focusing on only one request at a time--first, respondents are asked to classify part-time employees by job description in part a. Then they are asked to classify full-time employees by job description. Because this format repeats information at the point it is needed, we hypothesize that it will make the respondent's task easier. Of course, the disadvantage is that it lengthens the questionnaire, which may decrease a respondent's motivation to complete it. This example illustrates the fact that there can be competing forces at work when we design a questionnaire and that we clearly need to learn more about these forces.

9. Utilize single-question formats rather than matrix-question formats.

Question 1 in Example 15 asks respondents if they currently have the students for class listed down the left-hand side of the page. If the respondents do, then they are to answer three follow-up questions. Jenkins and Ciochetto (1993) conclude that this format presents respondents with too many tasks at once. Furthermore, it presents them with a choice, but provides little guidance for making the choice. They can choose to answer a full set of questions about one student at a time. In this case, they work across the rows. Or they can answer the same question for each of the students. In this case, they work down the columns.

In Example 16, the matrix format is replaced with a single array of questions pertaining to one student at a time, with the questions running down the page rather than across it. In this version, respondents need only be concerned with answering one question about one student at a time, and they need not deviate from moving down the page in search of the next question. Although the researchers recommend further work in this area, a small number of cognitive interviews showed that this is a more manageable task from the respondent's point of view. (This example also used a new skip instruction. We discuss skip instructions later.)

Additional research supporting the single-question format comes from both focus group and experimental research on the 1990 Decennial Census Questionnaire. A focus group examination of the Census Questionnaire in which respondents were asked to answer a series of questions for each member of their household in a matrix format (questions in left-hand column to be answered for household members listed across the top of the page) identified the matrix format as a barrier to response (Dillman et al., 1991). Furthermore, a revised questionnaire, which used a single-question rather than a matrix-question format, attained an improved response rate (Dillman et al., 1992).

10. Make headings and instructions at the top of a page more prominent than those in the middle of a page.

Respondents find transitions between topics helpful. A transition need not be complicated, it simply needs to be enough to warn the respondent the topic is about to change. For instance, in Example 17, the heading "SECTION

2--STAFFING PATTERNS" in the middle of the page was enough to convey to respondents that the topic was about to change (Jenkins et al., 1992a).

In contrast, Example 18 shows a transitional heading followed by an instruction that comes at the top of a page. Contrary to expectations, respondents tended to read transitional headings and instructions that came in the middle of a page, but few read information that fell at the top of a page.

We hypothesize that respondents may be exhibiting a similar kind of behavior here as they exhibited with the introductory information. Many respondents skipped over the introductory material in an effort to get to the questions, but once they were actively engaged in answering the questionnaire, they were more likely to read information put into their path. Perhaps respondents skip over information at the top of a page in an effort to get to the next question, but once they become involved in answering the questions, they are more likely to see other information.

11. Provide directions in a natural reading format and utilize graphical design features and composition to make the directions more salient.

In order to efficiently and accurately answer a self-administered questionnaire, respondents must be able to maneuver their way through the questionnaire. One very important instruction for doing this is the skip instruction. The problem with skip instructions, however, is that respondents commonly overlook them (Jenkins and Ciochetto, 1993; Turner et al., 1992; Gower, 1989).

Jenkins and Ciochetto (1993) conclude that respondents overlook skip instructions for two reasons, one of which is derived from the other. The primary reason respondents overlook the skip instruction is because they do not perceive it, but the reason they do not perceive it is because of the convoluted reading structure presented by the skip instruction. Item 29a shown in Example 19 illustrates this. A respondent begins to answer this item by first reading the question "Were there any teaching vacancies in this school for this school year, i.e., teaching positions for which teachers were recruited and interviewed?" Then they will move to the right-hand side of the page to answer the question, see the answer boxes, and continue to the right of these to read the answer choices "yes" and "no." The next step in the process is to choose one of these, say the "no" response, and to move back to the left of this to mark the answer box. Note what is happening at this moment--the respondents are moving away from the skip instruction. If the skip instruction has not been in some way made salient to respondents before they begin their journey back to the left, chances are they are never going to see it. Once they mark the answer box they are likely to conclude they are done answering this question and are going to begin to look for the next question.

Experimental data presented by Turner et al. (1992) confirm the hypothesis that respondents only see information to the right of an answer category if it is in some way made salient. Among other questionnaire design issues, Turner et al. studied the extent to which respondents and interviewers correctly executed branching instructions embedded in alternative versions of the 1990 National Household Survey on Drug Abuse (NHSDA) questionnaire. Turner et al. conclude that respondents were more likely to overlook a visually obscured branching instruction, as shown in question 1 of Example 20, than a visually salient one, as shown in question 5. They conclude that both the length and visual salience of questions a through e in question 5 assisted respondents in correctly following the branching instruction in question 5.

These studies suggest that directions need to be presented in a more natural reading format and graphical design features and composition should be used to make the directions more salient. Given this information, three alternative skip instruction formats are presented below, each of which appears to have advantages and disadvantages: (1) the salient skip instruction, (2) the intermediate skip instruction, and (3) the natural reading sequence skip instruction.

Salient Skip Instruction. Example 21 shows the salient skip instruction. Rather than having information placed to the right of the answer categories, directional arrows are placed to the left of both answer boxes. These arrows extend horizontally from each answer box towards the left-hand margin of the page and then turn vertically downwards. One of these arrows proceeds to the next question and the other ends with a verbal instruction within a shaded box.

This format was designed to overcome the conventional skip instruction's highly convoluted reading format (moving from left to right, right to left, back again to the right and finally, back to the left) and to replace a more-difficult-to-perceive verbal instruction with a more-easily-perceived combination visual/verbal instruction. In our judgement, the advantage of this format is that respondents may visually take in, if only briefly, the skip instruction information while moving from left to right in search of the answer categories.

The disadvantage, however, is that whereas respondents may be more likely to see this information, they also may be more likely to misunderstand it. It is possible that the wrong respondents (those who are supposed to continue to the next question) may mistakenly execute the skip instruction because of its visual salience, leading to a serious error--the omission of data. Another disadvantage with this format is that a question with a complicated skip instruction may become visually cluttered.

Intermediate Skip Instruction. Example 22 presents the intermediate skip instruction. This format relies on two features: (1) graphical instructions (an arrow) for going to the next question and (2) words to direct other respondents through a skip pattern. The two paths are further distinguished by originating the arrow from the left of the answer choice, and placing the words to the right.

In our judgment, the advantage that this format may have over the salient skip instruction is that respondents are unlikely to make the serious error of incorrectly executing the skip instruction. Another advantage is that it may not appear as visually cluttered to respondents. However, a small number of cognitive interviews suggest that it may not be as efficiently executed as the salient skip instruction. Just as with the conventional skip instruction, the word instruction to the right of the answer category may be overlooked at first. However, it is likely to be more efficiently executed than the conventional skip instruction, in which nothing but words are used off to the right of the answer choices. Therefore, this skip instruction format is a deliberate compromise between the conventional and salient skip instruction.

Natural Reading Sequence Skip Instruction. In contrast to the above skip instructions in which a respondent must move from left to right in search of the answer categories and then reverse this direction and move from right to left to answer the question, another possibility might be to establish a skip instruction format with a more efficient, natural, and logical flow. This format is shown in Example 23. As can be seen in this example, the answer boxes are placed to the right of the answer categories and the skip instructions to the right of the answer boxes. To maintain the vertical alignment of the answer boxes, the answer categories are right-justified rather than left-justified. Also, if the answer categories need to be double or triple-lined, as is the case with the category "Dropout/Chronic Truant (See Definition Below)" in question 1 of Example 23, then the answer box should follow the last of these lines. This is to help maintain the respondent's natural reading structure, for which we have been arguing all along.

This skip instruction seems to have several advantages over the preceding skip instructions. The first and probably best advantage is that the information is presented to respondents in the sequence they need it: first the answer categories, then the answer boxes, and finally, the skip instruction. Example 24 compares the reading format of the natural reading sequence skip instruction with that of the conventional skip instruction. As can be seen, respondents need not ever reverse their direction with the natural reading sequence skip format. Another advantage is that the natural reading sequence format is not cluttered looking.

A disadvantage, however, is that respondents may overlook bracketed skip instructions using this format. Although these instructions will be closer to the answer boxes in this format than they are in the conventional skip instruction format (that is, if the answer categories come between the answer boxes and the skip instruction), they be just far enough away from the answer boxes as to be out of the respondent's view.

Another disadvantage is that from an overall perspective, the questionnaire's vertical alignment is disrupted. In the previous formats, the questions, answer boxes, and categories are all left justified and begin in the same horizontal position on the page. Although vertical alignment of the questions can be maintained using the natural reading sequence skip instruction format, the answer categories will certainly not be vertically aligned. The answer boxes can be made to maintain vertical alignment within a question; however, they may not be able to maintain alignment from question to question, further disrupting the overall look of the questionnaire.

A final disadvantage with the natural reading sequence skip instruction is related to data processing. In this format, the location of the keycodes is problematic. One possibility is to place the keycodes before the answer category, but this puts them quite a distance from the answer box from the keyer's perspective. This may slow down production and/or increase keyer error. Another possibility is to place them either directly before or after the answer box, but this may confuse the respondent.

We have described skip instructions at some length because it is an area which is exceedingly important, but now lacks ideal solutions. The alternatives presented need extensive testing in large samples.

12. Utilize graphical design techniques to establish a clear path through the questionnaire for the respondent to follow.

Many questionnaires mix questions and information, utilizing space wherever it is available and thinking that so long as the information is presented, it will get read. As can be seen in Example 25, it is unclear to the respondent where to begin, and most important in what order the information is to be read.

Example 22 is a redesigned page from the Survey of College Graduates. Here the white answer spaces contrast with the light blue background. The message intended, and communicated by graphical layout rather than words, is to establish a visual path through the questionnaire by associating the white spaces with the "need to provide an answer."

13. Avoid using the same design feature to request different respondent actions.

The essence of this principle is to associate particular design features with what the respondent is being asked to do, and to be completely consistent with their use. For example:

13a. Use dark type for question stems and light type for response category options.
13b. Write all definitions and special instructions for a particular question in italics placed within parentheses.
13c. Use capital letters for words to be emphasized to the respondent in both questions and answers.

The important point here is not that capitals must be reserved for emphasis, and italics for instructions, or that bold type is better for questions than light type. Doing the opposite may work just as well--the issue is consistency, so that as a respondent gets into a questionnaire they begin to associate the chosen procedure with a particular piece of information or request for action.

14. Utilize variability in design features judiciously.

Closely associated with the need to be consistent is the need to limit variability. One would never consider writing a paragraph in which every word is written in different type fonts and sizes. Doing so would slow down the reader's comprehension. Instead, one should select a limited number of design elements and use them consistently.

15. Visually emphasize information the respondent needs to see and de-emphasize information the respondent does not need to see.

Coding information is a good example of this principle. In Example 26, the codes are bold and made even more prominent by encasing them in boxes. Not only that, but they are placed directly in the respondent's reading path. One result is that respondents may mentally process information irrelevant to them, thus making the task of responding more time consuming and difficult than necessary.

In Example 22 the light blue background is a 10 percent screen, and the coding information is printed in small numbers without boxes in 100 percent color. The respondent, who is already being guided "towards" the white answer spaces by black type of questions and answers and "away from" the blue background seems less likely to see or be confused by the dark blue lettering. Furthermore, the codes are placed outside the respondent's reading path. Yet, for a person who is searching for the blue code numbers, they are easily visible.

16. Utilize graphical layout of questions on the page to distinguish among different types of question structures; maintain consistency within types.

If a questionnaire begins by listing answer categories below the stem of a question vertically, like the "yes/no" answer categories in the first question in Example 27, it is undesirable to occasionally present answer choices horizontally, or even to sometimes use a second or third column of answer choices. Once a format is selected it needs to be followed consistently.

If one has a question like C9 and C10, where several items in a series are to be evaluated and the answer categories are the same for each item in the series, these answer categories should be placed to the right of the items and the respondent should be instructed to choose from among horizontally arranged categories in this case. Respondents should learn to choose from among vertical choices when the boxes are on the left and from among horizontal choices when boxes are to the right.

17. Provide descriptive captions either above, beneath, or to the right of blank answer spaces and utilize appropriate signs or symbols whenever numbers are requested.

When people are asked to report income, number of weeks worked, or other data by filling in blank spaces, inaccuracies may result from utilizing the wrong units or from not remembering exactly what was asked. Therefore, the answer spaces in Example 27 have captions to remind people what is being requested. For instance, C13 has the caption "Total 1991 earned income." In addition, the blank answer space has a dollar sign and ".00" in it to keep people from reporting cents, since they weren't wanted.

18. Utilize dominant graphical markings to provide the most important information needed by the respondent to guide them through the answering process.

This principle is violated in Example 25, where the "return to" instruction is predominant. It is also violated in Example 28 where the black marks used to optically scan the questionnaire are quite dominant. In neither case do the dominant marks effectively guide the answer process.

In Example 29, the dominant markings are the questionnaire's title, **THE 1992 NATIONAL CENSUS TEST**, followed by the **ARROW**, and the **PERSON 1** and **PERSON 2** headings. These dominant markings are meant to guide the respondent through the form.

19. Avoid the separation of questions through the use of lines and rectangles in favor of an open format in which the respondent's answering path is clearly shown.

Frequently designers of questionnaires utilize lines and rectangles to separate questions from one another. In general this practice makes questionnaires more, rather than less, difficult to answer. The use of rectangles, as shown in Examples 25 and 30, gives no clear indication of where to go next; the lines function in much the same way as "stop" signs, requiring one to stop and contemplate the next steps. This is especially the case in these two examples, where it is not readily apparent which box comes next. Consequently, the boxes require additional information, that is, the prominent section numbers. Also, the use of lines is one additional use of ink on a page which must then be cognitively processed by the reader, in contrast to white (or other background color) space which one can pass over without pausing to think about what it means.

In contrast, the formats used in Examples 22, 27, and 29 are open, using lines mostly to identify the page space in which answers are to be provided. These pages are easier for respondents to follow. Also, the respondent path is easily recognized, following the cultural norm of left to right within the defined answering space, and top to bottom on the page.

20. Structure and organize the questionnaire in such a way that it, first, makes sense to respondents and, second, avoids leaving the choice of the order in which questions get answered up to the respondent.

On the surface, this principal seems obvious and easy to implement. However, this may not be so, if the Census long and short forms are any indication. The Census long form is probably one of the most complex questionnaires in existence. It has a fold-out flap which asks for a listing of household members, followed by a matrix of short-form information and finally two pages of sample population questions for each person. This form involves a complex sequence of tasks, the order of which was traditionally dictated by Census needs to provide Congress with mandated information by the end of the census year.

A split-panel experiment with the long form, known as the 1990 Alternative Questionnaire Experiment (AQE), showed that the form's structure was not properly organized from the respondent's point of view (DeMaio et al., 1992). Along with the control form, which was identical to the 1990 Census long form, five experimental questionnaires were tested in the 1990 AQE. Two of the experimental questionnaires (Panels 2 and 3) incorporated many of the principles we have discussed concerning color, the consistent use of typeface and answer spaces, etc. However, three of the experimental questionnaires (Panels 4, 5, and 6) incorporated dramatic changes to the structure and organization of the form. Panel 4 became a matrix booklet in which the flap was eliminated and all of the person items were placed together. Panels 5 and 6 became "kits" in which individual questionnaires for each person in the household were placed in a folder.

The main finding was that "small" format changes alone (as incorporated in Panels 2 and 3) did very little to improve either item or overall response rates, but it took changes to the structure and organization of the questionnaire (as incorporated in Panels 4, 5, and 6) to make improvements. This suggests that "small" format changes are not enough to overcome the difficulty of completing a questionnaire that is not properly organized from the respondent's point of view. In addition, the Simplified Questionnaire Test (Dillman et al, 1992) and the Appeals and Long Form Experiment (Bates, 1993) confirmed this finding.

III. CONCLUSION

Little information on the design of self-administered questionnaires existed until relatively recently. That which did was based primarily on common sense and individual experience. Instead, it was the verbal language of interviewer-administered questionnaires that predominately captured the attention of researchers.

The evidence presented in this paper demonstrates, however, that we need to pay serious attention to the visual language of self-administered questionnaires in addition to the verbal. Toward this end, we need to develop a set of scientifically derived and experimentally proven graphic design principles to guide us in our quest to improve both response rates and the accuracy of responses. We hope that the principles we've developed are a first step in that direction. We have little doubt that the problems we've uncovered exist. However, because many of the solutions have not been tested, we openly admit that they are subject to challenge. Some of the solutions we've offered will stand the test of time; others will not. Undoubtedly, this is an area in need of further study and creative insight.

Finally, we also hope to expand upon our work here by exploring literature that has remained outside the domain of survey methodology to date-- most notably, the eye-movement and the graphical design literature. Knowing what we do now, it certainly seems that this literature may offer further insight into the self-administered question-response process.

REFERENCES

Bates, N. and DeMaio, T. 1989. "Using Cognitive Research Methods to Improve the Design of the Decennial Census Form." *Proceedings of the Bureau of the Census Fifth Annual Research Conference*, March 1989, pp. 267-290.

Bates, N. 1993. Personal Communication. Preliminary Results from Appeals and Long Form Experiment. U.S. Bureau of the Census. Washington, DC.

Belson, W. 1981. *The Design and Understanding of Survey Questions*. Gower.

Cialdini, R.B. 1984. *Influence: The New Psychology of Modern Persuasion*. New York: Quill.

Converse, J.M. and Presser, S. 1986. *SURVEY QUESTIONS Handcrafting the Standardized Questionnaire*. Newbury Park: Sage Publications, Inc.

DeMaio, T. and Jenkins, C. 1991. "Questionnaire Research in the Census of Construction Industries." *Proceedings of the Section on Survey Methods Research*, American Statistical Association, August 1991.

DeMaio, T.J. and Bates, N.A. 1992. "Redesigning the Census Long Form: Results from the 1990 Alternative Questionnaire Experiment" *Proceedings of the Section on Survey Methods Research*, American Statistical Association, August 1992.

Dillman, D.A., Reynolds, R.W., and Rockwood, T.H. 1991. *Focus Group Tests of Two Simplified Decennial Census Forms*. Technical Report 91-39. The Social and Economic Sciences Research Center. Washington State University. Pullman.

Dillman, D.A., Sinclair, M.D., and Clark, J.R. 1992. "Mail-back Response Rates for Simplified Decennial Census Questionnaire Designs." *Proceedings of the Section on Survey Methods Research*, American Statistical Association, August 1992.

Groves, R.M., Cialdini, R.B., and Couper, M.P. 1992. "Understanding the Decision to Participate in a Survey." *Public Opinion Quarterly*, 56: 475-495.

Jenkins, C. 1992. "Questionnaire Research in the Schools and Staffing Survey: A Cognitive Approach." *Proceedings of the Section on Survey Methods Research*, American Statistical Association, August 1992.

Jenkins, C., Ciochetto, S., and Davis, W. 1992a. "Results of Cognitive Research on the Public School 1991-92 Field Test Questionnaire for the Schools and Staffing Survey." Internal Census Bureau Memorandum, June 15, 1992.

Jenkins, C., Ciochetto, S., and Davis, W. 1992b. "Recommendations for Changes to the Public School 1991-92 Field Test Questionnaire for the Schools and Staffing Survey (SASS-3A) Based on Cognitive Interviews." Internal Census Bureau Memorandum, September 24, 1992.

Jenkins, C. and Ciochetto, S. 1993. "Results of Cognitive Research on the Multiplicity Question from the 1991 Schools and Staffing Survey Student Records Questionnaire." A report submitted to the National Center for Education Statistics, Census Bureau Memorandum, February 10, 1993.

Jenkins, C.R. and Dillman, D.A. 1993. "Combining Cognitive and Motivational Research Perspectives for the Design of Respondent-Friendly Self-Administered Questionnaires. " Revision of a Paper Prepared for Presentation at the Annual Meetings of the American Association for Public Opinion Research. May, 1993.m

Jobe, J.B., White, A.A., Kelley, C.L., Mingay, D.J., Sanchez, M.J. and Loftus, E.F. 1990. "Recall Strategies and Memory for Health Care Visits." *Milbank Memorial Fund Quarterly/Health and Society*, 68: 171-189.

Jobe, J.B., Tourangeau, R., and Smith, A.F. 1993. "Contributions of survey research to the understanding of memory. *Applied Cognitive Psychology*, 7: 567-584.

Lessler, J.T. 1989. "Reduction of Memory Errors in Survey Research: A Research Agenda." *Proceedings of the 47th Session of the International Statistical Institute*, Paris, France, pp. 303-322.

Newell, A. and Simon, H.A. 1972. *Human Problem Solving*. New Jersey: Prentice-Hall, Inc., pp. 65-66.

Petty, R.E. and Cacioppo, J.T.. 1986. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*. New York: Springer-Verlag.

Tourangeau, R. 1984. "Cognitive Science and Survey Methods" in T.B. Jabine, M.L. Straf, J.M. Tanur, and R. Tourangeau (eds), *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, Washington, DC: National Academy Press, pp. 73-100.

Turner, C.F., Lessler, J.T., George, B.J., Hubbard, M.L. and Witt, M.B. 1992. "Effects of Mode of Administration and Wording on Data Quality" in C.F. Turner, J.T. Lessler, and J.C. Gfroerer, (eds), *Survey Measurement of Drug Use Methodological Studies*, Washington, DC: National Institute of Drug Abuse, U.S. Department of Health and Human Services, pp. 221-243.

FORM SASS-3A

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS
ACTING AS COLLECTING AGENT FOR
U.S. DEPARTMENT OF EDUCATION
NATIONAL CENTER FOR EDUCATION STATISTICS

OMB No. 1850-0598
Approval Expires 12/31/92

PUBLIC SCHOOL QUESTIONNAIRE

SCHOOLS AND STAFFING SURVEY

1991-92 Field Test

This report is authorized by law (20 U.S.C. 1221e). Your answers will be kept strictly **confidential.** Results from this survey will appear in summary or statistical form only, so that individuals cannot be identified.

RETURN TO **Bureau of the Census, Current Projects Branch, 1201 East 10th Street, Jeffersonville, IN 47132-0001**

Please complete this questionnaire with information about the SCHOOL named on the label, and return it to the Bureau of the Census in the enclosed preaddressed envelope. Please return it within 3 weeks.

If the school's name or grade level is different from that indicated on the label or if you have any questions, please call the Bureau of the Census at 1-800-221-1204.

If the school named on the label is no longer in operation, mark (X) the box below and return this questionnaire to the Bureau of the Census in the enclosed envelope.

0011 1 ☐ School no longer in operation

(Please correct any error in name, address, and ZIP Code)

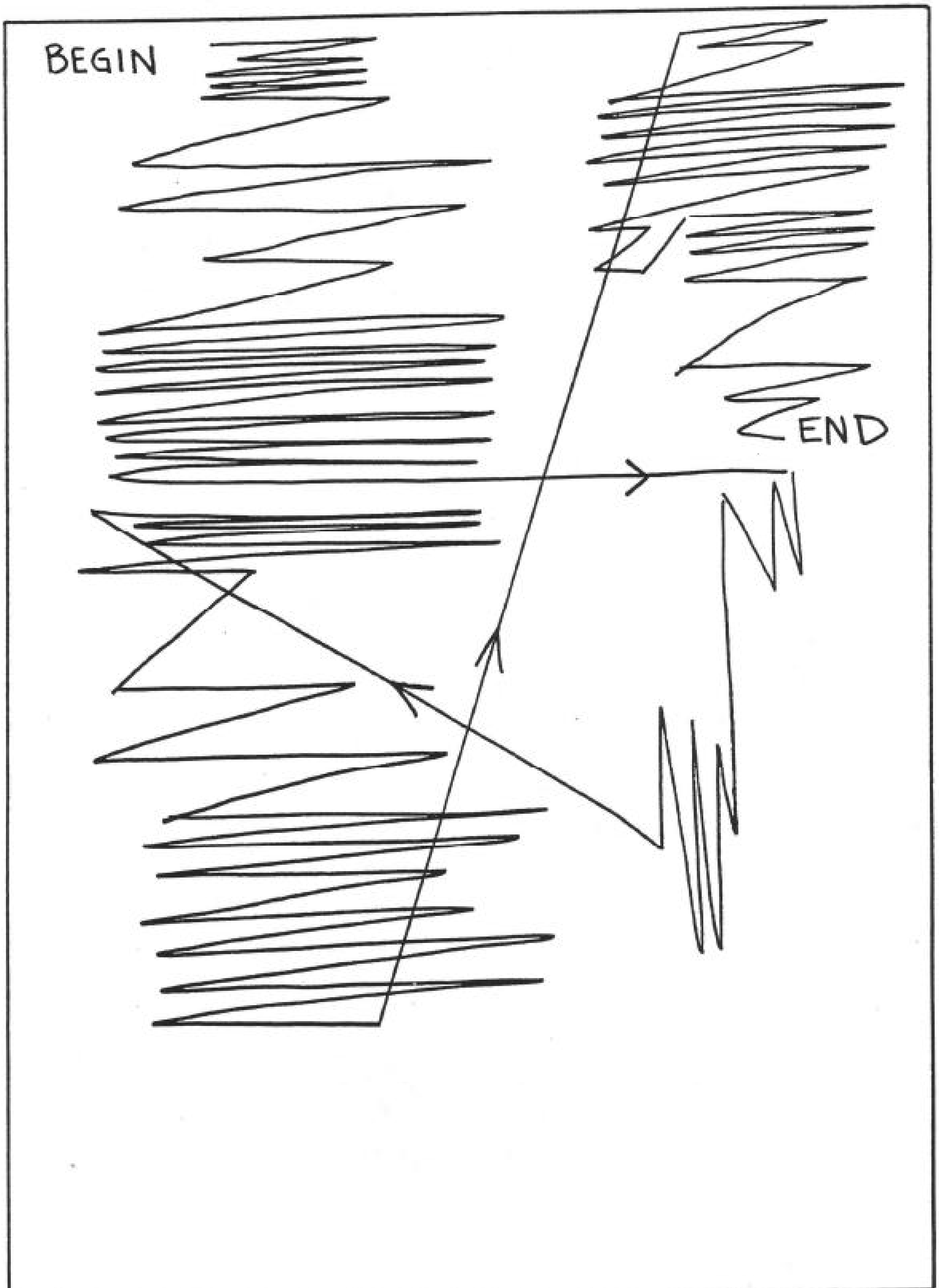
SEQ: 374
TEA: 31.4
ID #: 876842051 2

0 41 1 092 501 3 1 3A 27
PRINCIPAL
Martin High School 09200
1001 4th Street
Nevada, IA 50201

THIS SURVEY HAS BEEN ENDORSED BY:

American Association for Counseling and Development
American Association of School Administrators
American Federation of Teachers
Council of Chief State School Officers
National Association of Elementary School Principals
National Association of Secondary School Principals
National Education Association

Example 1. Cover Page of the Public School Questionnaire (Jenkins et al., 1992b).



Example 2. Diagrammatical Representation of the Public School Questionnaire Cover Page's Reading Structure.

FORM SASS-3X
(1-22-93)

OMB No. 1850-0598: Approval Expires 12/31/93

U.S. Department of Education
National Center for Education Statistics

SCHOOL QUESTIONNAIRE
SCHOOLS AND STAFFING SURVEY

1993 FIELD TEST

Conducted by:

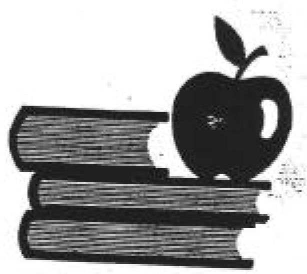
U.S. Department of Commerce
Bureau of the Census

PLEASE COMPLETE THIS QUESTIONNAIRE WITH INFORMATION ABOUT:

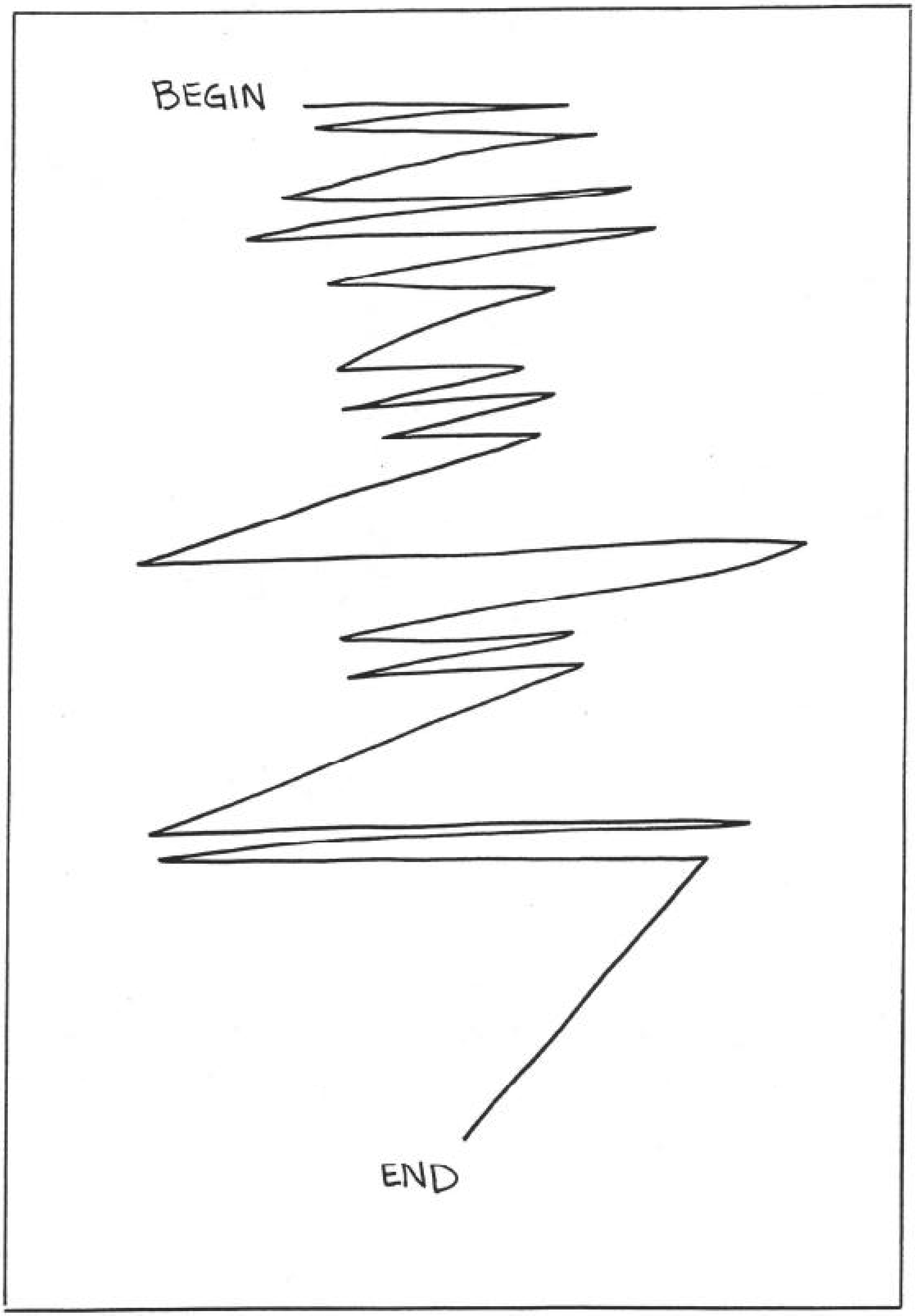
MARTIN HIGH SCHOOL
GRADES 9-12

If you have any questions, call the Bureau of the Census at 1–800–221–1204.

After completing this form, mail it to the Bureau of the Census in the preaddressed envelope provided. Please return it within 2 weeks.



Example 3. Redesigned Version of the Public School Questionnaire's Cover Page (adapted from Jenkins and Ciochetto, 1993).



Example 4. Diagrammatical Representation of the Redesigned Public School Questionnaire Cover Page's Reading Structure.

Dear Principal:

The National Center for Education Statistics (NCES) of the U.S. Department of Education requests your participation in the field test of the 1992–93 Schools and Staffing Survey. Your school is one of 900 public and private schools across the Nation selected to be in the sample.

The Schools and Staffing Survey, first conducted in school year 1987–88, and again in 1990–91, is an integrated set of surveys consisting of the Teacher Demand and Shortage Survey, the School Survey, the School Administrator Survey, and the Teacher Survey. These surveys are being conducted periodically to measure critical aspects of teacher supply and demand, the composition of the administrator and teacher work force, and the general status of teaching and schooling. The purpose of the School Survey is to obtain information about schools such as staff-pupil ratio, student characteristics, staffing patterns, and teacher turnover.

The U.S. Bureau of the Census is conducting the survey for the National Center for Education Statistics by the authority of Section 406(b) of the General Education Provisions Act, as amended (20 U.S.C. 1221e). The data will be treated as confidential and will be reported only in statistical summaries that preclude the identification of any individual participating in the surveys.

We are conducting this field test with a sample of schools. While this minimizes overall response burden, the value of each individual survey response is greatly increased because it represents many other schools. I, therefore, encourage you to participate in this voluntary survey by completing this questionnaire and returning it within 3 weeks to the **Bureau of the Census, Current Projects Branch, 1201 East 10th Street, Jeffersonville, IN 47132-0001,** in the preaddressed envelope enclosed for your convenience.

Thank you for your cooperation in this very important effort.

Sincerely,

Emerson J. Elliott
Acting Commissioner
National Center for Education Statistics

INFORMATION ABOUT YOUR PARTICIPATION

Public reporting burden for this collection of information is estimated to average one hour, including the time for reviewing instructions, gathering the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to the U.S. Department of Education, Information Management and Compliance Division, Washington, DC 20202-4651; and to the Office of Management and Budget, Paperwork Reduction Project 1850-0598, Washington, DC 20503.

Example 5. Introductory Information from the Public School Questionnaire (Jenkins et al., 1992b).

INSTRUCTIONS

Most of the items on this questionnaire are arranged so that the questions are on the left side of the page and the answer categories or spaces for written answers are on the right. Please answer the questions by marking the appropriate answer category with an X, or recording your answer in the space provided. We suggest that you use a pencil or a typewriter, rather than a pen or marker.

Notice that at the end of some answer categories and answer spaces, there are instructions to skip to a later item or to continue with the next item on the questionnaire.

Unless otherwise indicated, all questions refer to the 1991–92 school year.

If you are unsure about how to answer a question, give the best answer you can and make a comment in the "Remarks" space. Please include the item number.

If you have any questions, call the Bureau of the Census at 1–800–221–1204.

Return your completed questionnaire to the Bureau of the Census in the enclosed preaddressed envelope. Please return it within 3 weeks.

Please keep count of the time required to complete this questionnaire. At the end of the survey, you are asked to record the amount of time spent.

a. Please give your name, title, telephone number, and the most convenient days/times to reach you. This information will be used only if it is necessary to clarify any of your responses.

Name

Title

Telephone — Area code | Number

If necessary to reach you – *Specify* — Days | Time a.m. p.m.

b. Does this school serve students in ANY of grades 1 through 12 or comparable ungraded levels?

010 1☐ Yes – *Continue with c.*
2☐ No – ***Stop now** and return this questionnaire to the Bureau of the Census in the enclosed envelope.*

Thank you for your time.

c. Please check the identification number on address label – Is this your School State Identification Number?

011 1☐ Yes
2☐ No – *Provide the correct number below.*
012 ____________

Remarks

Example 5 Continued.

A. Please record the current time. At the end of the questionnaire you are asked to record the amount of time required to complete this questionnaire.

Current time: ____________

B. Does this school provide instruction for grade 9 or above?

009 1 ☐ Yes → What grading system is used to compute a student's grade point average (GPA)?

010

1 ☐ 0.0 to 4.0
2 ☐ 0 to 100
3 ☐ −1 to 3
4 ☐ Other specify ________

2 ☐ No → **Skip to information below**

WHY ARE WE CONDUCTING THIS SURVEY?

This questionnaire is the last in a series of surveys designed to obtain nationwide information on schools, staffing patterns, and student characteristics. We will treat your data as confidential and only use it to prepare statistical summaries.

WHO IS CONDUCTING THIS SURVEY?

The National Center for Education Statistics of the U.S. Department of Education requests your participation in this voluntary survey. The Bureau of the Census is conducting this survey by the authority of Section 406(b) of the General Education Provisions Act, as amended (20 USC 1221e).

INFORMATION ABOUT YOUR PARTICIPATION

Public reporting burden for this collection of information is estimated to average thirty minutes, including the time for reviewing instructions, gathering the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to the U.S. Department of Education, Information Management and Compliance Division, Washington, DC 20202−4651; and to the Office of Management and Budget, Paperwork Reduction Project 1850−0598, Washington, DC 20503.

Example 6. Revised Introductory Material Beginning with Questions Rather Than Prose (Jenkins and Ciochetto, 1993).

Item 15 — DOLLAR VALUE OF BUSINESS DONE IN 1989

		Mil.	Thou.	
a. What was the value of all construction work done by this establishment in 1989? *Report your best estimate using either billings, revenues, receipts, or other estimate of value of construction work done. Refer to instructions for further explanation.* Reported data based on — *Mark (X) in appropriate box.* 136 1 ☐ Billings 3 ☐ Receipts 2 ☐ Revenues 4 ☐ Other — *Specify* ______ ***INCLUDE*** • New construction • Additions, alterations, or reconstruction • Special trades contracting work • Maintenance and repair work • Land development and improvement • Installation and service of equipment • Buildings and other structures built for sale, excluding the value of the land • Construction work on own account ***EXCLUDE*** • The cost of industrial and other specialized machinery which are not an integral part of a structure	137			☐ 0
b. What were your receipts from kinds of business other than those reported in line a above in 1989? ***INCLUDE*** • Architectural and engineering services • Construction management/consulting services • Wholesale trade • Retail trade • Realty services • Rental of machinery, equipment, or buildings to others • Manufacturing • Transportation • Materials sold to contractors • Sale of land	139			☐ 0
c. TOTAL dollar value of business done in 1989 — *Sum of lines a and b* →	140			☐ 0

Example 7. The Original Version of the Dollar Value of Business Item from the Census of Construction Industries Questionnaire (DeMaio and Jenkins, 1991).

Item	Key	Dollars in thousands ($)	Mark (X) if "0"
13a. For this establishment in 1989, —			
(1) what were the receipts or billings for all contract construction work done for others? Exclude the dollar value of items purchased by this establishment that were installed in a building but were not part of its structure, such as production machinery, furniture, etc.	141		☐ 0
(2) what was the estimated dollar value of speculative construction work done on residential and other building projects which you sold or intended to sell? **Include the estimated dollar value of —** • work actually done in 1989, whether buildings were sold or not. • work in progress, that was actually done in 1989. • all improvements to land associated with these building projects done by or for you in 1989. **Exclude the estimated dollar value of —** • work done before January 1989 and after December 1989. • land. Even though land would generally be included in the value of your building projects, the value of the land is not considered construction work done. • work done in 1989, for rent or lease.	142		☐ 0
(3) what was the estimated dollar value of construction work done for this establishment's own use, i.e., not intended for sale, rent, or done under contract for others?	143		☐ 0
(4) what was the dollar value of construction work done? Sum lines (1) through (3).	137		☐ 0
b. What was the dollar value of receipts or billings for all other business activities done by this establishment in 1989? **Include —** • architectural services • building on your own land for rent or lease • construction management services • engineering services • manufacturing • mining • real estate agents and managers • rental of construction machinery or equipment to others • retail trade • subdividing and preparing your own land into lots, for sale, rent, or lease • transportation • wholesale trade • other business activities	139		☐ 0
c. What was the total dollar value of all business done by this establishment in 1989? Sum lines 13a(4) and 13b.	140		☐ 0

Example 8. The Revised Version of the Dollar Value of Business Item from the Census of Construction Industries Questionnaire (DeMaio and Jenkins, 1991).

Item 5 — NUMBER OF EMPLOYEES

How many paid employees, permanent or temporary, full-time or part-time, were on the payroll of THIS ESTABLISHMENT during the pay periods shown at the right? *Include those on paid sick leave, paid holidays, and paid vacations as well as those actually working. If a corporation, include salaried officers and executives of this establishment. If unincorporated, exclude proprietors or partners. Include all employees for whom you file quarterly withholding statements. Do not include your subcontractors or their employees.*

	Number of employees of this establishment during the pay period including the 12th of —			
	March 1989	May 1989	August 1989	November 1989
a. Construction workers — ***Include*** — • Laborers • Apprentices • Journeymen • Craftsmen • Equipment operators and mechanics • Truck drivers and helpers • On-site record keepers • Others engaged directly in construction operations, including supervisors up through the working foreman level	101	102	103	104
b. All other employees — ***Include*** — • Executives • Professionals • Technicians • Purchasing • Accounting • Personnel • Office staff • Supervisors above the working foreman level	105	106	107	108
c. TOTAL NUMBER OF EMPLOYEES — ***Sum of lines a and b above*** →	109	110	111	112

Example 9. The Original Version of the Number of Employees Item from the Census of Construction Industries Questionnaire (DeMaio and Jenkins, 1991).

Your answers to 5 through 9 should be based on all employees for whom you filed withholding statements (Federal Tax Return Form 941). Do NOT include your subcontractors or their employees.

5. During the pay periods which include the 12th of March, May, August, and November 1989, —	Number of employees			
	Pay period including the 12th of March 1989	Pay period including the 12th of May 1989	Pay period including the 12th of August 1989	Pay period including the 12th of November 1989
a. how many construction workers were on the payroll of this establishment? **Include —** • Working foremen • Job-site record keepers • Laborers • Apprentices • Journeymen • Craftsmen • Equipment operators and mechanics • Truck drivers and helpers • Others engaged directly in construction	101	102	103	104
b. how many other employees were on the payroll of this establishment? **Include —** • Supervisors • Job-site and home office clerical and maintenance staff • Personnel staff • Purchasing agents • Accounting staff • Technicians • Architects • Engineers • Professionals • Executives • Others engaged in non-construction activities	105	106	107	108
c. how many total employees were on the payroll of this establishment? Sum lines a and b.	109	110	111	112

Example 10. The Revised Version of the Number of Employees Item from the Census of Construction Industries Questionnaire (DeMaio and Jenkins, 1991).

4. What percent of K–12 students enrolled in this school are male? *Record the percent in whole numbers, not tenths; do not enter a decimal point. Do NOT include prekindergarten or postsecondary students.*	055 ________________ %

4. What percent of the students enrolled in this school are male?

Record the percent in whole numbers, not tenths; do not enter a decimal point. Do NOT include prekindergarten or postsecondary students.

055 ________________ %

4. Excluding prekindergarten or postsecondary students, what percent (in whole numbers) of the students enrolled in this school are male?

055 ________________ %

Example 11. Horizontally Aligned Question Format (Top View), Vertically Aligned Question Format with Stand Alone Instructions (Middle View), and Vertically Aligned Question Format with Incorporated Instruction (Bottom View).

2. How many students (in head counts) were enrolled in THIS SCHOOL (the school named on the questionnaire label) in grades K–12 or comparable ungraded levels – *Include only students enrolled in the school named on the questionnaire label. Do NOT include prekindergarten or postsecondary students.* **a. On or about October 1 of THIS SCHOOL YEAR?**	048 ________ Students
b. On or about October 1 of LAST SCHOOL YEAR?	049 ________ Students 0☐ School not operating in Fall 1990

Example 12. The Student Enrollment Item from the Public School Questionnaire (Jenkins et al., 1992b).

SECTION 2 — STAFFING PATTERNS — Continued

30. How many employees hold full- or part-time positions in this school in each of the following categories?

If an employee holds a position in more than one of the categories, count that person as part-time in each category that applies.

	FULL-TIME	PART-TIME
INSTRUCTIONAL STAFF		
a. Instructional staff – Instructional aides (paraprofessionals who assist classroom teachers)	255 ______ 0☐ None	256 ______ 0☐ None
INSTRUCTIONAL SUPPORT STAFF		
b. Instructional coordinators and supervisors (including curriculum specialists)	257 ______ 0☐ None	258 ______ 0☐ None
c. Librarians/Media specialists	259 ______ 0☐ None	260 ______ 0☐ None
d. Library/Media center aides	261 ______ 0☐ None	262 ______ 0☐ None
e. Guidance counselors	263 ______ 0☐ None	264 ______ 0☐ None
f. Vocational-technical counselors	265 ______ 0☐ None	266 ______ 0☐ None
SUPPORT SERVICES STAFF		
g. Administrators: **(1) Principal(s)**	267 ______ 0☐ None	268 ______ 0☐ None
(2) Vice Principal(s)	269 ______ 0☐ None	270 ______ 0☐ None
(3) Other managers – e.g., business	271 ______ 0☐ None	272 ______ 0☐ None
h. Administrative support staff – Clerical and nonmanagerial support staff	273 ______ 0☐ None	274 ______ 0☐ None
i. Student support services staff – Professionals and supervisory staff providing noninstructional services to students, including health, psychology, social work, or attendance	275 ______ 0☐ None	276 ______ 0☐ None
j. All other support staff (not reported in other categories, such as health aides, maintenance, bus drivers, security, and cafeteria workers)	277 ______ 0☐ None	278 ______ 0☐ None

Example 13. The Classification of Employees Item from the Public School Questionnaire (Jenkins et al., 1992b).

PART-TIME EMPLOYEES

26a. How many employees hold part-time positions in this school in each of the following categories? Please read through all of the categories listed below before starting to answer.

INCLUDE AS PART TIME
- o Employees who work part-time at this school only.
- o Employees you share with other schools within or outside of the school district.
- o Employees who perform more than one function at this school, e.g., a teaching principal would be counted once as a part-time teacher and again as a part-time principal.

	PART-TIME
1. Administrators:	
(a) Principals	☐ None or ____
(b) Vice Principal(s)	☐ None or ____
(c) Other managers, such as business	☐ None or ____
2. Instructional coordinators and supervisors, such as curriculum specialists	☐ None or ____
3. Guidance counselors-- needs a definition	☐ None or ____
4. Vocational-technical counselors-- needs a definition	☐ None or ____
5. Librarians/Media specialists	☐ None or ____
6. Teachers Do not include as teachers-- Other employees listed in this item, unless they also teach. Teachers who teach only prekindergarten students	☐ None or ____
7. Student support services staff, such as school psychologists, social workers, occupational therapists, speech therapists, nurses, and truant officers	☐ None or ____
8. Library/media center aides	☐ None or ____
9. Teacher aides	☐ None or ____
10. Student teachers	☐ None or ____
11. Clerical and nonmanagerial support staff	☐ None or ____

Example 14a. Redesigned Version of the Classification of Employees Item from the Public School Questionnaire, Part A, Part-Time Status (Jenkins et al., 1992b).

FULL-TIME EMPLOYEES

26b. **How many employees hold full-time positions in this school in each of the following categories?** Please read through all of the categories listed below before starting to answer.

		FULL-TIME
1.	Administrators:	
	(a) Principals	☐ None or ____
	(b) Vice Principal(s)	☐ None or ____
	(c) Other managers, such as business	☐ None or ____
2.	Instructional coordinators and supervisors, such as curriculum specialists	☐ None or ____
3.	Guidance counselors-- needs a definition	☐ None or ____
4.	Vocational-technical counselors-- needs a definition	☐ None or ____
5.	Librarians/Media specialists	☐ None or ____
6.	Teachers Do not include as teachers-- Other employees listed in this item. Teachers who teach only prekindergarten students	☐ None or ____
7.	Student support services staff, such as school psychologists, social workers, occupational therapists, speech therapists, nurses, and and truant officers	☐ None or ____
8.	Library/media center aides	☐ None or ____
9.	Teacher aides	☐ None or ____
10.	Student teachers	☐ None or ____
11.	Clerical and nonmanagerial support staff	☐ None or ____
12.	Cafeteria workers	☐ None or

Example 14b. Redesigned Version of the Classification of Employees Item from the Public School Questionnaire, Part B, Full-time Status (Jenkins et al., 1992b).

STUDENT CLASS SCHEDULES WITH SELECTED TEACHERS

JOHN JONES

	1. Excluding study halls and free periods, does John Jones currently have this student for class?	2. Excluding study halls and free periods, please list the classes in which John Jones currently has this student.	3. How many times per week does this class meet?	4. How many total students are enrolled in this class?
KAYE STEWART	[ ] yes	1.		
		2.		
		3.		
	[ ] no			
MARIE LEARY	[ ] yes	1.		
		2.		
		3.		
	[ ] no			
SUZANNE FLANIGAN	[ ] yes	1.		
		2.		
		3.		
	[ ] no			

Example 15. Questions Developed for the Student Records Questionnaire Using a Matrix Format (Jenkins and Ciochetto, 1993).

STUDENT 1's NAME

1a. Excluding homeroom, study halls, and free periods, do you currently teach this student?

010 1 ☐ Yes → 2 ☐ No → **Skip to Item 2a**

b. Do you teach multiple subjects to this student all or most of the day?

011 1 ☐ Yes → **Skip to Item 2a** 2 ☐ No →

c. Excluding homeroom, study halls, and free periods, please list the classes that you teach this student and the number of times per week that each class meets.

Class name	Meetings per week

STUDENT 2's NAME

2a. Excluding homeroom, study halls, and free periods, do you currently teach this student?

012 1 ☐ Yes → 2 ☐ No → **Skip to Item 3a**

b. Do you teach multiple subjects to this student all or most of the day?

013 1 ☐ Yes → **Skip to Item 3a** 2 ☐ No →

c. Excluding homeroom, study halls, and free periods, please list the classes that you teach this student and the number of times per week that each class meets.

Example 16. Redesigned Version of the Questions Developed for the Student Records Questionnaire Using a Single Question Format (Jenkins and Ciochetto, 1993).

SECTION 1 — SCHOOL CHARACTERISTICS — Continued

23a. Does this school offer a general program for students who do not plan to attend college?

214 1☐ Yes – *Continue with b*
2☐ No – *Skip to item 24a*

b. How many students in grades 10–12 are enrolled in this program?

215 _________ Students
0☐ None

24a. LAST SCHOOL YEAR, how many students were enrolled in 12th grade?

216 _________ Students – *Continue with b*
0☐ No 12th graders in 1990–91 – *Skip to the note above item 26a*

b. How many students were graduated from the 12th grade last year? Include 1991 summer graduates.

217 _________ Students – *Continue with c*
0☐ None – *Skip to the note above item 26a*

c. How many of last year's graduates applied to two- or four-year colleges?

218 _________ Graduates
0☐ None

25a. Does this school offer job placement services for graduating seniors?

219 1☐ Yes
2☐ No

b. Does this school have a "Tech-Prep" program, i.e., vocational-technical instruction in the last two years of high school designed to prepare students for two years of vocational instruction at the postsecondary level?

220 1☐ Yes
2☐ No

SECTION 2 — STAFFING PATTERNS

NOTE: *For items 26–28, INCLUDE itinerant teachers and long-term substitutes. DO NOT INCLUDE student teachers, teacher aides, short-term substitutes, teachers who teach ONLY prekindergarten or postsecondary students, and other non-teaching staff (administrators, other professionals such as counselors and librarians, and support staff) unless they also teach part-time. Report in head counts, not FTEs.*

26a. How many K–12 teachers have FULL-TIME teaching positions at THIS school?

221 _________ Full-time teachers
0☐ None

b. How many K–12 teachers have PART-TIME teaching positions at THIS school?

222 _________ Part-time teachers
0☐ None

c. What is the total number of K–12 teachers at this

Example 17. Transitional Heading "Section 2 -- Staffing Patterns" Falling in the Middle of a Page on the Public School Questionnaire (Jenkins et al., 1992b).

SECTION 1 — SCHOOL CHARACTERISTICS — Continued

NOTE – Items 2–7 refer to grades K through 12. Do not include prekindergarten or postsecondary students or grades in answering these questions.

2. How many students (in head counts) were enrolled in THIS SCHOOL (the school named on the questionnaire label) in grades K–12 or comparable ungraded levels –

Include only students enrolled in the school named on the questionnaire label. Do NOT include prekindergarten or postsecondary students.

a. On or about October 1 of THIS SCHOOL YEAR? — 048 ________ Students

b. On or about October 1 of LAST SCHOOL YEAR? — 049 ________ Students

0☐ School not operating in Fall 1990

3. How many K–12 students in this school are –

Do NOT include prekindergarten or postsecondary students.

a. American Indian, Alaskan Native *(Aleut, Alaskan Indian, Yupik, Inupiat)***?** — 050 ________ Students 0☐ None

b. Asian or Pacific Islander *(Japanese, Chinese, Filipino, Korean, Asian Indian, Vietnamese, Hawaiian, Guamanian, Samoan, other Asian)***?** — 051 ________ Students 0☐ None

c. Hispanic, regardless of race *(Mexican, Puerto Rican, Cuban, Central or South American, or other Hispanic culture or origin)***?** — 052 ________ Students 0☐ None

d. Black (not of Hispanic origin)? — 053 ________ Students 0☐ None

e. White (not of Hispanic origin)? — 054 ________ Students 0☐ None

4. What percent of K–12 students enrolled in this school are male?

Record the percent in whole numbers, not tenths; do not enter a decimal point. Do NOT include prekindergarten or postsecondary students. — 055 ________ .0 %

5. How many K–12 students were absent the most recent school day?

Include both excused and unexcused absences. Do NOT include prekindergarten or postsecondary students. — 056 ________ Students 0☐ None

6. How many days are in the school year for students in this school?

Example 18. Transitional Heading and Advanced Instruction Coming at the Top of a Page on the Public School Questionnaire (Jenkins et al., 1992b).

29a. Were there teaching vacancies in this school for this school year, i.e., teaching positions for which teachers were recruited and interviewed?

230 — 1 ☐ Yes – *Continue with b*
2 ☐ No – *Skip to item 30*

b. Did this school have any teaching vacancies this school year that could not be filled with a teacher qualified in the course or grade level to be taught?

231 — 1 ☐ Yes – *Continue with c*
2 ☐ No – *Skip to item 29d*

c. Which of these methods did this school use to cover the vacancy(ies)?

Mark (X) all that apply.

232 — 1 ☐ Cancelled planned course offerings
233 — 2 ☐ Expanded some class sizes
234 — 3 ☐ Added sections to other teachers' normal teaching loads
235 — 4 ☐ Assigned a teacher of another subject or grade level to teach those classes
236 — 5 ☐ Used long-term and/or short-term substitutes
237 — 6 ☐ Used part-time or itinerant teachers
238 — 7 ☐ Hired a less qualified teacher
239 — 8 ☐ Other – *Specify* ↗

____

____

____

Example 19. Conventional Skip Instruction from the Public School Questionnaire (Jenkins et al., 1992b).

1. Have you ever drunk a beer, a glass of wine or a wine cooler, a shot of liquor, or a mixed drink with liquor in it? If you have only had sips from another person's drink, answer "no."
MARK ONE BOX.

1 ☐ Yes → a. Think about the very first time you drank a beer, glass of wine or wine cooler, a shot of liquor, or a mixed drink. Not counting sips you might have had from someone else's drink, how old were you the first time you drank an alcoholic beverage?

_______ years old

CONTINUE WITH QUESTION 2.

2 ☐ No

GO TO QUESTION 2.

YOUR 12-MONTH REFERENCE DATE IS: ______________________

5. The next questions are about the past 12 months—the period from the date written above up to and including today.

Did you smoke a cigarette during the past 12 months?
MARK ONE BOX.

1 ☐ Yes → a. During the past 12 months, have you smoked cigarettes every day or almost every day for two or more weeks in a row?
MARK ONE BOX.

1 ☐ Yes 2 ☐ No

b. During the past 12 months, have you felt that you needed or were dependent on cigarettes?
MARK ONE BOX.

1 ☐ Yes 2 ☐ No

c. During the past 12 months, have you needed larger numbers of cigarettes to get the same effect?
MARK ONE BOX.

1 ☐ Yes 2 ☐ No

d. During the past 12 months, have you have you tried to cut down on your use of cigarettes?
MARK ONE BOX.

1 ☐ Yes 2 ☐ No

e. During the past 12 months, have you felt sick or had withdrawal symptoms because you stopped or cut down on cigarettes?
MARK ONE BOX.

1 ☐ Yes 2 ☐ No

CONTINUE WITH QUESTION 6 ON PAGE 6.

2 ☐ No

GO TO QUESTION 6 ON PAGE 6.

Example 20. Branching Instructions in Questions 1 and 5 of the 1990 National Household Survey and Drug Abuse Questionnaire (Turner et al., 1992).

SECTION 1 – SCHOOL CHARACTERISTICS – CONTINUED

11. Please indicate whether each of the following programs or services is currently available at this school **to the students included in item 6**, either during or outside of regular school hours and regardless of funding source.

a. English as a Second Language – Students with limited English proficiency are provided with intensive instruction in English.

051 1 ☐ Yes

2 ☐ No

SKIP to b.

How many students participate in this program?

052 0 ☐ None or ________ Students

b. Bilingual education – Native language is used to varying degrees in instructing students with limited English proficiency. *For example, transitional bilingual education and structured immersion. Do not include foreign language classes or foreign language immersion programs.*

053 1 ☐ Yes

2 ☐ No

SKIP to c.

How many students participate in this program?

054 0 ☐ None or ________ Students

Example 21. "Salient" Skip Instruction (adapted from Jenkins and Ciochetto, 1993).

PART A – Employment Status During the Week of April 12–18, 1992

A1. Were you working for pay or profit during the week of April 12–18, 1992? This includes being self-employed or temporarily absent from a job (e.g., illness, vacation or parental leave), even if unpaid.

PGM2

001 1☐ Yes – *Skip to A6*

2☐ No

A2. Did you look for work at any time during the 5 weeks between March 8 and April 12, 1992?

002 1☐ Yes

2☐ No

A3. What was your MAIN reason for not working during the week of April 12–18?

Mark (X) One

003 1☐ Retired –*Skip to A5*

2☐ On layoff from a job

3☐ My work is seasonal

4☐ Student

5☐ Family responsibilities

6☐ Chronic illness or permanent disability

7☐ Could not find work or believed no suitable jobs available in my field

8☐ Waiting for new job to begin within 30 days

9☐ Waiting for school to begin

10☐ Did not need or want to work

11☐ Other *Specify*

- -

A4. Had you previously RETIRED from any position (e.g., mandatory retirement or early retirement)?

004 1☐ Yes

2☐ No – *Skip to Part B on Page 4*

A5. When did you retire?

Month Year

005 ___|___ 19 ___|___

NOW SKIP TO PART B ON PAGE 4

A6. During the week of April 12–18, 1992 were you working full time or part time?

009 1☐ Full time (usually worked a total of 35 or more hours per week) – *Skip to A9*

2☐ Part time (usually worked less than 35 hours per week)

A7. Were you seeking full-time work during the week of April 12–18, 1992?

010 1☐ Yes

2☐ No

A8. What was your MOST important reason for holding a part-time position during the week of April 12–18, 1992?

Mark (X) One

011 1☐ Full-time position not available

2☐ Worked part time to accommodate spouse's/partner's job or career

3☐ Worked part time for other family-related reasons

4☐ Preferred part-time position for other reason *Specify*

- -

A9. Although you were working during the week of April 12–18, 1992, had you previously RETIRED from any position (e.g., mandatory retirement, early retirement)?

012 1☐ Yes

2☐ No – *Skip to A11*

A10. When did you retire?

Month Year

013 ___|___ 19 ___|___

A11. For whom did you work during the week of April 12–18, 1992? (IF YOU HAD MORE THAN ONE JOB THAT WEEK: Please answer for the job you considered your principal employment.)

Employer Name

014 -

Street

- -

City/Town

- -

State/Foreign Country

- -

015 x☐ *MARK (X) HERE IF YOU WERE SELF-EMPLOYED*

Example 22. "Intermediate" Skip Instruction from the Survey of College Graduates

Fill in the first student's name from the cover page on the line below.

STUDENT 1's NAME

1. What is this student's current status at this school?

Enrolled	☐1	Skip to 2
Suspended	☐2	Skip to 2
Expelled	☐3	Skip to NEXT STUDENT ON PAGE ...
Transferred	☐4	Skip to NEXT STUDENT ON PAGE ...
Dropout/Chronic Truant (See definition below)	☐5	Skip to NEXT STUDENT ON PAGE ...
Deceased	☐6	Skip to NEXT STUDENT ON PAGE ...
Other	☐7 Specify ______	Skip to 2

2. Is this student male or female?

Male ☐1
Female ☐2

3. What is this student's race/ethnicity?

American Indian or Alaskan Native ☐1
Asian or Pacific Islander ☐2
Hispanic, regardless of race ☐3
Black (not of Hispanic origin) ☐4
White (not of Hispanic origin) ☐5

4a. Excluding homeroom, study halls, and free periods, is this student currently taught by [Teacher 1]?

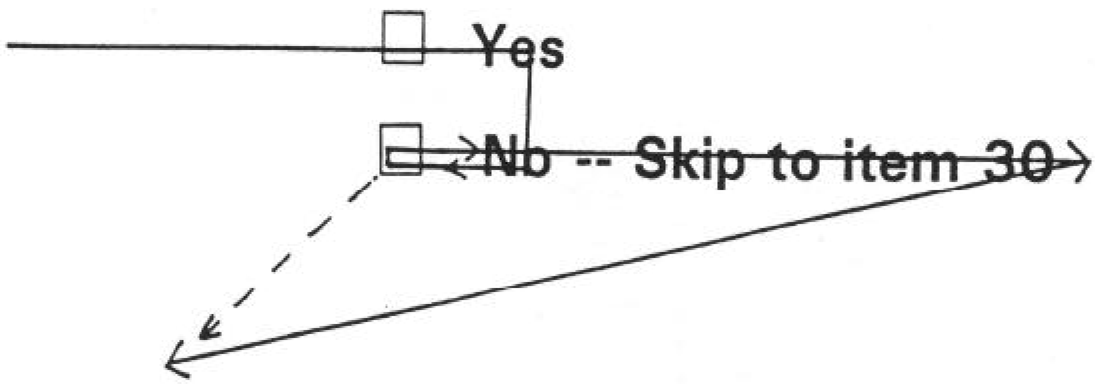
Yes ☐1 Skip to 4b
No ☐2 Skip to 5a

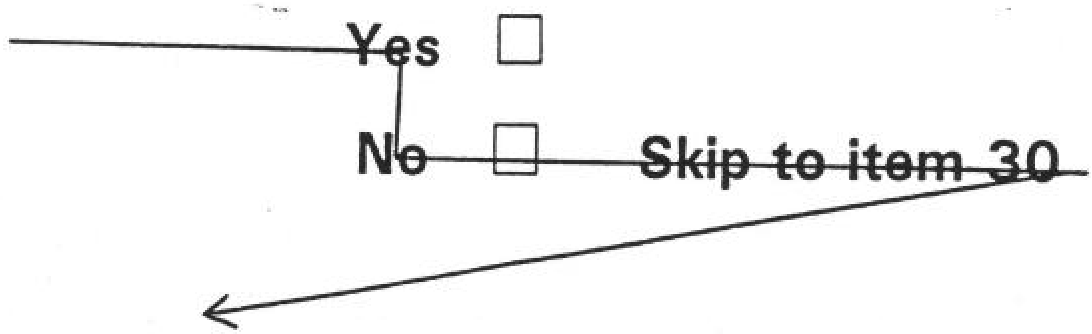
4b. Does [Teacher 1] teach this student all or most of the day?

Yes ☐1 Skip to 5a
No ☐2 Skip to 4c

4c. ...

Example 23. "Natural Reading Sequence" Skip Instruction (Jenkins and Ciochetto, 1993).





Example 24. Diagrammatical Representation of the Conventional (Top View) and "Natural Reading Sequence" (Bottom View) Skip Instruction Reading Structures.

FORM **IPEDS-IC-2**
(6-1-91)

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS
ACTING AS COLLECTING AGENT FOR THE
U.S. DEPARTMENT OF EDUCATION
NATIONAL CENTER FOR EDUCATION STATISTICS

INTEGRATED POSTSECONDARY EDUCATION DATA SYSTEM

INSTITUTIONAL CHARACTERISTICS SURVEY 1991–92

OMB No. 1850-0582: Approval Expires 2/29/92

NOTE – This form is authorized by law (20 U.S.C. 1221e–1). While you are not required to respond, your cooperation is needed to make the results of this survey comprehensive, accurate, and timely.

Public reporting burden for this collection of information is estimated to average 1.0 hours per response but may range from 30 minutes to 2.0 hours depending on whether the information is readily accessible to machine readable files. The burden estimate includes the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to the U.S. Department of Education, Information Management and Compliance Division, Washington, DC 20202-4651, and to the Office of Management and Budget, Paperwork Reduction Project 1850-0582, Washington, DC 20503.

Please read the accompanying instructions before completing this survey form. Respond to each item on this report in the space provided. Certain responses are preprinted. These responses were provided by your institution on the previous IPEDS Institutional Characteristics Survey form. If a response is preprinted, verify that it is correct. If a preprinted response is incorrect, cross out the existing, incorrect response with a single line and clearly indicate the correct response. ***MAKE YOUR CHANGES IN RED*** *so they are easily identified. Be sure to update the enrollment and tuition questions. Certain terms in the various questions are defined in the Glossary which begins on page 5 of the instructions.*

If there are any questions about this form, contact Elaine Kroe, NCES, at (202) 219-1361 or the Bureau of the Census IPEDS representative at (301) 763-4947.

RETURN TO

Date due: September 1, 1991

INSTITUTIONAL IDENTIFICATION 1991-92

Please correct errors in the name, address, and ZIP Code listed below.

1. Name of institution covered by this report — 001
2. UNITID — 002
3. Address *(Number and street name)* — 003
4. Name of county or independent city — 004

City — 005 | State — 006 | ZIP Code — 007

5. Congressional district — 008
6. Name of chief administrator — 009 | Title — 010
7. Name of respondent — 011 | Telephone number — 012
8. Telephone numbers ▶ General information — 013 | Financial aid office — 014 | Admissions office — 015

Part A — TYPES OF EDUCATIONAL OFFERINGS

1. Which of the following types of instruction/programs does your institution offer? *Mark (X) all that apply.*

Example 25. Cover Page of the 1991-92 Integrated Postsecondary Education Data System Questionnaire.

Part II — EDUCATION AND TRAINING — Continued

11. If you are a student attending a college or university, mark your status.

132
1 ☐ Student, full-time
2 ☐ Student, part-time
3 ☐ Not currently a student

12a. Which of these kinds of training did you participate in during 1980 or 1981?

Mark (X) the appropriate year for each type of training you received.

1980 (133)	1981 (134)	KIND OF TRAINING
1 ☐	1 ☐	Military training applicable to present civilian occupation
2 ☐	2 ☐	Extension or correspondence courses applicable to present civilian occupation
3 ☐	3 ☐	Courses at employer's training facility
4 ☐	4 ☐	Courses at adult education center
5 ☐	5 ☐	Courses presented in conjunction with professional meetings
6 ☐	6 ☐	Courses presented by professional training organizations (commercial or non-profit)
7 ☐	7 ☐	Other training
0 ☐	0 ☐	None

b. Were continuing education units (CEU's) or other forms of recognized credit units earned as a result of the above training (in item 12a)?

1980: 135 — 1 ☐ Yes, 2 ☐ No
1981: 136 — 1 ☐ Yes, 2 ☐ No

Part III — EMPLOYMENT STATUS

13. During the week of May 9, 1982, were you —

137
1 ☐ Working full time (35 hours or more per week in at least one position) — *SKIP to 17a*
2 ☐ Working part time — *GO to 14*
3 ☐ Not working, but seeking work — *SKIP to Part IV*
4 ☐ Not working and not seeking work — *SKIP to 15*

14. Were you seeking full-time work?

138
1 ☐ Yes
2 ☐ No
} *SKIP to 17a*

15. Did you look for work at any time during the 3 weeks PRIOR to the week of May 9, 1982?

139
1 ☐ Yes
2 ☐ No

16. What was the MAIN reason you were not working or not seeking work during the week of May 9, 1982?

Mark (X) only one box.

140
1 ☐ On layoff from a job
2 ☐ On vacation or otherwise temporarily absent from a job for health or personal reasons
} *GO to 17a*
3 ☐ Retired
4 ☐ Student
5 ☐ Family responsibilities
6 ☐ Chronic illness or permanent disability
7 ☐ Could not find work or believed no jobs available in my particular field
8 ☐ Did not want to work
9 ☐ New job to begin within 30 days
10 ☐ Waiting for school to begin
11 ☐ Other — *Specify* ______
} *SKIP to Part IV*

17a. During the week of May 9, 1982, were you working at (or on layoff from) a position related to the natural sciences, social sciences, or engineering?

141
1 ☐ Yes — *SKIP to Part IV*
2 ☐ No — *GO to b*

Example 26. A Page from a Conventionally Designed Questionnaire.

C7. Do you currently belong to any national professional societies or associations?

119 1 ☐ Yes

2 ☐ No

C8. In the 2 years between April 1990 and April 1992, did you attend any work-related workshops, seminars, or other work-related training activities?

- *Do **not** include college courses – these will be discussed in Part D.*
- *Do **not** include professional meetings unless you attended a special training session conducted at the meeting/conference.*

120 1 ☐ Yes

2 ☐ No – *Skip to C12*

C9. In which of the following areas did you attend work- related workshops, seminars, or other work-related training activities?

Mark (X) Yes or No for each

		Yes	No
121	**a.** Management or supervisor training. .	1 ☐	2 ☐
122	**b.** Technical training in my occupational field	1 ☐	2 ☐
123	**c.** General professional training (e.g., public speaking, business writing). . .	1 ☐	2 ☐
124	**d.** Other work-related training *Specify*	1 ☐	2 ☐

_ _

_ _

C10. For which of the following reasons did you attend training activities between April 1990 and April 1992?

Mark (X) Yes or No for each

		Yes	No
125	**a.** To acquire further skills or knowledge in my current occupational field.	1 ☐	2 ☐
126	**b.** To acquire skills or knowledge in a different field	1 ☐	2 ☐
127	**c.** For licensure/certification.	1 ☐	2 ☐
128	**d.** To increase opportunities for promotion/advancement/higher salary	1 ☐	2 ☐
129	**e.** To learn skills or knowledge needed for a recently acquired position (e.g., orientation).	1 ☐	2 ☐

C11. What were your TWO most important reasons for attending training activities?

132 _ _ _ Most IMPORTANT REASON from Question C10 – *(Enter LETTER)*

133 _ _ _ SECOND most IMPORTANT REASON from Question C10 – *(Enter LETTER)*

C12. During the 52 weeks of the 1991 calendar year, how many weeks were you –

NUMBER OF WEEKS

134 _ _ \| _ _ **Working,** including weeks of **paid** vacation, **paid** sick leave, and military service?

135 _ _ \| _ _ **On an unpaid** leave of absence from a job?

136 _ _ \| _ _ **Not working** but seeking work?

137 _ _ \| _ _ **Not working** and **not** seeking work?

5 \| 2 **TOTAL**

C13. What was your TOTAL EARNED income, BEFORE deductions for 1991?

Include all wages, salaries, bonuses, overtime, commissions, consulting fees, net income from businesses, summertime teaching or research, post-doctoral appointment, or other work associated with scholarships.

138 $ _ _ _ _ _ _ _ _ _ _ _ .00

Total 1991 earned income

Example 27. A Page from a Redesigned Questionnaire.

PLEASE ALSO ANSWER HOUSING QUESTIONS ON PA

	PERSON 1	PERSON 2
Please fill one column ➡ for each person listed in Question 1a on page 1.	Last name First name — Middle initial	Last name First name — Middle initial
2. How is this person related to PERSON 1? Fill ONE circle for each person. If **Other relative** of person in column 1, fill circle and print exact relationship, such as mother-in-law, grandparent, son-in-law, niece, cousin, and so on.	START in this column with the household member (or one of the members) in whose name the home is owned, being bought, or rented. If there is no such person, start in this column with any adult household member.	If a RELATIVE of Person 1: ○ Husband/wife ○ Natural-born or adopted son/daughter ○ Stepson/stepdaughter ○ Brother/sister ○ Father/mother ○ Grandchild ○ Other relative If NOT RELATED to Person 1: ○ Roomer, boarder, or foster child ○ Housemate, roommate ○ Unmarried partner ○ Other nonrelative
3. Sex Fill ONE circle for each person.	○ Male ○ Female	○ Male ○ Female
4. Race Fill ONE circle for the race that the person considers himself/herself to be. If **Indian (Amer.)**, print the name of the enrolled or principal tribe. → If **Other Asian or Pacific Islander (API)**, print one group, for example: Hmong, Fijian, Laotian, Thai, Tongan, Pakistani, Cambodian, and so on. → If **Other race**, print race. →	○ White ○ Black or Negro ○ Indian (Amer.) (Print the name of the enrolled or principal tribe.) ○ Eskimo ○ Aleut Asian or Pacific Islander (API) ○ Chinese ○ Japanese ○ Filipino ○ Asian Indian ○ Hawaiian ○ Samoan ○ Korean ○ Guamanian ○ Vietnamese ○ Other API ○ Other race (Print race)	○ White ○ Black or Negro ○ Indian (Amer.) (Print the name of the enrolled or principal tribe.) ○ Eskimo ○ Aleut Asian or Pacific Islander (API) ○ Chinese ○ Japanese ○ Filipino ○ Asian Indian ○ Hawaiian ○ Samoan ○ Korean ○ Guamanian ○ Vietnamese ○ Other API ○ Other race (Print race)
5. Age and year of birth **a.** Print each person's age at last birthday. Fill in the matching circle below each box. **b.** Print each person's year of birth and fill the matching circle below each box.	a. Age: 0 ○ 0 ○ 0 ○; 1 ○ 1 ○ 1 ○; 2 ○ 2 ○; 3 ○ 3 ○; 4 ○ 4 ○; 5 ○ 5 ○; 6 ○ 6 ○; 7 ○ 7 ○; 8 ○ 8 ○; 9 ○ 9 ○ b. Year of birth: 1 ● 8 ○ 0 ○ 0 ○; 9 ○ 1 ○ 1 ○; 2 ○ 2 ○; 3 ○ 3 ○; 4 ○ 4 ○; 5 ○ 5 ○; 6 ○ 6 ○; 7 ○ 7 ○; 8 ○ 8 ○; 9 ○ 9 ○	a. Age: 0 ○ 0 ○ 0 ○; 1 ○ 1 ○ 1 ○; 2 ○ 2 ○; 3 ○ 3 ○; 4 ○ 4 ○; 5 ○ 5 ○; 6 ○ 6 ○; 7 ○ 7 ○; 8 ○ 8 ○; 9 ○ 9 ○ b. Year of birth: 1 ● 8 ○ 0 ○ 0 ○; 9 ○ 1 ○ 1 ○; 2 ○ 2 ○; 3 ○ 3 ○; 4 ○ 4 ○; 5 ○ 5 ○; 6 ○ 6 ○; 7 ○ 7 ○; 8 ○ 8 ○; 9 ○ 9 ○
6. Marital status Fill ONE circle for each person.	○ Now married ○ Separated ○ Widowed ○ Never married ○ Divorced	○ Now married ○ Separated ○ Widowed ○ Never married ○ Divorced

Example 28. Original Census Questionnaire Using a Matrix Format.

THE 1992 NATIONAL CENSUS TEST

Four very brief questions about each person living in your household on April 1, 1992.

Please complete both sides of this questionnaire and return in the postage-paid envelope.

OMB No. 0607-0728
Approval Expires 09/30/92

4916 Livingstone St.
Bremer, WA 90001

A. Please answer the four brief questions for each person living at this address → as of Wednesday, April 1, 1992.

DON'T INCLUDE PERSONS WHO:

- Usually live somewhere else
- Are away in an institution such as a prison, mental hospital, or nursing home
- Are college students living somewhere else while attending college
- Are in the Armed Forces and live somewhere else
- Stay somewhere else most of the week while working

BE SURE TO INCLUDE PERSONS WHO:

- Usually live here, such as family members, housemates and roommates, foster children, roomers, boarders, and live-in employees
- Are temporarily away on a business trip, on vacation, or in a general hospital
- Are college students staying here while attending college
- Are in the Armed Forces and live here
- Are newborn babies still in the hospital
- Are children in boarding schools below the college level
- Stay here most of the week while working, even if they have a home somewhere else
- Have no other home and are staying here on April 1

(Start with the household member, or one of the household members, in whose name this house or apartment is owned, being bought, or rented. If there is no such person, start with any adult household member.)

PERSON 1

NAME ______ ______ ______
(Print) Last First Initial

1. What is this person's sex?
☐ Female
☐ Male

2. What is this person's date of birth?
☐☐ Month ☐☐ Day ☐☐☐☐ Year

(It is important to answer both Questions 3 and 4)

3. Is this person of Spanish/Hispanic origin?
☐ No, not Spanish/Hispanic
☐ Yes, Mexican, Mexican-Am., Chicano
☐ Yes, Puerto Rican
☐ Yes, Cuban
☐ Yes, other Spanish/Hispanic *(Print group)* → ______

4. What is this person's race?
☐ White
☐ Black or Negro
☐ American Indian *(Print name of enrolled or principal tribe)* → ______
☐ Eskimo
☐ Aleut
☐ Chinese
☐ Japanese
☐ Filipino
☐ Asian Indian
☐ Hawaiian
☐ Samoan
☐ Korean
☐ Guamanian
☐ Vietnamese
☐ Other Asian or Pacific Islander *(Print race)* → ______
☐ Some other race *(Print race)* → ______

PERSON 2

NAME ______ ______ ______
(Print) Last First Initial

4. What is this person's race?
☐ White
☐ Black or Negro
☐ American Indian *(Print name of enrolled or principal tribe)* → ______

Example 29. Redesigned Census Questionnaire Using An Individual Space Format.

DUE BY JANUARY 1, 1991

OMB No. 0607-0672: Approval Expires 06/30/91

FORM 80-A17 (10-10-90)

U.S. DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

UNITED STATES CENSUS OF AGRICULTURE

NOTICE — Response to this inquiry is required by law (title 13, U.S. Code). By the same law YOUR REPORT TO THE CENSUS BUREAU IS CONFIDENTIAL. It may be seen only by sworn Census employees and may be used only for statistical purposes. Your report CANNOT be used for purposes of taxation, investigation, or regulation. The law also provides that copies retained in your files are immune from legal process.

In correspondence pertaining to this report, please refer to your Census File Number (CFN)

Please complete this form and RETURN TO BUREAU OF THE CENSUS, 1201 East Tenth Street, Jeffersonville, IN 47133

Note — If your records are not available, reasonable estimates may be used. If you cannot file by January 1, a time extension request may be sent to the above address. Include your 12-character Census File Number (CFN) as shown in your address label in all correspondence to us.

If you received more than one report form, enter extra Census File Number(s) here and return extra copies with your completed report.
A –
A –

CENSUS USE ONLY: 035 036 037 038 039 040 041 042

80-A17

CFN 08603545545 EI 86-0354554
6 24 04 10 00 802100 F 8002
NAVAJO 08603545545 4H 00100

L. David Sinclair
4916 Livingstone St.
Bremer, WA 90001

Please correct errors in name, address, and ZIP Code. ENTER street and number if not shown.

SECTION 1

1. At **any** time **during 1990**, did you do any of the following:
   - operate a farm or ranch (including greenhouse and nursery)?
   - sell your farm or ranch?
   - retire from farming or ranching?

   1 ☐ Yes 2 ☐ No

2. At **any** time **during 1990**, did you grow or raise any:
   - crops (including **hay**, tobacco, fruits, vegetables, nuts, berries, etc.) for **sale**?
   - livestock or poultry (including **horses**, cattle, hogs, sheep, goats, fish, etc.) for **sale**?
   - crops, livestock, or poultry for **home use or as a hobby**?

   3 ☐ Yes 4 ☐ No

If you answered YES to EITHER of these questions, go to SECTION 2.
If you answered NO to BOTH of these questions, go to SECTION 10.

SECTION 2 ACREAGE IN 1990

Report land owned, rented, or used by you, your spouse, or by the partnership, corporation, or organization for which you are reporting. *Include ALL LAND, REGARDLESS OF LOCATION OR USE — cropland, pastureland, rangeland, woodland, idle land, house lots, etc.*

	None	Number of acres
1. All land owned	☐	043
2. All land rented or leased FROM OTHERS, including land worked by you on shares, used rent free, in exchange for services, payment of taxes, etc. *Include leased Federal, State, and railroad land. (DO NOT include land used on a per-head basis under a grazing permit.)*	☐	044
3. All land rented or leased TO OTHERS, including land worked on shares by others and land subleased. *Also complete item 5 below.*	☐	045
4. **Acres in "THIS PLACE"** — *ADD acres owned (item 1) and acres rented (item 2), then SUBTRACT acres rented TO OTHERS (item 3), and enter the result in this space.*		046

If the entry is zero, please refer to the Information Sheet, section 2.

5. Of the land you rented or leased to others, how many acres did you own? None ☐ 053 ______ Acres

6. In what county was the largest value of your agricultural products raised or produced? County name ______ State ______

SECTION 3 LAND USE and IRRIGATION

• PART A — How were the ACRES in this place used in 1990?

	None	Number of acres
1. Cropland harvested — *Include all land from which crops were harvested or hay was cut, and all land in orchards, citrus groves, vineyards, and nursery and greenhouse crops.*	☐	787
2. Cropland on which all crops failed — *(Exception: Do not report here land in orchards and vineyards on which the crop failed.)*	☐	790
3. Cropland idle, cropland used for cover crops, or cropland in cultivated summer fallow	☐	793
4. Cropland used only for pasture, woodland pastured, and other pastureland and rangeland	☐	795
5. All other woodland, wasteland, houselots, etc. not reported in items 1 through 4 above	☐	797

SECTION 4

• PART A — CROPS HARVESTED from "THIS PLACE" in 1990
(Do not include crops grown on land rented to others.)

	None	Acres harvested	Quantity harvested	Gross value of crops sold — Dollars	Cents
1. Hay crops — a. Alfalfa and alfalfa mixtures	☐		Tons dry	$	00
b. Small grain hay	☐		Tons dry	$	00
c. Wild hay	☐		Tons dry	$	00
d. Other hay — Specify kind ______	☐		Tons dry	$	00
2. Corn for grain or seed	☐		Bu.	$	00
3. Soybeans for beans	☐		Bu.	$	00
4. Wheat for grain	☐		Bu.	$	00
5. Tobacco — all types	☐	/10	Lbs.	$	00
6. Potatoes, Irish — *(Do not include those grown for home use.)*	☐	/10	Cwt.	$	00

7. All vegetables for sale *(Do not include those grown for home use.)*	None	Total acres	Dollars	Cents
	☐	/10	$	00
Specify kind(s)		/10		
		/10		

8. All fruit and nut orchards, vineyards, and berries	None	Total acres	Quantity harvested	Dollars	Cents
	☐	/10		$	00
Specify kind(s)		/10	Lbs.		
		/10	Lbs.		

9. Other crops — *For additional crops, enter the crop name and code from the list below. Report quantity harvested in the unit specified with crop name.*

Crop name	Code	Acres harvested	Quantity harvested	Gross value of crops sold — Dollars	Cents
				$	00
				$	00

If more space is needed use a separate sheet of paper.

Crop name	Code	Crop name	Code
Barley for grain (bushels)	079	Oats for grain (bushels)	076
Corn for silage or green chop (tons, green)	070	Sorghum for grain-milo (bushels)	082
Cotton (bales)	091	Other crops (pounds) – Specify	782

• PART B — NURSERY and GREENHOUSE CROPS GROWN FOR SALE on "THIS PLACE" in 1990

From the list below, enter the crop name and code for each crop grown.

Crop name	Code	Square feet under glass or other protection	Acres in the open in 1990 — Whole acres	Tenths	Sales in 1990 — Dollars	Cents

Example 30. Cover Page of the 1990 Census of Agriculture Questionnaire (Photo-reduced to 74 Percent of Original Size).

DISCUSSION

Jared B. Jobe
National Center for Health Statistics

These excellent papers approach the topic of self-administered questionnaires from different perspectives: Lessler and O'Reilly discuss the use of audio computer-assisted self-administered questionnaires for sensitive questions; Jenkins and Dillman describe the use of graphic design principles, as well as cognitive and motivational factors in the design of paper and pencil self-administered questionnaires. Its a privilege to have the opportunity to discuss their results, relate their findings to the literature, and offer some suggestions for future work.

Presentation by Lessler and O'Reilly

Lessler and O'Reilly describe the evidence that self-administered questionnaires result in more reports of sensitive behaviors. They describe the disadvantages of self-administered questionnaires; some of these problems were discovered by cognitive testing. They describe computer-assisted self-administered interviews (CASI), and audio CASI data collection, as well as some of audio CASI's advantages. Finally, they presented the results of an experiment conducted during the National Survey of Family Growth (NSFG) Cycle 5 Pretest.

The first issue that I'll address is respondents' ratings of their preferences of method of administration. Lessler and O'Reilly describe a small-scale study by O'Reilly, Hubbard, Lessler, Biemer, and Turner (forthcoming) showing that respondents preferred the audio CASI method. A greco-latin square design was used. Thus, all respondents' ratings were unbiased for the three self-administration methods tested. In the present study by Lessler and O'Reilly, respondents also recommended audio CASI for reporting abortions. However, such a finding should be interpreted with caution. Respondents could only compare audio CASI to the computer-assisted interview. Contrast that result with those from a collaborative study between the National Center for Health Statistics (NCHS) and the National Opinion Research Center (NORC) (Jobe, Pratt, Tourangeau, Baldwin, and Rasinski, forthcoming). In the NCHS-NORC study, focus group respondents recommended focus groups for collecting sensitive information, and respondents in a CASI pretest recommended CASI. Both the Lessler and O'Reilly study and the NCHS-NORC study lack appropriate comparison groups.

All these studies impose large demand characteristics on respondents (see e.g., Orne, 1969). Therefore, my conclusion is that respondents will usually prefer the administration method that they have just experienced. Thus, the audio CASI may not be as strongly preferred as Lessler and O'Reilly's results would suggest.

A second issue is whether method of administration effects occur when collecting data about sensitive topics. This is a timely question. The O'Reilly *et al.* study, described by Lessler and O'Reilly, found that the two CASI methods, audio and video, produced more reports of marijuana and cocaine use than the paper and pencil self-administered questionnaire. Few differences in sexual behaviors were noted. In Lessler and O'Reilly's study, the 178 respondents reported 48 abortions when using audio CASI, compared to 42 abortions during the regular interview.

This result can be contrasted with results of the NCHS-NORC study mentioned above (Jobe *et al.*, forthcoming). We crossed computer-assisted and paper and pencil interview modes with interviewer and self administration. Computerization had no effects on a variety of reports of sensitive questions. Self administration, however, resulted in more reports of some sensitive behaviors: As shown in the top panel of Table 1, more sex partners were reported for the last year, last five years, and lifetime with self administration. As shown in the middle panel of Table 1, more condom use was reported in the last 30 days and past year (marginally significant) with self administration. As shown in the bottom panel of Table 1, more respondents reported a sexually transmitted disease with self administration. Neither computerization nor self administration had any main effects on reports of abortions or drug use.

Table 1

Mean sexual partners and rates of sexually transmitted diseases by method of administration

	Method of Administration		
	Self-Administered	Administered by Interviewer	Ratio
Number of Sexual Partners			
Past Year	1.71	1.44	1.19
Past Five Years	3.87	2.82	1.37
Lifetime	6.51	5.43	1.20
Condom Use			
Past 30 Days	46.7%	35.3%	1.32
Past Year	23.8%	17.9%	1.33
Sexually Transmitted Diseases	22.0%	17.0%	1.29

Source: National Center for Health Statistics

Recently, Boekeloo, Schiavo, Rabin, Conlon, Jordan, and Mundt (1994) reported that patients at a sexually transmitted disease clinic reported more high-risk sex behaviors to 2 of 16 questions for audio CASI compared to a written self-administered questionnaire. Both were superior to a face-to-face interview. They also found fewer missing responses with the audio CASI.

Thus, the studies I have discussed and several others, indicate that self-administered questionnaires may result in more reports of sensitive behaviors than interviewer-administered questionnaires. However, not all sensitive behaviors are reported more frequently in every study showing self-administration effects. Moreover, there is some evidence that computer or audio CASI is slightly superior to other forms of self-administered questionnaires. We clearly need more studies comparing different methods of administration, especially comparing written, video, and audio self administration. My hypotheses are that, across studies, the effects will be small and equivocal among those three, but that effects will be larger and more consistent for self administration over interviewer administration.

A third issue raised by Lessler and O'Reilly's presentation is the use of incentives in sensitive surveys. They manipulated incentives as a variable, in addition to audio CASI. Table 2 shows additional data from their study, reported by Mosher and Duffer (1994). Self administration and incentives worked in an additive fashion: The highest percentages of respondents reported an abortion with a $20 incentive and audio CASI (30%), and with a $40 incentive and no audio CASI (29%). Intermediate levels of abortion reporting were found with audio CASI and no incentive (25%), and with a $20 incentive and no audio CASI (22%). The lowest level of reporting was found with no audio CASI and no incentive (14%).

Table 2

National Survey of Family Growth Cycle 5 Pretest

Group	% Reporting Abortion	No. in Group
In-home, No $, No Audio CASI	14%	(n = 96)
In-home, $20, No Audio CASI	22%	(n = 72)
In-home, No $, Audio CASI	25%	(n = 98)
In-home, $20, Audio CASI	30%	(n = 80)
Off Site, $40, No Audio CASI	29%	(n = 147)

Source: National Center for Health Statistics

In the NCHS-NORC study (Jobe *et al.*, forthcoming), incentives were used only with neutral site interviews, and respondents interviewed in their own home were not paid. We found no main effects on the incentive/site variable for any of the tested sensitive behaviors. Thus, in one study incentives had an effect, and in another study incentives had no effect. Clearly, more research is needed on the effects that incentives have on responding to sensitive questions.

Presentation by Jenkins and Dillman

Jenkins and Dillman presented 20 principles for designing self-administered questionnaires. In her conclusions, Ms. Jenkins states, "Little information on the design of self-administered questionnaires existed until relatively recently. That which did was based primarily on common sense and individual experience." What is significant and interesting about this statement is that it is so close to statements researchers made about the design of interviewer-administered questionnaires before cognitive psychology began to make an impact a decade or so ago.

I am excited by their approach. Their principles have a high degree of face validity. Potentially, attention to graphic design features as well as cognitive and motivational factors could improve self-administered questionnaires as much as cognitive interviews have improved interviewer-administered questionnaires. However, in order for this to occur, two major differences must be overcome between how cognitive psychologists approached questionnaire design and how Jenkins and Dillman have approached self-administered questionnaires.

The first difference is that, from the beginning, cognitive psychologists involved in questionnaire design have utilized the theories and results from cognitive psychology (for reviews, see Jobe and Mingay, 1991; Jobe, Tourangeau, and Smith, 1993). These scientific citations helped convince people that questionnaire design could be more of a science and less of an art. Researchers integrated basic and applied cognitive research on language comprehension, memory encoding and retrieval, frequency and magnitude estimation, heuristics, and decision processes. These are described in articles and books dating back to the beginning of the survey research-cognitive science collaboration (e.g., Hippler, Schwarz, and Sudman, 1987; Jabine, Straf, Tanur, and Tourangeau, 1984; Moss and Goldstein, 1979).

A prime example is an excellent discussion by Fred Smith (Smith, 1991). He described how cognitive laboratory research on free recall, on frequency estimation, and on magnitude estimation applied to the respondents' tasks of recalling their previous day's intake, estimating the frequencies with which they eat foods, and estimating the sizes of their portions. Awareness of the literature on cognitive theory and research has resulted in the use of these theories and application of results in questionnaire design research. Researchers who are not cognitive psychologists have been able to this knowledge to design excellent cognitive experiments on questionnaire design.

Jenkins and Dillman state that a major reason for their papers is to encourage experimental research on the issues raised by their 20 principles. Theirs is a laudable goal, and an attainable one. A paper with these well thought out principles will encourage more research, if it is well grounded scientifically. A necessary next step for Jenkins and Dillman is to integrate significant research, some cognitive, some social, which is applicable to their principles. Relevant research has been conducted on reading comprehension (e.g., Graesser and Bower, 1990), eye movement (e.g., Carpenter and Just, 1983), respondent effort (Krosnick, 1991), politeness (e.g., Grice, 1975), and impression management (e.g., Schwarz, 1993), to name a few areas.

The second difference between this presentation and the cognitive approach is that cognitive psychologists have well described the applicable methodology so that other people can use it. The most prominent example of this is the cognitive interview (e.g., Lessler, Tourangeau, and Salter, 1989; Willis, Royston, and Bercini, 1991). Although different types of cognitive interviews are used, the one most frequently used in questionnaire design is the concurrent think aloud with probes. This methodology has been described sufficiently so that the largest federal statistical agencies, university survey laboratories, and private survey organizations now use cognitive interviews and do it well. It is not very difficult to learn, although there are individual differences in skill at conducting cognitive interviews.

Jenkins and Dillman have NOT described appropriate techniques for all their principles so that others can use them. For example, they refer to graphic design principles in their paper. But, after reading this paper and a much longer version of the same paper--I am unable to describe these graphic design features. In Principle two they state about Example 3, "This cover page uses natural reading format and graphical design features." The same problem occurs on other principles such as numbers 11 and 12. Principle 11 uses the same two terms, and yet they are never defined. For these principles to be helpful, they must not merely give examples of how the Census Bureau successfully solved questionnaire design problems for a particular survey (I am impressed with their success), but they must educate people so that they can use them on their own questionnaires.

Several of the principles ARE self explanatory and easy to implement. For example, I developed a solution to the problem of multi-task formats identified in Principle 8 and illustrated in Examples 13 and 14. In the solution described by Jenkins and Dillman, the respondent must still perform two mental calculations at a time. In contrast, another solution would be to ask respondents to report the total number of employees in each category, such as teachers, guidance counselors, and teachers aides. Then the respondent can be asked to divide the employees in each category into full time and part time workers. Not only does this solution require the respondent to perform only one mental task at a time, but it more closely matches how the information is likely to be organized in the respondent's long-term memory. The solution I just described also illustrates my earlier point that knowledge of relevant scientific literature can make these principles more effective.

A third issue, and one that illustrates the effectiveness of the principles, is the split-ballot experiment. The one described by Jenkins and Dillman used 5 experimental questionnaires, and is a dramatic example of how these principles can be tested experimentally and shown to be effective. This study demonstrated that large structural and organization resulted in large improvements in item and response rates. Note also that smaller changes produce smaller results.

I can conclude by stating that these 20 principles have the potential to revolutionize the design of self-administered questionnaires. However, the long-term effectiveness of these principles may be determined, at least in part, by how their scientific underpinnings are explicated, and by how their everyday use is described.

ACKNOWLEDGEMENT: The author is greatly indebted to Douglas Herrmann for his many valuable suggestions on an earlier version of this paper.

REFERENCES

Boekeloo, B. O., Schiavo, L., Rabin, D. L., Conlon, R. T., Jordan, C. S. and Mundt, D. J. (1994). Self-reports of HIV risk factors by patients at a sexually transmitted disease clinic: Audio vs. written questionnaires. *American Journal of Public Health, 84,* 754-760.

Carpenter, P. A. and Just, M. A. (1983). What your eyes do while your mind is reading. In K. Rayner (Ed.), *Eye movements in reading: Perceptual and language processes*. New York: Academic Press.

Graesser, A. C. and Bower, G. H. (1990). (Eds.). *The psychology of learning and motivation: Inferences and text comprehension.* New York: Academic Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Syntax and semantics: Speech acts 3* (pp. 41-58). New York: Academic Press.

Hippler, H. J., Schwarz, N. and Sudman, S. (Eds.). (1987). *Social information processing and survey methodology*. New York: Springer-Verlag.

Jabine, T. B., Straf, M. L., Tanur, J. M. and Tourangeau, R. (Eds). (1984). *Cognitive aspects of survey methodology: Building a* *bridge between disciplines*. Washington, DC: National Academy Press.

Jobe, J. B. and Mingay, D. J. (1991). Cognition and survey measurement: History and overview. *Applied Cognitive Psychology, 5,* 175-192.

Jobe, J B., Pratt, W. F., Tourangeau, R., Baldwin, A. K. and Rasinski, K. A. (forthcoming). Effects of interview mode on sensitive questions in a fertility survey. In L. Lyberg *et al.* (Eds.), *Survey measurement and process quality*.

Jobe, J. B., Tourangeau, R. and Smith, A. F. (1993). Contributions of survey research to the understanding of memory. *Applied Cognitive Psychology, 7,* 567-584.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5,* 213-236.

Lessler, J., Tourangeau, R. and Salter, W., (1989). Questionnaire design in the cognitive research laboratory. *Vital and Health Statistics,* Series 6, No. 1 (DHHS Publication No. PHS 89-1076). Washington, DC: U.S. Government Printing Office.

Mosher, W. D. and Duffer, A. P., Jr. (1994, May). *Experiments in survey data collection: The National Survey of Family Growth Pretest.* Paper presented at the meeting of the Population Association of America, Miami.

Moss, L. and Goldstein, H. (Eds.). (1979). *The recall method in social surveys.* London: NFER Publishing Co.

O'Reilly, J. M., Hubbard, M. L., Lessler, J. T., Biemer, P. P. and Turner, C. F. (forthcoming). Audio and video computer-assisted self interviewing: Preliminary test of new technology for data collection. *Journal of Official Statistics.*

Orne, M. T. (1969). Demand characteristics and the concept of quasi-controls (pp. 143-179). In R. Rosenthal and R. L. Rosnow (Eds.), *Artifact in behavioral research.* New York: Academic Press.

Schwarz, N. (1993, November). *What respondents learn from questionnaires: The survey interview and the logic of conversation.* Third Annual Morris Hansen lecture presented to the Washington Statistical Society, Washington, DC.

Smith, A. F. (1991). Cognitive processes in long-term dietary recall. *Vital and Health Statistics,* Series 6, No. 4 (DHHS Publication No. PHS 92-1079). Washington, DC: U.S. Government Printing Office.

Willis, G. B., Royston, P. N. and Bercini, D. (1991). The use of verbal report methods in the applied cognitive laboratory. *Applied Cognitive Psychology, 5,* 251-267.

DISCUSSION

Roger Tourangeau
National Opinion Research Center

1. Introduction

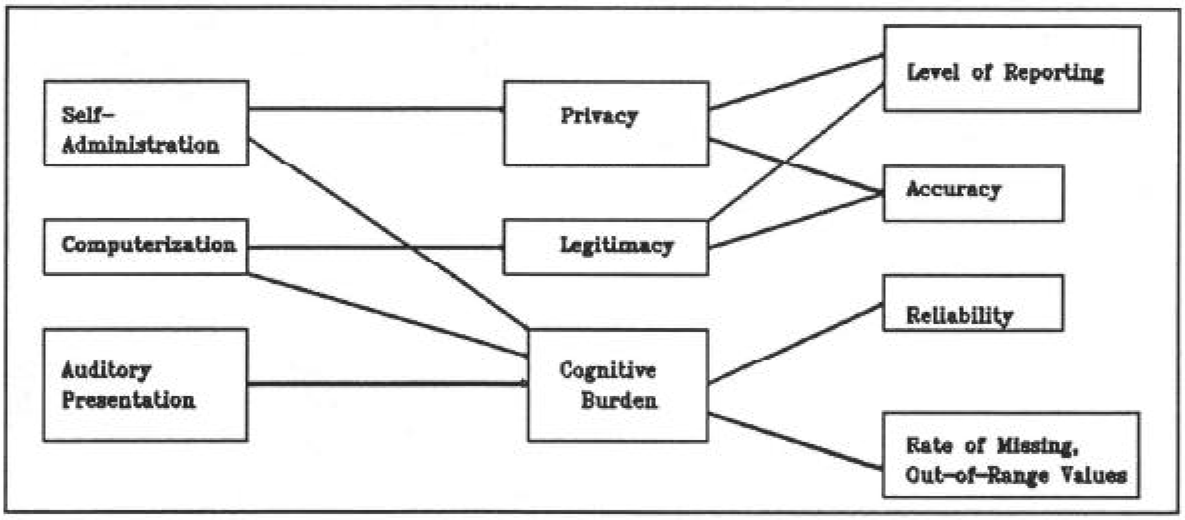
Both of the papers in this session concern the impact of the mode of data collection--in particular, the effects of self-administration--on data quality. The papers share a more specific concern with the difficulties respondents may have in reading survey questions and following skip patterns and other instructions for completing the questions. The two papers explore very different approaches to addressing these problems. The paper by Jenkins and Dillman describes some procedures for making it easier to read questionnaires; the paper by Lessler and O'Reilly discusses a method that eliminates the need for the respondent to read the questions entirely.

Although a good deal has been written about the impact of mode of interviewing on the results obtained (see, for example, Bradburn et al., 1991; deLeeuw and van der Zouwen, 1988; Groves and Kahn, 1979; Hochstim, 1967), there is no general model of the effects of the different methods of collecting survey data. The various popular methods of collecting survey data--in telephone or face-to-face interviews or in self-administered questionnaires--differ on several important dimensions; further, each of these basic procedures can be carried out on paper or using a computer. I suggest that the different modes of data collection vary on at least three key psychological dimensions:

> 1) **Cognitive demands.** Conventional paper-and-pencil questionnaires require either the respondent or the interviewer to read the questions and to follow the instructions; this requirement may sometimes exceed the reading abilities of respondents or interviewers.
>
> 2) **Level of privacy.** By eliminating the need for respondents to tell the interviewer their answers, self-administered questionnaires may reduce respondent concerns about the interviewer's reaction or about other family members overhearing sensitive information.
>
> 3) **Perceived importance of the study.** Laptop computers are still a novelty for most of the population, and the use of laptops in face-to-face data collection may enhance the perceived importance or objectivity of the study.

A general model incorporating all three dimensions is depicted in Figure 1. According to the model, features of the method of data collection (such as use of self-administered questions) affect the three psychological dimensions (e.g., level of privacy), which in turn affect data quality (willingness to report accurately about sensitive behaviors). Much of what is currently known about the different modes of data collection is captured in the model.

Figure 1. Path Model of Mode Effects



2. Paper by Lessler and O'Reilly

The paper by Lessler and O'Reilly reports results from a study that compared computer-assisted personal interviews (CAPI) with interviews in which the computer administered the questions directly to the respondent via earphones (audio computer-assisted self-interviewing, or audio-CASI). This study was done as a pretest for Cycle V of the National Survey of Family Growth (NSFG), and the pretest comparison naturally focusses on the two modes of data collection most likely to be used in that survey. Unfortunately, the pretest was not designed to separate out the effects of the several key advantages that audio-CASI offers relative to other modes of data collection (computerization, auditory presentation, self-administration). By comparing audio-CASI and CAPI, the pretest mainly examines the impact of self-administration rather than the other variables distinguished in Figure 1.

A fair number of studies have already shown that self-administration increases the level of reporting of sensitive behaviors. For example, Turner, Lessler, and DeVore (1992) demonstrated increases in reported drug use with self-administered questions, a finding replicated by Schober and colleagues (Schober et al., 1992). London and Williams (1990) find that more abortions are reported on a self-administered questionnaire than in a face-to-face interview (see also Mott, 1985). A study recently conducted by NCHS and NORC demonstrates increased reporting of sexual behaviors in self-administered questionnaires (see Table 1). That study compared face-to-face interviews and self-administered questionnaires in both a computer-assisted and conventional paper-and-pencil formats; as is apparent in Table 1, the effects of self-administration were larger and more consistent that those of computerization.

Table 1. Mode Effects in the Women's Health Study

Experimental Group	**Mean Reported Sexual Partners**		
	Past Year	Past 5 Years	Lifetime
Self-Administered Questions	1.72	3.88	6.54
Conventional (SAQ)	1.56	3.37	6.88
Computer-assisted (CASI)	1.89	4.40	6.25
Interviewer-Administered Q's	1.44	2.82	5.43
Conventional (PAPI)	1.56	2.86	4.58
Computer-assisted (CAPI)	1.36	2.79	6.27

Note: Each mean based on approximately 240 interviews; total rows for self- and interviewer-administration are based on approximately 500 completed interviews.

The results of the NSFG pretest on the differences in abortion reporting by mode are not very dramatic--audio-CASI increased the proportion of the sample cases who reported an abortion to 27.1% as compared to 23.8% when those same cases were interviewed via CAPI. Moreover, the audio-CASI abortions questions were different from those in the CAPI questionnaire, and they were administered after the respondents had already completed the CAPI interview. So the results from the study are perhaps better characterized as suggestive than definitive. In any case, they are certain to be useful to those charged with making practical decisions about the NSFG design.

Despite any weaknesses in the evidence regarding the advantages of audio-CASI, I predict that this technology will quickly be widely adopted by survey organizations. Having the capability, we will embrace the objective. The advantages of the new technology are, in some sense, too clear cut to require a lot of experimental confirmation--computerization virtually eliminates skip errors, self-administration minimizes privacy concerns, and auditory presentation eliminates the need for respondents to be literate. All in all, audio-CASI is a package that should prove irresistible.

3. Paper by Jenkins and Dillman

The paper by Jenkins and Dillman proposes 20 principles for improving the readability of self-administered questionnaires. The sensible tone of these recommendations reminded me of the admonitions in Tufte's *The Visual Display of Quantitative Information* and of the advice offered to writers in Strunk and White's *Elements of Style*. I did, however, wish that the authors had followed one additional principle; here is my proposed addition to the list:

Principle 21. Follow the Lord's example; never present more than ten commandments at any one time.

I am not advocating that Jenkins and Dillman abandon any of their principles! But I do think that they might try to formulate some larger principles from which their more specific guidelines follow. As I read their paper, it seemed to me that their recommendations reflected four key underlying axioms. First, the flow of a questionnaire should follow the natural reading order of the respondents. In English, this means questions should flow from left to right and from top to bottom. Second, questionnaires should use familiar, readily-understood graphical conventions. For instance, the same design element should always cue the same respondent action. Third, the questionnaire should call attention to the key information (via boldfacing and other methods). Finally, there should be a clear path for respondents to follow. Graphical features should emphasize this path. Table 2 groups 19 of the 20 principles discussed by Jenkins and Dillman under these four general themes.

I found almost all of their recommendations quite compelling. The one major exception involved matrix items--for example, questions that are asked for each family member or for each event of a given type. Jenkins and Dillman argue against giving respondents the choice on how to proceed through the matrix, and this may be the best way to ensure that they answer every question.

Table 2. Four Underlying Axioms

1) Where possible, take into account the natural reading sequence (left-to-right, top-to-bottom).

   --Include key information in the question, not after it (5)

   --Align questions and answers vertically (6)

   --Make top headings more prominent than those in the middle (10)

2) Use easily understood graphical conventions.

   --Use familiar formats (1)

   --Use same design feature to request the same action (13)

   --Avoid variability (14)

   --Use different layouts to distinguish different types of questions (16)

3) Call attention to the **key information** (15).

   --Present only the most relevant information (2)

   --Feature questions rather than explanations (4)

   --Put instructions where they are needed (7)

   --Put captions, units for answers where they will be seen (17)

4) Establish a clear path through the questionnaire (12).

   --Avoid multi-task questions (8)

   --Avoid matrix questions (9)

   --Use graphical instructions (such as arrows) to make the path salient (11)

   --Use graphical features to emphasize the path (18)

   --Avoid separating questions with lines and boxes (19)

   --Provide structures that make sense leave the R no choice! (20)

Note: Numbers correspond to those used by Jenkins and Dillman.

The counterargument is that there are sometimes good reasons for letting respondents follow their natural chain of associations in recalling specific incidents. For some respondents, it may be easiest to recall events that involve one person before recalling those involving the next person; for others, however, an event involving one family member may trigger the recall of similar events involving a different family member. As a result, there could be advantages to letting respondents follow whatever order the flow of memories seems to impose. Only further work can determine whether the advantages of imposing an order on the questions outweigh those of letting the respondents select the order they find most congenial.

The work that Jenkins and Dillman are doing is, in my view, quite important. Mail questionnaires are likely to remain a major method for collecting survey data for the foreseeable future. As this paper demonstrates, our current practice in developing these questionnaires often falls far short of the ideal. At NORC, instructions on self-administered questionnaires are sometimes put in boxes to distinguish them from the questions. During cognitive pretests, I have found that respondents often use the box as a cue for identifying material they need not bother to read. So, I agree with Jenkins and Dillman in thinking that bad graphical design can lead serious errors.

4. A Final Point

The method developed by Lessler and O'Reilly and the principles articulated by Jenkins and Dillman share an underlying goal--that of improving data quality, primarily by reducing missing data. One of the main advantages of audio-CASI over other methods of self-administration is that the software automatically computes which item the respondent is to answer next; this eliminates data that are missing due to incorrectly skipped items. Similarly, many of Jenkins and Dillman's principles stress methods to make sure that respondents answer all the applicable questions by making it easier for the respondents to figure which questions they are supposed to answer. Although audio-CASI has other noteworthy features and the principles proposed by Jenkins and Dillman will help address other response problems (such as questions that are misunderstood rather than missed entirely), a major objective of both approaches is the reduction of missing data. Valuable though this endeavor is, I cannot help but wonder whether this is the most pressing data quality issue that we face. The development of computer-assisted data collection methods has greatly increased the capital requirements of survey organizations; I sometimes wonder whether the gains in terms of data quality have provided a return commensurate with the investment. Perhaps it would make more sense to worry about whether respondents answer the questions accurately than to worry so much about whether they answer at all.

REFERENCES

Bradburn, N. M., Frankel, M., Hunt, E., Ingels, J., Schoua-Glusberg, A., Wojcik, M., and Pergamit, M. (1991), A comparison of computer-assisted personal interviews with personal interviews in the National Longitudinal Survey of Labor Market Behavior—Youth Cohort. In *Proceedings of the U.S. Bureau of the Census Annual Research Conference*, Washington, D.C.: U.S. Bureau of the Census.

deLeeuw, E.D. and van der Zouwen, J. (1988), Data quality in telephone and face to face surveys: A comparative meta-analysis. In R. Groves et al. (Eds.), *Telephone Survey Methodology*, New York: Wiley.

Groves, R.M. and Kahn, R.L. (1979), *Surveys by Telephone*, New York: Academic Press.

Hochstim, J. (1967), A critical comparison of three strategies of collecting data from households. *Journal of the American Statistical Association, 62*, 976-989.

London, K., & Williams, L. (1990, May), A comparison of abortion underreporting in an in-person interview and self-administered questionnaire. Paper presented at the Annual Meeting of the Population Association of America, Toronto, Canada.

Mott, F. (1985), *Evaluation of fertility data and preliminary analytic results from the 1983 survey of the National Longitudinal Surveys of Work Experience of Youth*. A report to the National Institute of Child Health and Development by the Center for Human Resources Research, January, 1985.

Schober, S., Caces, M., Pergamit, M., & Branden, L. (1992), Effects of mode of administration on reporting of drug use in the National Longitudinal Survey. In C. Turner, C., J. Lessler, & J. Gfroerer (Eds.), *Survey Measurement of Drug Use: Methodological Studies* (pp. 267-276). Rockville, Md.: National Institute on Drug Abuse.

Tufte, E. (1983), *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Turner, C., Lessler, J., and Devore, J. (1992). Effects of mode of administration and wording on reporting of drug use. In C. Turner, J. Lessler, & J. Gfroerer, (Eds.), *Survey Measurement of Drug Use: Methodological Studies* (Pp. 177-220). Rockville, Md.: National Institute on Drug Abuse.

Session 10

STATISTICAL USES OF ADMINISTRATIVE RECORDS

Improving Data Quality Through Increased Data Sharing

Edward A. Trott[1]
Bureau of Economic Analysis

In this time of major cutbacks throughout the federal government the Bureau of Economic Analysis (BEA), along with other statistical agencies, is being asked to improve the quality of statistical information and its availability. This is a big challenge. It will require a lot of thinking "outside the lines," followed by action.

To emphasize this point, I offer the following quote;

> "In a rapidly changing world, the best solution is not to keep redesigning the organizational chart; it is to melt the rigid boundaries between organizations. The federal government should organize work according to customers' needs and anticipated outcomes, not bureaucratic turf. It should learn from America's best-run companies, in which employees no longer work in separate, isolated divisions, but in project- or product- oriented teams.
>
> To do so, the government must make three changes. It must give federal workers greater decision making authority, allowing them to operate effectively in cross-cutting ventures. It must strip federal laws of prohibitions against such cooperation. And it must order agencies to reconsider their own regulations and tradition-bound thinking."[2]

One of the recommendations of the National Performance Review is the elimination of legislative barriers to the exchange of business data among federal statistical agencies. This recommendation is referenced as DOC11 and cites the Census Bureau, Bureau of Labor Statistics, and the Bureau of Economic Analysis as participating agencies. It further recommends the reduction of reporting burden on American business.

Today I will mention several areas in BEA's international, national, and regional accounts that utilize data from the Internal Revenue Service (IRS), Bureau of the Census (BOC), and Bureau of Labor Statistics (BLS)--where we are today--and then take a look at where we could be if more barriers to information exchange are removed.

---

[1]The author is the Assistant to the Chief, Regional Economic Analysis Division, Bureau of Economic Analysis, United States Department of Commerce. He is grateful for the contributions of Ralph Kozlow, International Investment Division; Brooks Robinson and Kenneth Petrick, National Income and Wealth Division; and Wallace Bailey, Regional Economic Measurement Division to the various parts of the paper.

[2]"*Creating a Government That Works Better and Costs Less*; *The Report of the National Performance Review*." Penguin Books, USA, 1993.

I.Current Uses

International Accounts

BEA and the Census are authorized to exchange certain confidential statistical data with one another for the purposes of augmenting and improving the quality of BEA's Title 22 (direct investment and international services) data and Census' Title 13 (economic census and related programs) data, under the Foreign Direct Investment and International Financial Data Improvements Act of 1990. Under the same act, the Bureau of Labor Statistics is authorized, for the purposes of augmenting and improving its data, to have access to selected information collected by BEA. Only data collected directly by BEA can be shared with Census and BLS; thus, for example, data BEA obtains from Census cannot, in turn, be shared with BLS. Similarly, only data collected directly by Census can be shared with BEA. Thus, BEA cannot have access to administrative records data obtained by Census from the IRS.

BEA and Census have made extensive use of their limited authorization to exchange data with one another and, as a result, have significantly augmented and improved the data collected in their business statistics programs. Several data exchanges have been for the purpose of obtaining detailed, establishment-level data on foreign direct investment in the United States. In 1992, BEA information on the identity of enterprises, or companies, that were foreign owned was linked to the Census' establishment-level data on the Standard Statistical Establishment List covering all U.S. establishments, to obtain detailed BOC data (State by 4-digit SIC) for 1987 on the number, employment, payroll, and shipments or sales of foreign-owned and all U.S. establishments. In 1993, BEA and Census released detailed data (mostly at the 3-digit SIC level by State) for 1989 and 1990, covering nearly all the data items on Census' Annual Survey of Manufactures (ASM) for foreign-owned and all U.S. manufacturing establishments. In 1994, BEA and Census will release detailed 1988 and 1991 ASM data for foreign-owned and all U.S. establishments. These exchanges have enabled statistical analyses that offer important new insights into the operations of foreign-owned companies and their impact on particular sectors of the U.S. economy.[3]

Other data exchanges between BEA and BOC have been for the purpose of supplementing the mailing lists for BEA's mandatory international surveys. BEA obtained confidential Census data on the identity of companies that are foreign owned, of companies that have foreign affiliates, and of companies that export services, in order to supplement mailing lists for BEA surveys of inward and outward direct investment and of international services transactions, respectively. The use of Census data to supplement the mailing lists for BEA surveys of international services transactions were particularly useful, because a significant number of potential new respondents were added

---

[3] See, for example, "Foreign Direct Investment in the United States: Establishment Data for 1987," in the October 1992 issue of the *Survey of Current Business*, and "Characteristics of Foreign-Owned U.S. Manufacturing Establishments," in the January 1994 issue of the *Survey*.

to BEA's mailing list as a result. (To further this purpose, questions on whether a given company or establishment sold, or exported, any services to foreign persons were included on a number of additional 1992 economic census surveys at BEA's request.)

In addition, BEA and the BOC are currently studying the feasibility of linking BEA's data on foreign-owned enterprises to Census' establishment-level product and product class data from the economic censuses, and product-by-country merchandise export data from its Exporter Data Base. The feasibility of other data link projects between the two Bureaus is also being discussed.

As mentioned above, in a parallel data link project, BEA's data on foreign direct investment in the United States have been linked to BLS data on all U.S. businesses. The initial results of that link, released in 1992 by BLS[4], provided data for 1989 and 1990 on the number, employment, and payroll of foreign-owned establishments. In October 1993, BLS released information on the occupational structure of foreign-owned manufacturing establishments for 1989[5].

BEA and the Department of the Treasury are also implementing limited data exchanges. BEA collects its data on direct investment and international services transactions, and Treasury collects its data on portfolio investment, under the same act (P.L. 94-473, as amended) and, in many cases, from the same U.S. companies. In order to ensure that companies are reporting correctly to both agencies, the Secretaries of Commerce and Treasury obtained OMB authorization to exchange and compare certain data. OMB approval for the data exchanges was obtained in January 1993, but BEA and Treasury are still working out the details necessary to implement the data exchanges. Note that any additional data exchanges between BEA and Treasury will require this same slow implementation process (beginning with letters from the Secretaries of Commerce and Treasury requesting OMB authorization, followed by negotiation between the agencies on the details for sharing the data.) It is hoped the time and resources required to implement data exchanges would be greatly reduced or eliminated if broad data sharing legislation were enacted.

National Accounts

BEA currently is accessing both noncorporate and corporate Statistics of Income (SOI) data via letters of agreement between IRS and BEA and the Secretary of Commerce and the IRS Commissioner as permitted by statutes. The

---

[4]United States Department of Labor. Bureau of Labor Statistics. "Employment and Wages in Foreign-Owned Businesses in the United States, 1989." Press Release, USDL 92-473 (July, 1992). and (*ibid*) "Employment and Wages in Foreign-Owned Businesses in the United States, Fourth Quarter 1990." Press Release, USDL 92-663 (October 1992).

[5]United States Department of Labor. Bureau of Labor Statistics. "New Research on Occupations in Foreign-Owned Manufacturing Establishments in the United States." Press Release, USDL 93-455 (October 1993).

Bureau's relationship with IRS has been in place for many years--dating back to 1944. At that time the Secretary of the Treasury, Henry Morganthau Jr., signed an executive order granting Commerce staff access to both noncorporate and corporate returns for statistical purposes. This executive order was superseded by statute in the early 1980's and access to SOI data from then until now has been limited in scope.

Following is a list of some of the specific topics that were recently researched via corporate return access--unless otherwise noted, adjustments to data resulting from this analysis improved the current year NIPA estimates of profits, interest, and dividends by adjusting extrapolating series to better conform to SOI definitions:

Identified discrepancies between SOI tabulations and trade association data in profits reported by life insurance companies for 1986-1987. Once identified, the SOI tabulations could be adjusted for changes in the tax law.

Identified differing reporting practices for commercial bank bad debt losses in SOI and FDIC tabulations and adjusted the current year FDIC series.

Reconciled pension and profit-sharing reporting practices between SOI and IRS Form 5500 preparatory to revising estimation methodology and data sources.

Separated business and interest receipts of credit agencies to more accurately measure interest flows.

Identified public utility joint partnership formations to ensure unduplicated reporting of new plants.

Analyzed captive finance companies' consolidation practices by parent corporations to ensure unduplicated reporting of income items.

Made adjustments to shift data for individual corporations to different industry classifications, which are more consistent with NIPA definitions and other data sources.

A secondary use of the individual company reports has been in the analysis of peculiar movements in SOI data. Because BEA has direct access to the IRS' corporate tax return information (not for the Foreign Direct Investment project), company by company panel comparisons have been compiled to review several questions regarding reporting practices, consolidation methods, or compilation errors.

Currently, actual use of noncorporate tax information has been limited to a study to determine the reporting patterns of partnership "pass-through" income reported on schedule K and by the individual partners on their Form 1040's. Results of the study led to a revision of the NIPA estimates of noncorporate partnership income and a redesign of the Form 1065 to explicitly determine the amounts of Schedule K income distributed to partners.

An additional area of data sharing between BEA and another federal agency--the Census Bureau--involves the use of the Quarterly Financial Reports (QFR) to adjust and analyze the quarterly NIPA estimates. Through a memorandum of agreement with the Census, BEA's sworn Census employees have continued to access individual company reports since 1982. Prior to 1982 BEA had access to the reports through an agreement with the Federal Trade Commission.

Primary use of this access is for the quarterly NIPA estimates. Each quarter BEA staff provide a list to Census of large non-recurring income or expense items which have been noted in a company's individual QFR. These items--mainly capital related and not associated with current production--are not treated as current income or expense items in the NIPA's. Census reviews the company report to determine if the questioned item was reported as a component of the indicator used to estimate pretax profits, or if it was excluded via several other reporting options. If the amount was included, Census identifies the exact pretax impact, which often differs from the amount in published after-tax profit reports, and the industry where the company is classified in their tabulations. The QFR also is used to augments the BEA company list with additional companies Census discovers during the review process of the quarterly reports.

In 1989 BEA entered into another cooperative arrangement with the Bureau of the Census; to explore the feasibility of using BOC data to prepare construction price indexes. In 1989, BEA staff--after being sworn in as Census agents--and the Census staff of the Construction Statistics Division (CSD) began to work on producing improved construction price indexes. The CSD provided micro-level construction data that were either collected through the Survey of Construction or the Value of New Construction Put in Place Survey. BEA staff used the data to test prospects for producing hedonic price indexes for residential and nonresidential buildings.

BEA staff produced a multifamily structures hedonic price index in 1991 for 1978-1989, and BEA began using it to prepare constant dollar estimates of the multifamily structures component of GDP at the time of the 1991 comprehensive revision of the NIPA's. Prior to this time, BEA used the BOC's Single-family Hedonic Price Index for Houses Under Construction. An article that describes the development of the index was published.[6]

Regional Accounts

For the preparation of BEA's estimates of personal income by State and county, the Regional Economic Measurement Division (REMD) receives several statistics tabulated by county from IRS Form 1040. The data are coded geographically and tabulated by the Census Bureau; the IRS, SOI Division provides these unpublished tabulations to BEA. The sample used by SOI for the Form 1040 statistics that they publish is too small to yield data for regional use. The

[6]See U.S. Department of Commerce, Bureau of Economic Analysis. "A Price Index for New Multifamily Housing." *Survey of Current Business*. Washington, DC: U.S. Government Printing Office, February 1993.

tabulations of dividends, taxable interest, and gross rents are the principal bases for REMD's estimates of the monetary components of property income. The tabulations of wages and salaries are vital for adjusting the components of personal income that are initially estimated on a place-of-work basis--wages and salaries, other labor income, and most of personal contributions for social insurance--to a place-of-residence basis. The tabulations of number of returns, number of exemptions, and adjusted gross income are used for REMD's quality control of the tabulations that are used directly in the estimates.

Proprietors' income and employment are important parts of the State and county estimates of personal income and employment prepared by REMD. Annual filings to the IRS by sole proprietors and partnerships are the only direct source of data for regional nonfarm proprietors' income and employment. The sample-based statistics from these filings that are prepared and published by the SOI Division are used for the calculation of national nonfarm proprietors' income. For the State and county estimates, REMD has acquired, under a contractual arrangement, individual records for sole proprietors from IRS Form 1040 Schedule C and for partnerships from IRS Form 1065. These records were then geographically and industrially coded, tabulated and combined to form the basis for REMD's State and county nonfarm proprietors' estimates. REMD-SOI data contracts have been in place for tax years 1981-83 and 1987-91.

II.Future Uses: How BEA Programs Would Be Improved Upon Implementation of DOC11

For illustrative purposes, several examples of how BEA's economic accounts can be improved with increased data sharing are given below. These are only a few of the examples that could be cited.

International Accounts

First, increased data sharing would improve the data obtained in the data link projects on foreign direct investment in the United States that BEA conducts with the BOC and BLS. As discussed earlier, BEA has shared its information on the identity of foreign-owned companies with both Census and BLS, to obtain those agencies' detailed data, by industry and State, for companies that are foreign owned. However, at present, Census and BLS cannot share their data with one another either directly or through BEA. For the data items released that were defined most similarly by both Census and BLS (number of operating establishments and employment), there are very large unexplained differences between the two agencies' data for foreign-owned establishments. For example, within manufacturing, the differences between Census and BLS data for 1990 are often at least 20 percent at the 2-digit SIC classification level; at the 3-digit and 4-digit SIC levels, the differences are often even larger. (Other industries may show even larger differences than manufacturing.) With such large unexplained differences, the results of the data link projects, and the conclusions based upon them, may be questioned. Reasons for differences may include different definitions, differences in industry coding, errors in the data, or that one agency was better able than the other agency to link its establishments to a given foreign-owned company. Because Census and BLS cannot share these data with one another currently, the exact reasons for the differences cannot now be determined. With increased data sharing, the reasons for the differences would be learned, and appropriate corrections to the data sets, if necessary, could be made.

Second, BEA's ability to analyze the data from the BEA-Census data link project would be improved if BEA had access to IRS data on Census' file of foreign-owned establishments. BEA employees are allowed to obtain limited administrative records data directly from the IRS, but they are not allowed to access the same data at the Census Bureau, because redisclosure of administrative record data to BEA is prohibited. (For 1987, IRS data account for about 15% of the estimate of employment for all foreign-owned establishments combined.) The Commerce Department must analyze and report to Congress on the employment, market share, value added, productivity, profitability, etc. of foreign-owned business enterprises compared to all U.S. business enterprises (P.L. 101-533 section 3(c)(1)). In order that BEA may efficiently perform these analyses, they need access to the full foreign direct investment (FDI) data file at Census. (The IRS has said it will support a regulatory change that will allow BEA access to the full FDI file, but this change may take considerable time to implement and will only facilitate this one project.)

Third, full data sharing would improve BEA's direct investment data if it permitted BEA access to complete IRS data on foreign-owned U.S. businesses and on U.S. businesses that have foreign affiliates. Both BEA and the IRS have released data on the net income, assets, sales, and number of foreign-owned U.S. companies and of U.S. companies that have foreign affiliates, and the data of the two agencies frequently differ substantially. Some of these differences might be eliminated if BEA were able to obtain information from the IRS for comparing with its own information on which companies are foreign owned and which companies own foreign affiliates. Also, appropriate corrections to BEA's data sets could be made, and BEA's sample frames could be improved.

National and Regional Accounts

First, for companies that own more than one establishment, Census and BLS both annually "map" all the individual establishments on their registers to the owning enterprises. Census obtains data needed for this purpose from the mandatory annual Report of Organization survey; BLS obtains the needed data from the Multiple Worksite Report (executed by each of the States and funded by BLS) and from commercial sources, such as Dun and Bradstreet. Under data sharing, only one agency would need to gather the information and share it with the other. This change would result in one set of company/establishment relationships that BEA could use to allocate data available only for companies to an establishment industry level.

Second, REMD currently makes intensive use of tabulations of ES-202 wage and employment data in estimating the wage-related components of personal income and employment for States and counties. Any improvement in the geographic or industrial classification of the ES-202 data that might result from ES-202/Standard Statistical Establishment List integration would increase the accuracy of BEA's regional estimates.

Third, full data sharing between Census and BLS would improve the derivation of the sampling frame for BLS' Producer Price Index. BLS uses the ES-202 file for its sample frame. The price indexes would be improved if the probability of inclusion in the sample were based on its shipments or sales, because it would allow for more precise weighting of the data that are collected. Producer Price Indices are used extensively in preparing constant dollar estimates for several expenditure components of GDP and for the industry distribution of GDP.

Fourth, with full data sharing, the industry coding used for calculating both the BLS producer price indexes and the Census shipments data could be fully consistent. At present, BEA must apply BLS' price indexes, which are based upon BLS' industry codes, to Census' shipments data, which are based on Census' industry codes, in calculating real GDP by industry.

Fifth, increased data sharing will improve BEA's I-O tables and estimates of gross State product and GDP by industry. In developing its detailed industry and State estimates, BEA utilizes Census data on establishment receipts and value added, and BLS data on employment and payroll. As discussed in connection with BEA's estimates of real GDP, the accuracy and usefulness of the I-O tables and of BEA's estimates of gross State product and GDP by industry would be considerably improved through consistent industry coding of establishments. These estimates would also be improved with full BEA access to unsuppressed Census Bureau estimates of value added, cost of materials, etc.; BEA now must indirectly estimate these items for suppressed Census Bureau data cells.

Lastly, REMD has identified additional statistics from the Individual Income Tax Returns (see under "Current uses") that would strengthen its current estimates and would help to extend its current estimates to prepare for the adoption of measures more consistent with the System of National Accounts (SNA).

The transfer of these data series to BEA do not require additional legislative action, only that the data be requested from the IRS. A relatively new and growing income entry on the individual tax form--interest from municipal bonds--will yield data to replace proxy information currently used to estimate the nontaxable interest component of personal income at both the State and county level. The tabulation of pensions and annuities received would be essential for the estimation of the corresponding components of the SNA aggregates.

III Conclusions

From the current uses section of this paper, it is apparent that even the limited amount of data sharing permitted under existing legislation has already produced a remarkable increase in the amount and quality of data and analyses, with no increase in respondent burden and minimal increased costs to the statistical agencies. The examples of future uses suggest improvements in data sharing that will benefit both the customers and staff of BEA as the recommendations of DOC11 are met.

It is apparent that significant duplication and inefficiency exist in business statistics data collection programs currently. It is particularly important that we continue our efforts to promote fuller data sharing in light of the dual challenges of limited agency resources and having to monitor developments in a rapidly changing world.

The opportunity to improve the accuracy of the income and product accounts produced by BEA is here. I believe that with the implementation of DOC11, estimates prepared not only at BEA but throughout the government will be improved as to their accuracy, timeliness, and will be less costly to produce.

Health Reform Information Systems: Great Expectations, Uncertain Prospects

Edward L. Hunter
National Center for Health Statistics

Introduction

In the opening keynote address, Graham Kalton described Roger Herriot (who organized this session and died suddenly this spring) as someone that had a vision beyond what we did every day, a person that encouraged us to look up from our work occasionally and check for shifting paradigms. Roger recognized that the types of information systems proposed for health reform may well represent such a paradigm shift. The purpose of this paper is to convey a sense of the great expectations that have accompanied discussion of health reform, the importance of these proposals for research and statistics, and to discuss some of the uncertainty over whether these expectations will be realized.

Health reform

The national debate over health care reform is reaching a crescendo this summer in Washington. More than a dozen congressional committees are weighing decisions that might - or might not - lead to wholesale changes in the U.S. health care system and, as a result, in our Nation's health information systems. The work of these committees is focused on piecing together legislation that can pass both Houses before members leave the Capitol to face voters in November.

While each health reform proposal may get there by a different route, most share the same common goals: improve health insurance security for the 38 million persons without health insurance, and for those that have insurance but risk losing it due to illness or unemployment; improve access to health care; control and even reduce health care costs, and even improve the health of the American people.

Among the questions yet to be resolved:

1) Will health reform result in universal coverage for health insurance, or will more incremental approaches be adopted?

2) If there are standard health insurance packages, how generous will the benefits be?

3) How will coverage for the uninsured be financed? Options include a mandate that employers provide coverage, a mandate that individuals buy insurance, or broad-based tax increases.

4) To what extent will the Government be involved in regulating activity in the health care system?

Major proposals still under consideration include modifications to the President's proposed Health Security Act and more modest or market-oriented alternatives proposed by moderate Democrats and Republicans. At this point in the debate, it is not yet clear what type of reform - if any - will be enacted.

Role of information in the current health reform debate

It does seem clear, however, that issues of health reform will continue to dominate our agenda for the coming years regardless of what happens this session of the Congress. And it is clear that information - or the lack thereof - has become central to the emerging debate over health reform. Questions, such as who lacks health insurance and why; how and why do businesses cover employees; how is insurance related to use of services, and to health; what drives health costs up; and what will expanded benefits cost, have become central to crafting a legislative package. Yet, these are questions for which our current information systems - both survey and administrative - can provide only partial answers.

Importance of information in a reformed health system

Nearly all advocates of health reform visualize a new health care system in which information plays an increased role. Since reform will almost certainly be implemented at the State level, State governments will need better information to tailor their approaches. Health plans will need better data to select member institutions and facilities, and manage the quality and cost of patient care. Consumers will need far greater information on quality of services, satisfaction, and outcomes in order to make informed choices between competing plans. Further down the road, States will need to assure that health plans are meeting standards - and the Federal Government will need to assess the expected and unexpected impacts of reform. Finally, there will almost certainly be a debate over what went right - or wrong - that can only be conclusive if there is adequate information.

There is a growing, and encouraging, consensus among the sponsors of competing health reform proposals regarding several key information provisions. Most important among these are the need for administrative simplification - digging health providers and payers out from under the ever-growing pile of forms, and the need for greater uniformity in the way health transactions are recorded and reported. Nearly all parties to the debate recognize that we need to hold providers and health plans more accountable, that we need better information on health outcomes to guide practice, and that consumers will play a greater role as information users.

The Health Security Act: a national framework for health information

President Clinton's proposed "Health Security Act"[1] includes the most ambitious - and controversial - information proposals of any plan currently being considered. The balance of this paper will outline these proposals, and address some of the implications if current legislative deliberations result in enactment of provisions resembling these proposals.

The architects of the Health Security Act regard the availability of health information as essential to the success of managed competition, the basis for the Act. Toward that end, the Act calls for a national framework under which standardized, nationwide information would be collected for all patients, providers, and encounters with the health care system. This system would be built on a consensus around core, minimum data sets that can be used by multiple parties; standardization and simplification of currently burdensome systems of administrative and payment records; unique identification of individuals and providers to facilitate linkages; national legislation to protect the privacy of records in the system; and regional data centers to process the records and provide for access for research and statistical purposes.

Information on individuals enrolled in health insurance plans

There are three major elements in this design: enrollment, encounter, and administrative data. In the first, information would be obtained on each individual eligible for health insurance when they are enrolled in a health plan. At a minimum, this information would include identifying information (including name and address); a unique identification number; and additional demographic information such as age, sex, and (hopefully) race/ethnicity. Information on the individual's source of coverage would be included, as would any other information deemed necessary to adjust premium payments to health plans based on the individual's risk. It is possible that this list could eventually expand to include information on the individual's health status and medical history.

Data on encounters with the health care system

In the second element of the framework, a standard, minimum data set will be collected at the point of each encounter with the health care system. This data would begin with that now provided on claims forms, such as reason for visit, diagnoses, procedures, prescriptions, followup, and disposition. But since care will increasingly be delivered by prepaid systems that do not require individual claims, information would also be obtained for all such encounters. Demographic data will already have been obtained for individuals through the enrollment system, so it would not be necessary to collect such information at each encounter.

---

[1] HR 3600, "The Health Security Act of 1994", Title V.

Debate over implementing this encounter data system will focus on the length and content of the minimum data set, the extent to which an encounter might be used to obtain risk factor or other information not immediately relevant to the clinical encounter, and the types of coding and classification systems that will be used to translate recorded information into usable statistics. To make this information of maximum use for research purposes, encounter data will need to go beyond what we currently collect through the payment system, and be of better quality.

Data on plans and providers

Third, in order to facilitate the operation of the health care system, including payment of claims, certification of plans, etc., administrative data on the characteristics and operations of health plans and providers will be included in the system. This may range from the multiple affiliations that plans and providers will have with each other, to the types of services provided and patients served.

Privacy and data access

To summarize, the information framework envisioned in the Health Security Act provides for demographic information on all individuals eligible for health insurance, information on the characteristics of all health providers, and a small set of information on each encounter an individual has with the health care system.

Such a system has enormous potential - the potential to meet a wide range of research needs, and the potential - if badly designed or implemented - to jeopardize privacy protection that Americans have come to expect.

As a result, privacy protection is being given a great deal of attention by those currently debating the Health Security Act. The Administration is committed to the enactment of strong confidentiality protection,[2] and there are serious efforts being made in the Congress to craft comprehensive privacy legislation for medical records.[3]

The privacy debate over the health reform information network is critical to the eventual ability of researchers to use data that may be part of the network, for a variety of reasons. First, without addressing the public's concerns through enactment of strong protection, the Congress may not be willing to create such a network in the first place. Second, if the

---

[2] Testimony of Nan D. Hunter, Deputy General Counsel, DHHS, before the Subcommittee on Information, Justice, Transportation, and Agriculture, Committee on Government Operations, U.S. House of Representatives, April 20, 1994

[3] HR 4077, "Fair Health Information Practices Act of 1994," introduced by Representative Gary Condit (D-CA), is one example of pending legislation.

public doesn't have confidence in the privacy of their medical records, it will be impossible to assure the quality and accuracy of the information they provide to the health care system. Third, in an effort to reassure the public on privacy, it is possible that the Congress will over-react and place restrictions on access to data that will unnecessarily restrict research and statistical uses. And fourth, even with the best of intentions, it is a tough job to craft legislation that balances privacy with access to data, and we must make efforts to ensure that there are no inadvertent limits are placed on our ability to use these sources.

Statistical uses of *ADMINISTRATIVE* records?

The Federal statistical community argues that information produced through any system, including purely administrative data systems, should serve multiple purposes, and be put into a form that is conducive to research and compilation of statistics. It is an encouraging development that the Administration and the Congress are giving serious consideration to ways in which the health reform information framework can be put to maximum public use.

In fact, it is not at all clear that this system should be considered an *ADMINISTRATIVE* system - since research and statistics are among the most important uses the system is being designed to meet.

Federal statistical agencies have a strong history of making data available to the public, while protecting privacy, by producing anonymous, person level public use files. It is encouraging that this history is guiding policy development for access to data in the health reform network. As envisioned by the authors of the Health Security Act, regional data centers will be responsible for creating general-purpose files, as well as responding to individual researchers with special requirements, while being given special privacy authorities to assure the protection of individual records.

Potential uses of data from the health reform network

Public health provides a number of useful examples of how this network might serve research needs. The network will be broad and comprehensive, but will lack detailed diagnostic and clinical information typically included in patient charts. In effect, it will be a "mile wide and an inch deep." But this "mile" of data can fulfill many important public health needs, including ones we have great difficulty meeting today.

It can provide a broad picture of the population, and reasonably complete coverage of the delivery of health services. The existence of a population base for these records also provides greater opportunity to link medical events to denominators, and develop better incidence/prevalence data on certain diseases. It allows us to address subgroups of the population with greater confidence.

Examples of health applications

One specific example is the tracking of progress in improving the immunization of preschool children, a national public health priority on which we will spend hundreds of millions of dollars in the coming years. From enrollment files, we will know the ages of children that should be receiving vaccines - and will be able to use encounter data to determine immunization rates by health plans and local areas. Likewise, we will be able to monitor our success in promoting other preventive services, such as encouraging women of certain age groups to have a mammogram, and men of appropriate ages to have cholesterol screening. With this information, we will be able to better target intervention programs to prevent disease.

If a major health reform is implemented, we will also experience significant change in nearly every aspect of health care and, potentially, the health of the public. The data in this network will help us to monitor health status and outcomes, produce routine measures of quality of care, and assess changes in the organization, financing, and delivery of health services.

Potential uses in other areas

Since the network will include information on a broad population basis through the enrollment records, it may prove to be a valuable tool for demographic research. These files will, at a minimum, include a variety of demographic and geographic items, may provide us a picture of the family status of individuals, and may even record changes in individual circumstances over time.

Since enrollment for health insurance will be primarily employment-based, the network will include considerable information about the employment status and occupation of covered individuals and their family members. This may prove to have applications for labor market and employment-related research.

And, the network will include detailed information on health care providers, including facilities, professionals, and other organizations. This will provide the basis for research on the establishments and workforce of a large segment of the economy.

Potential uses for Census

The broad population coverage anticipated for the enrollment file has obvious potential uses in conducting censuses of populations. Congressional testimony by the Administration reflects current thinking that <u>some</u> access to identifiable information in the network by statistical agencies may be appropriate.[4] As the Bureau of the Census looks toward the use of other administrative sources of data for population censuses, the health insurance enrollment file may prove to be among the most complete source of demographic information of the population. This is said, of course, with full recognition of the limitations of this system, which will be discussed briefly below.

It is possible that some provisions allowing for access to identifiable records for such purposes may be built into the health security act as it is considered by the Congress. However, it is also likely that these sorts of uses will be carefully specified to ensure that the public is not left with the perception that there will be widespread sharing of identifiable records for non-health related purposes. It is also important to remember that the Act was written with privacy protection in mind, and limits use of identifiable files to health applications except in limited circumstances.

Applications for survey research

To this point, this review has not focused on the limitations of the proposed health reform information system, but clearly, there are many. For example, it is likely that even a universal entitlement to insurance coverage will not entice all eligible individuals to participate - leaving enrollment files short of complete population coverage. We know that certain population groups are likely to be underrepresented, or excluded. We know from other administrative data sets that we have reason to suspect the quality of reporting of certain types of data.

Finally, since the framework relies principally on the reporting of encounters with the health care system, it will lack information on events, conditions, and other health issues of interest that occur outside the health care delivery system.

For the purposes of this paper, however, as we speculate on what this system might do for research and statistics, it is assumed that we can devise methods to compensate or adjust for these limitations. For example, it will be critical to augment the network with surveys and other data collection mechanisms to fill critical data gaps and provide quality checks on administrative reporting.

The information framework, then, will both rely on data collected in surveys, and can at the same time facilitate the conduct of surveys.

---

[4] Testimony of Nan D. Hunter, Deputy General Counsel, DHHS, before the Subcommittee on Census, Statistics, and Postal Personnel, Committee on Post Office and Civil Service, U.S. House of Representatives, March 16, 1994

Surveys required for health reform

At the National Center for Health Statistics, and elsewhere in the Federal statistical establishment, we also face an enormous task in developing surveys to provide information for implementing and evaluating health reform, and we have reason to hope that new systems will facilitate this work. These surveys will be essential to augment information available through the network, since no single data source will be able to provide the depth of data needed for the types of analyses we will need to be performing. We are already beginning to address some of these needs with development of new surveys, such as the National Employer Health Insurance Survey.

Sampling frames

The first, and most significant benefit of the new system will be in the construction of sampling frames. Many current surveys rely on sampling approaches that are time consuming and inefficient. The three components of the network will facilitate sampling as follows:

1) **Population sampling from the enrollment data base,** allowing for selection of sample individuals by demographic characteristics such as age, race, sex, geographic location, occupation, and possibly other proxies. Sampling and screening for rare and non-clustered population groups, currently difficult and expensive, would be greatly facilitated. Similarly, it would be possible to quickly identify and survey selected population groups, allowing for quicker turnaround for topics of current policy or research interest.

2) **Provider sampling from administrative records.** Many current health provider surveys (e.g., hospitals, physicians, nursing homes) construct sampling frames from lists provided by professional associations, phone directories, licensing agencies, or private marketing firms. These lists are subject to considerable error and possible bias, and are costly and time consuming to create. Sampling from complete and accurate lists of providers, with detailed characteristics already included in administrative files, would make provider surveys faster, less costly, and of greater quality.

3) **Sampling based on conditions, diagnoses, or procedures** that can be identified through encounter records. For example, cohorts of individuals that received specific medical treatments can be identified for followup or interviews to assess outcomes. Similarly, persons with certain diagnoses or conditions could be identified to conduct studies of access to appropriate medical care; persons with only limited use of the health care system could be identified and included in studies of barriers to access to care.

In each of these examples, it is clear that extensive research will be required to assess the extent to which sampling frames created entirely from administrative records systems will adequately represent the universe, and approaches will need to be developed to augment these frames. It is also clear that there will continue to be an important role for traditional sampling approaches where administrative frames are not adequate.

Other potential uses - record linking

Access to enrollment and encounter data in the network can also facilitate efficient linkage of detailed health status, risk, and behavior information that we will obtain through population-based surveys to the utilization and outcomes data included in encounter files.

These files can facilitate longitudinal studies, in which individuals could be enrolled in cohorts according to characteristics in the enrollment files (e.g., occupation), or based on encounters. Followup of theses cohorts could then be conducted through surveys and through analysis of encounter data.

Conclusion: prospects

To conclude, the information system described in the health security act, and touched upon in other health reform proposals, holds enormous potential for those of us interested in research and statistics. Data from enrollment, encounter, and provider data systems will augment or replace existing approaches, and provide survey researchers with new tools for sampling, conducting longitudinal followup studies, and linking survey data with outcomes. Along with a variety of new survey approaches, this information system provides us with the opportunity to fill many long-standing gaps in our understanding of the health care system and public health, and there are a variety of potential applications beyond health.

Getting there from here requires that at least four difficult questions be answered in the affirmative: 1) Can Congress actually enact a broad reform that will change the health care system? 2) In enacting any information requirements, can an appropriate balance between privacy and access to data be found that will allow the system to be useful? 3) Can all the affected interests come together to agree on standards and minimum data sets that include items of use for statistical purposes? and finally, 4) Can we develop and manage the technology for handling and protecting the volume of records that will result, and turn these records into usable statistics?

Improvements to Economic and Health Statistics: Discussion

Miron L. Straf[1]
Committee on National Statistics
National Academy of Sciences--National Research Council

It is a privilege and pleasure to discuss the two papers by Edward A. Trott, Jr., and Edward L. Hunter. Both papers are excellent and exceptionally well prepared. As such they tributes to the memory of Roger Herriot, to which we have dedicated this session. On behalf of all of us at this conference and the many others who knew or were influenced by Roger, I thank our speakers for their quality presentations.

Discussion of "Improving Data Quality Through Increased Data Sharing: The National Performance Review (NPR) Initiative," by Edward A. Trott, Jr.

Ted Trott is being modest. The data sought are even more important than one might glean from his paper. As our economy becomes an increasingly a global one, our current system of economic statistics becomes woefully out of date. Let me give you one example from the report of one of our Committee on National Statistics panels, *Behind the Numbers: U.S. Trade in the World Economy* (National Academy Press, 1992). The traditional balance of payments framework classifies transactions by geographical boundaries, so that when foreign affiliates of U.S. firms sell to other foreign firms, no effect is recorded in our trade balance (See Figure 1-1). By supplementing reports on exports and imports across geographic boundaries with information on other international business activities, the panel estimated that the difference in what the U.S. bought from and sold to foreigners in 1987 was $64 billion, less than half the reported $148 billion trade deficit for that year.

Trott shows how sharing data can improve efficiency and accuracy, but the benefits are more. In most of the cases that he mentions, more data are or would be created than the sum of what we have in separate agencies.

Sharing data also promotes new research and serves to test new theories and methods. It helps us improve our models and better understand sources of error. It furthers the use of empirical studies and other analyses in public policy formulation and evaluation. And it respects the respondents who provided the data by seeing that their information benefits society in the most effective way.

---

[1]The remarks herein are those of the author and not necessarily those of the Committee on National Statistics nor of the National Academy of Sciences--National Research Council.

Trott's paper is an excellent documentation of costly problems from not sharing and the benefits from sharing data. It also provides some innovative approaches to the problems of missing data. One example raises a question asked in the session with George Duncan and Nancy Kirkendall on confidentiality and disclosure limitation. The Census Bureau can't provide detailed micro data to the Bureau of Economic Analysis (BEA) and so suppresses some cells in the tables provided. BEA indirectly estimates these missing data. Suppose it does so through a model based on some assumptions. The Census Bureau has the original data and can determine whether the model holds or the assumptions are wrong. Should it inform the BEA?

My friends who are into Transcendental Meditation provide some important advice for missing data. Once, when I was delivering a lecture on multiple imputation methods for nonresponse, one of them intonated to the audience that the methods were unnecessary, because, if you search really hard, you can find the missing data within yourselves.

And, to me, that is Trott's message to the federal statistical system. Only, as he points out, you really don't have to search very hard.

Thus, we are led to ask if what we need is really going to be achieved by a data sharing agreement in fulfillment of recommendation 11 to the Department of Commerce from the National Performance Review. Perhaps, but not likely in our lifetime.

What is needed is a new framework to provide the incentives for finding the ways for data sharing to occur while fulfilling the intent of confidentiality law. the Committee on National Statistics explored one approach recently with a number of statistical agency heads. The basic issue is that there needs to be a quid pro quo for one party to share information with another. Truly cooperative agreements are rarely built on one-way streets. But what can be offered in return for data provided? One approach is to return analyses that depend on the data. But more can be offered.

A statistical agency can engage in designing, developing, and even managing a data collection program of another agency, so that it can obtain important data for statistical purposes. What is more the data program of the other agency could be for administrative purposes. Thus, BEA or the Census Bureau could design for the Treasury a tax data base for tax policy or even for enforcement purposes that could also provide statistical data. It's not far fetched. As David Binder has reminded us here, Statistics Canada already does it.

Discussion of "Health Reform Information Systems: Great Expectations, Uncertain Prospects," by Edward L. Hunter.

Hunter's paper is a consummate exposition of one of the greatest challenges to the federal statistical system: the information system that may arise in a reformed health care system. Ed lays out, in as much detail as has been developed, the data to be collected and the potential uses of them. I want to complement his description with another perspective that raises some issues for the federal statistical system.

The structure of the data is in three parts: enrollment, encounter, and claims. In addition, data would be compiled on health care providers. Consumer information would be issued as some sort of report card on plans and providers. Here measures of the quality of health care services become important.

The regional data centers may be entirely new public-private ventures. They may be the places where public-use files are compiled and where linkages to other data, such as surveys, are performed. Here privacy and confidentiality are crucial.

The health care information system envisioned can provide many benefits. The information is needed for physicians to make accurate diagnoses and recommend appropriate treatment. But the information is also needed to monitor trends that affect the costs of health care, to plan for the changes in the needs for health care, and to achieve a better understanding of how behavioral and social factors are related to health and health care coverage.

It is important to note however, that, although health care information alone may provide information on health outcomes, it is not possible to understand what may have caused the outcomes unless data can be combined with other data, such as from surveys. The information serves other purposes too, including being a valuable source of social, economic, and demographic characteristics of our population. However important these purposes are, however, the health care information system would be driven by the needs to implement the health care system and not by research and statistical needs.

So that the public might benefit from access to the information for research and statistical purposes, the Committee on National Statistics expressed concerns to the Congress about legislative provisions that were proposed. As Hunter points out, some of these provisions could preclude data for important research and statistical purposes.

The Committee's first and foremost concern is that privacy and confidentiality of health care information be adequately protected. It is not necessary to sacrifice either confidentiality or the benefits of information: both are possible if legislation provides for responsible access and demonstrated, effective means to protect confidentiality. The Committee also argued that health care legislation can protect confidentiality of information and yet permit important research and statistical uses of that information by

- Prohibiting data about an individual that are collected or maintained for research and other statistical uses from being used in any administrative or enforcement action affecting that individual. This principle is referred to as *functional separation*.

- Extending confidentiality protection to identifiable data about individuals, wherever the data are maintained.

- Providing sanctions against unauthorized disclosures by any user.

- Authorizing access to health care data about individuals for research and statistical purposes whenever confidentiality can be assured.

- Creating an independent federal advisory body charged with fostering a climate of enhanced protection for all federal data about persons and responsible data dissemination for research and statistical purposes.

The Committee conveyed these points in letters to members of Congress working on health care legislation. The result is that bills were modified to permit access to health care information for public health research and for research on behavioral and social factors affecting health. Without these changes, the viability of several of our major national surveys would be threatened.

Many difficult problems remain, however. Who, for example, rules on access? One proposal is that Institutional Review Boards grant permission for access. Some parties, such as the Institute of Medicine, would not even permit access with consent. The concern is that, if access were allowed with consent, then employers might require it as a condition of employment.

Hunter talks about the system enhancing the legally mandated reporting to local and state health departments. Not everyone wants to facilitate providing this information, however, because of confidentiality concerns.

Hunter also shows the many benefits of the data in and of themselves. We must take care however, that people do not get the impression that a single, large anonymous file can serve most needs. Such a public-use tape might do so for public health purposes, but many policy purposes require different data to be combined or data to be combined in other ways. And public use tapes cannot be combined with further data that may be needed, such as information from Social Security earnings records.

The examples Hunter gives of uses for demographic research, for labor market and employment-related research, and for providing information on health care providers shows the utility of the data to other agencies: the Census Bureau, the Bureau of Labor Statistics, and the National Center for Health Statistics. What are these agencies doing to assure access to the data? I am not aware of any major initiative.

Many serious confidentiality problems remain. For example, is it appropriate to screen the system to develop a sampling frame of people with a disease like AIDS? Are we going to allow them to be contacted? How? Should we permit contact by a call or letter that might go to a member of their household?

A specific confidentiality practice that can be harmful is when an agency specifies that information will not be used for purposes other than for which it was collected . That is a bad principle, and some bills before Congress have adopted it. Such a blanket prohibition denies many legitimate research and statistical uses. No one can foresee all potential uses of data that would benefit society. If such prohibitions were enacted and enforced, society would lack important information that it could obtain only at greater cost through new data collection that might further intrude on individual privacy.

Despite Hunter's expectations, the administration may have missed the boat on privacy and confidentiality provisions. Bills are moving now quickly through the Congress, and Representative Condit's bill can stand on its own with or without a health reform bill.

Nevertheless, the challenges for implementing such an ambitious data system are before us, and the statistical agencies can offer a special expertise here. We can look toward a health care information system to allow through research and statistics the means of providing the information required by stakeholders represented by the six P's: policy makers, public health officials, payers, providers, patients, and the public.

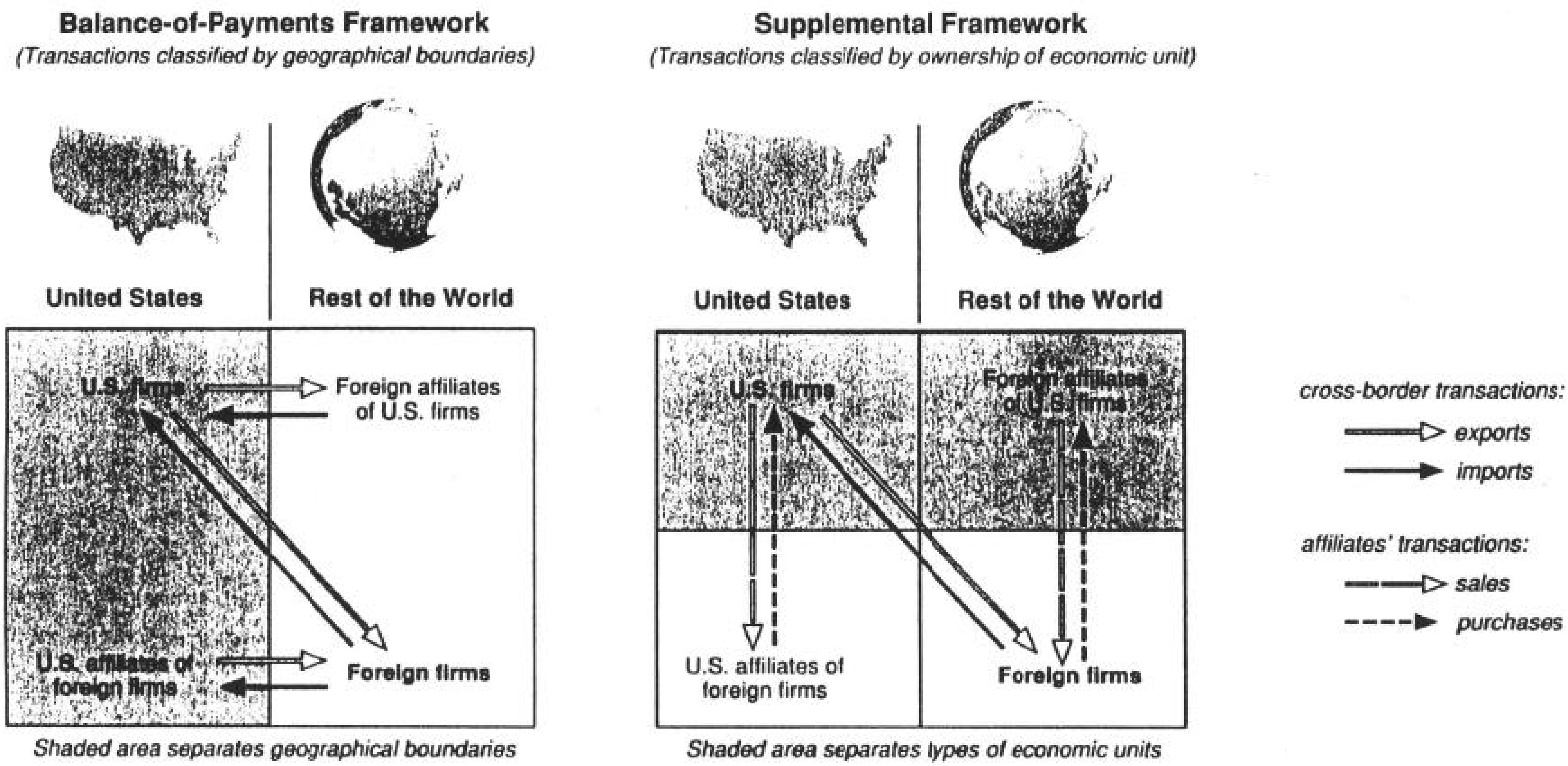


FIGURE 1-1 The balance-of-payments framework and a proposed supplemental framework.

Source: Kester, Anne Y. (1992). *Behind the Numbers: U.S. Trade in the World Economy*, Report of the Panel on Foreign Trade Statistics, Committee on National Statistics, National Academy of Sciences--National Research Council. Washington, D.C.: National Academy Press.

Session 11

SMALL AREA ESTIMATION

Small Area Estimation for the National Health Interview Survey Using Hierarchical Models

DONALD MALEC
National Center for Health Statistics; Hyattsville, MD

J. SEDRANSK
State University of New York; Albany, NY

1. Introduction

There is a continuing need to assess health status, practices and resources at both the national level and subnational levels. Estimates of these health items help determine the demand for quality health care and the access individuals have to it. Although NCHS survey data systems can provide much of this information at the national level, little can be provided directly at the subnational level, except for a few large states and metropolitan areas. The need for State and substate health statistics exists, however, because health and health care characteristics are known to vary geographically. Also, health care planning often takes place at the state and county level.

Using a hierarchical model, our focus is on the development of state estimators using data from the National Health Interview Survey (NHIS). Information on health status, practices and resources is collected annually in the NHIS and direct national estimates of these items are also produced annually. The NHIS is a multistage, personal interview sample survey. The current sample design uses 1,983 primary sampling units (PSU's), each PSU consisting of a single county or a group of contiguous counties (minor civil divisions are used instead of counties in New England and Hawaii). The population of 1,983 PSU's is stratified and approximately 200 are sampled with probability roughly proportional to their population sizes. Within each sampled PSU clusters of households are formed and sampled. Areas within a PSU with a high concentration of blacks are oversampled. The NHIS is a cross-sectional survey: each year, a new sample containing approximately 50,000 households and 120,000 individuals is selected. (For additional details about the design of the NHIS see Massey et al. 1989.) Although the total sample size is large, the sample size in most states is too small to produce direct estimates that are sufficiently precise.

Malec and Sedransk (1985) have described Bayesian methodology appropriate for the analysis of some multi-stage sample surveys when the variables are normally distributed. We have extended this methodology to accommodate binary random variables, the predominant variables in the NHIS. Our model is similar to that of Wong and Mason (1985). However, the objective in Wong and Mason (1985) is inference about parameters in the model rather than finite population quantities. While Dempster and Tomberlin (1980) investigate small area estimation methods for binary random variables they, like Wong and Mason, provide an empirical Bayes rather than a fully Bayes solution. Since empirical Bayes procedures often account for only a fraction of the error correctly represented in a fully Bayes approach, we prefer the latter. Recent advances in numerical methods (e.g., the Gibbs sampler) permit the employment of a full Bayesian analysis; see, e.g., Gatsonis, et al. (1993), Malec and Sedransk (1993a), and Malec, Sedransk, and Tompkins (1993).

The notation and model are described in Section 2 while the estimation methodology is presented in Section 3. Section 4 describes the techniques for fitting the proposed models, and displays the final model using data from the 1987 NHIS on utilization of physician care. There is a comparison of alternative estimators in Section 5, and evaluation of the proposed methodology is described in Section 6.

2. Model Specification

The model in (2.1), (2.2) and (2.3) below includes the most important features of the sample design. Our objective is to produce accurate point estimates and appropriate measures of variability by accounting for geographic variability of the response and using available covariate information.

Let Y_{ij} denote a binary random variable for individual j in county i where $i=1,\ldots, L$ and $j=1,\ldots,N_i$. Within county i and conditional on the p_{ij}, the Y_{ij} are assumed to be independent Bernoulli random variables; i.e.,

$$Pr(Y_{ij}=y_{ij}) = p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}, y_{ij}\in\{0,1\}. \qquad (2.1)$$

Given the vector of M covariates corresponding to individual j, $X_{ij}^t = (X_{ij1}, \ldots, X_{ijM})$, and $\underline{\beta}_i$, it is assumed that

$$\ln\{p_{ij}/(1-p_{ij})\} = X_{ij}^t\underline{\beta}_i. \qquad (2.2)$$

To allow for the possibility of a linear regression between each element of $\underline{\beta}_i$ and a set of covariates, $\underline{Z}_i^t=(Z_{i1},\ldots, Z_{ic})$, available at the county level, assume

$$\underline{\beta}_i \sim N(G_i\underline{\eta},\Gamma) \qquad (2.3)$$

where, conditional on $\underline{\eta}$ and Γ, the $\underline{\beta}_i$'s are independent and $G_i=\mathrm{Diag}(\underline{g}_{i1}^t, \underline{g}_{i2}^t, \ldots, \underline{g}_{iM}^t)$ and $\underline{g}_{ij}^t$ is a row vector of dimension c_j containing a subset of covariates from $\underline{Z}_i^t$. Additionally, $\eta^t= (\eta_{11}, \ldots, \eta_{1c_1}, \eta_{21}, \ldots, \eta_{2c_2}, \ldots, \eta_{M1}, \ldots, \eta_{Mc_M})$, conforming to the dimensions of G_i, and Γ is an MxM positive definite matrix. Finally, reference prior distributions are assigned to $\underline{\eta}$ and Γ; i.e.,

$$p(\underline{\eta},\Gamma) \propto \text{constant}. \qquad (2.4)$$

Taking $\Gamma=0$ provides a specification that is consistent with the basic assumption of synthetic estimation. In the following "synthetic estimation" refers to the use of (2.1), (2.2) with $\underline{\beta}_i=G_i\underline{\eta}$, and $p(\underline{\eta}) \propto$ constant.

We include as variables in (2.2) those individual level characteristics that provide the best prediction for p_{ij}, and are reliably estimated at the county level in non-census years. Candidate variables at the county level (i.e., G_i in (2.3)) include the variables used to define the NHIS strata (defined at the county level). By predicting for demographic groups within counties and then weighting by postcensal population estimates, estimates are automatically weighted to fixed population totals.

3. Estimation Methodology

3.1 Bayesian Predictive Inference

In this paper, our objective is to make inference about finite population means. By first summing the Y_{ij}'s within a county and then within a state, the population mean within a state can be expressed as

$$\theta = \sum_i \sum_{j=1}^{N_i} Y_{ij}/N. \tag{3.1}$$

Formula (3.1) can represent either a mean for the entire state or for a subpopulation. The first sum is over the collection of counties within the state, while N_i is the size of the population or subpopulation in county i. Here, $\Sigma_i N_i=N$.

In (2.2) we use the variables age, sex and race because these are the only variables for which reliable estimates are available at the county level for non-Census years. In this case, (3.1) can be simplified. Suppose that in the population there are K different values of the vector $\underline{X}_{ij}$. Then write $\underline{X}_{ij} = \underline{X}(k)$ for all ij having pattern k (k=1,...K).

From (3.1),

$$\theta=\Sigma_{(i,j)\in s}\frac{y_{ij}}{N}+\sum_i\sum_{k=1}^{K}\left(\frac{N_i(k)-n_i(k)}{N}\right)\overline{Y}_{ik}^{(ns)} \tag{3.2}$$

where $N_i(k)$ and $n_i(k)$ are, respectively, the population and sample sizes in county i with $\underline{X}_{ij} = \underline{X}(k)$ and $\overline{Y}_{ik}^{(ns)}$ is the mean of the nonsampled individuals with demographic characteristic k in county i.

Letting $\underline{y}_s$ denote the vector of sample observations, we emphasize the first two moments of θ, $E(\theta|\underline{y}_s)$ and $Var(\theta|\underline{y}_s)$.
From (3.2),

$$E(\theta|\underline{y}_s)=\sum_{(i,j)\in s}\frac{y_{ij}}{N}+\sum_i\sum_{k=1}^{K}\left(\frac{N_i(k)-n_i(k)}{N}\right)E(p_{ik}|\underline{y}_s) \tag{3.3}$$

where $p_{ik} = \exp\{\underline{X}^t(k)\underline{\beta}_i\}/[1+\exp\{\underline{X}^t(k)\underline{\beta}_i\}]$

and

$$\begin{aligned}Var(\theta|\underline{y}_s)=&\sum_i\sum_{k=1}^{K}\{N_i(k)-n_i(k)\}E\{p_{ik}(1-p_{ik})|\underline{y}_s\}/N^2\\&+Var\left(\sum_i\sum_{k=1}^{K}\{N_i(k)-n_i(k)\}p_{ik}|\underline{y}_s\right)/N^2.\end{aligned} \tag{3.4}$$

3.2 Numerical Evaluation

Since the posterior moments of θ are nonlinear functions of $\{\underline{\beta}_i:i=1,\ldots,L\}$, and the posterior distribution,

$$f(\{\underline{\beta}_i:i=1,\ldots,L\},\underline{\eta},\Gamma|\underline{y}_s), \tag{3.5}$$

cannot be expressed in a simple form, numerical evaluation is needed. We generate from (3.5) R sets of parameters, $\Omega = \{\Omega^{(r)}:r=1,\ldots,R\}$, where $\Omega^{(r)} = \{\{\underline{\beta}_i^{(r)}:i=1,\ldots,L\},\underline{\eta}^{(r)},\Gamma^{(r)}\}$. Then we evaluate the $p_{ij}^{(r)}$ using (2.2), and obtain estimates of $E(\theta|\underline{y}_s)$ and $Var(\theta|\underline{y}_s)$,

$$\hat{E}_\theta=\sum_{(i,j)\in s}y_{ij}/N+R^{-1}\sum_{r=1}^{R}\left[\sum_{(i,j)\notin s}p_{ij}^{(r)}/N\right] \tag{3.6}$$

and

$$\begin{aligned}\hat{V}_\theta=&R^{-1}\sum_{r=1}^{R}\left[\sum_{(i,j)\notin s}p_{ij}^{(r)}(1-p_{ij}^{(r)})+\{\sum_{(i,j)\notin s}p_{ij}^{(r)}\}^2\right]/N^2\\&-\{R^{-1}\sum_{r=1}^{R}\left[\sum_{(i,j)\notin s}p_{ij}^{(r)}/N\right]\}^2\ .\end{aligned} \tag{3.7}$$

This numerical evaluation is accomplished using a Gibbs sampler; see Malec and Sedransk (1993b) for details.

4. Variable Selection

Using data from the 1987 NHIS we select the variables to be included in (2.2) and (2.3) where the binary variable Y has Y=1 if there has been at least one visit to a physician during the past twelve months. We proceed in two steps by first fitting an individual-level model using (2.1) and (2.2), and then considering the county-level model in (2.3).

Our initial objective is to ascertain the general form of (2.2). We do this by ignoring county variation and estimating $\underline{\beta}$ in the "national" model, (4.1). If $\underline{X}_{ij} = \underline{X}(k)$, (2.1) and (2.2) are replaced by

$$Pr(Y_{ij}=y_{ij}) = p_k^{y_{ij}}(1-p_k)^{1-y_{ij}}, y_{ij} \in \{0,1\},$$

and

$$\ln\{p_k/(1-p_k)\} = \underline{X}^t(k)\underline{\beta}. \tag{4.1}$$

First, we obtain estimates based on the saturated model where the sample proportion of individuals in class k, $\hat{p}_k$, is used to estimate p_k. Figure 1 shows the effect of age, race and sex on $\ln\{\hat{p}_k/(1-\hat{p}_k)\}$.

The variation in log odds in Figure 1 corresponds to an expected pattern. First, for a given sex and age, the probability of a physician visit is generally larger for whites than for nonwhites. Second, the general patterns are similar for both races for a given sex. For males, the probability of a physician visit decreases steadily until about age 22.5, and then increases steadily. (Recall that we are using five year age groups.) For females, physician visits decrease steadily until age 12.5 and then increase to about age 27.5. Physician visits remain roughly constant from 27.5 until 62.5 and then increase steadily.

Due to the complex form seen from Figure 1, various spline models, linear in age, were used. A fixed knot spline can be defined as a linear model (Smith 1979) and, hence, used in (2.2). We include the possibility of a knot at each five year age group. The general model investigated included all possible splines that are linear in age, a race effect, a sex effect, a race by sex interaction and, finally, all interactions between these categorical variables and the linear age splines. The set of possible variables is

1) Categorical variables: Intercept, race (r), sex (s) and race by sex (rs)
2) Linear age splines: $X_a(k)=\max(0, \text{age}(k)-a)$, $a=0,5,10,\ldots,85$ and age(k) is the age for individuals in class k.
3) Categorical by age-spline interactions: r by $X_a(k)$, s by $X_a(k)$, rs by $X_a(k)$.

To determine a subset of terms to include in (4.1) the SAS forward stepwise logistic regression procedure, PROC LOGISTIC, was used. This procedure selects variables for inclusion and exclusion using a residual chi-squared test. Since the sample size is approximately 120,000 persons, variables possibly having only a small effect may be included in the model. To determine the total number of variables to use in the model a quantity like R^2 was used. Define the deviance D^2 for the model M_1 as $(\text{Dev}(M_1)-\text{Dev}(M_0))/(\text{Dev}(M_s)-\text{Dev}(M_0))$, where M_0 is the null model (with only an intercept term) and M_s is the saturated model (a parameter is fitted for each age by race by sex group). Note that $0 \leq D^2 \leq 1$, and equals R^2 for the linear model. The variables, intercept, $sX_{15},\ldots,r$, listed in the table below were included in the model. Adding other variables does not increase the value of D^2 appreciably (note the small contributions of the next best variables, rX_0 and sX_{45}, to D^2).

Variable	Intercept	sX_{15}	X_0	X_{15}	sX_{25}	X_{55}	X_{25}	r	rX_0	sX_{45}
Cumulative D^2	0.00	17.17	22.09	58.88	75.02	87.41	91.55	94.41	95.37	95.70

To check the fit of the model, partial residuals were plotted. Corresponding to each observation there is the residual, $r_{ij} = (y_{ij} - \hat{p}_{ij})/\hat{p}_{ij}(1-\hat{p}_{ij})$, which is then averaged over subsets of interest. A typical residual plot has, for a given domain (e.g., sex by race), r_{ij} averaged over all individuals of a given age plotted against age. The particular form of the residual is used because it will estimate a missing term in the logistic model (see Fienberg and Gong's comment to Landwehr, Pregibon and Shoemaker 1984). The residual plot in Figure 2 indicates that the eight variable model provides a good fit to the data. The one large remaining residual (for Black males, aged 85+) corresponds to an estimate based on a very small sample size.

The second step in the data analysis is to identify county-level covariates that affect an individual's probability of visiting a doctor, after having removed the effects due to the individual level covariates. To do this, we combined the individual level and county level models in (2.2) and (2.3) but set $\Gamma=0$. Then

$$\ln\{p_{ij}/(1-p_{ij})\} = X_{ij}^t G_i \eta \ . \tag{4.2}$$

To reduce the scale of this investigation we consider only the eight individual level variables identified earlier. As indicated by (4.2), we allow main effects of county-level variables and interactions of these county covariates with the individual-level variables. The collection of county covariates that we considered are ones included in the Area Resource File or county mortality file, and thought by subject-matter specialists to be relevant. We have also included county variables related to the formation of the NHIS strata. The procedure we used was to force the eight individual-level variables into the model, and let the SAS stepwise logistic regression procedure add variables. (We have also used graphical methods as described in Malec and Sedransk 1993a and Malec, Sedransk, and Tompkins 1993.) We found no county-level covariates that increased D^2 appreciably. However, there is still considerable county-to-county variation to be captured by (2.3) with $G_i = I$. For other dependent variables (e.g., health status), county-level covariates play a more significant role.

5. Comparison of Alternative Estimators

In this section we use data from the 1987 NHIS to compare the Bayes estimates with the standard alternatives, synthetic and design-based estimates.

For the largest states, the conventional design-based estimates should have relatively small variances, and there should be good agreement between them and estimates based on (3.2). In Figure 3 we plot, for each state and type of estimator (design-based, Bayes, synthetic), the estimated percent of the state population who visited a physician against state sample size. The Bayes estimates (based on a normal approximation to the posterior distribution) are close to the design-based estimates for the largest states, as one would hope. For the same states, the synthetic estimates are always further from the design-based estimates than are the Bayes estimates. As the state sample sizes become smaller the design-based estimates become increasingly unreliable, and the Bayes estimates look less like the design-based estimates, and more like the synthetic estimates. We have also used this same model to produce state estimates of the percent visiting a physician for subpopulations such as persons 65+, non-whites and females. These estimates exhibit the versatility of Bayes estimates; the between county variability, based on Γ, is different for these three cases, leading to different amounts of "gaining of strength". See Malec and Sedransk (1993b) for details and estimates.

Corresponding to Figure 3, Figure 4 is a plot for the 51 areas of posterior standard deviations vs. state sample sizes where we consider both the hierarchical Bayes (formulas (2.1) - (2.4)) and "synthetic" estimates. For the states with smaller populations, the standard deviations based on the hierarchical Bayes model more properly account for the uncertainty associated with inference about θ.

6. Evaluation

We have investigated whether the conventional sample weights are informative. Figure 5 is a partial residual plot similar to Figure 2. (For this analysis, $r_{ij} = \{y_{ij} - E(p_{ij}|\underline{y}_s)\}[E(p_{ij}|\underline{y}_s)\{1-E(p_{ij}|\underline{y}_s)\}]^{-1}$.) The ordinate of each point is the average residual for all individuals having a sample weight within the range centered at the corresponding abscissa. There is no evidence that the model should include the sample weight as a covariate.

Since the frequency of persons who visit a physician is not available for the entire NHIS population, it is not possible to compare the small area estimates with the true state values. However, by removing a portion of the sample, cross-validation methods can be used to assess how well the model and estimation procedure predict the part of the sample that has been deleted.

The cross-validation procedure that we plan to use is described below. Define the set of sampled elements that are set aside as "A". Let $\underline{y}_A$ denote the vector of observations that correspond to the elements in A and $\underline{y}_{(A)}$ the remaining sampled elements. Also, $\underline{Y}_A$ is the random variable (with observed value $\underline{y}_A$) that represents the removed elements. The predictive distribution, $f(\underline{Y}_A|\underline{y}_{(A)})$, can be used to make comparisons between the observed data, $\underline{y}_A$, and the values of $\underline{Y}_A$ predicted from the model. Specific functions comparing $\underline{Y}_A$ and $\underline{y}_A$, denoted $g(\underline{Y}_A|\underline{y}_A)$, can be defined to evaluate features of the predictions. (See Gelfand, Dey and Chang 1991 for a general review of Bayesian model assessment.)

We shall remove sets of sample elements in ways that permit us to see if our model captures the most important features of the NHIS data. Our evaluation will be based on how well the model predicts the deleted sample,

$$\theta_{AU} = \frac{\sum_{i\in U}\sum_{k=1}^{K}\sum_{j\in A_{ik}} y_{ikj}}{\sum_{i\in U}\sum_{k=1}^{K} n_{Ai}(k)}$$

where the first sum is over all counties in state "U", A_{ik} denotes the set of deleted individuals in demographic group k and county i, and $n_{Ai}(k)$ is the size of A_{ik}. Two choices for the error in prediction are

$$g_{1U}(\underline{Y}_A, \underline{y}_A) = (\theta_{AU} - E(\theta_{AU}|\underline{y}_{(A)}))^2$$

and

$$g_{2U}(\underline{Y}_A, \underline{y}_A) = \left(\frac{\theta_{AU} - E(\theta_{AU}|\underline{y}_{(A)})}{E(\theta_{AU}|\underline{y}_{(A)})}\right)^2.$$

To evaluate how well the model can predict the error of the estimate one may use

$$g_{3U}(\underline{Y}_A, \underline{y}_A) = \frac{g_{1U}(\underline{Y}_A, \underline{y}_A) - E(g_{1U}(\underline{Y}_A, \underline{y}_A | \underline{y}_{(A)})}{E(g_{1U}(\underline{Y}_A, \underline{y}_A) | \underline{y}_{(A)})}.$$

Numerical results from this cross-validation will appear in a forthcoming report.

References

Dempster, A.P., and Tomberlin, T.J. (1980), "The Analysis of Census Undercount From a Postenumeration Survey," Proceedings of the Conference on Census Undercount, Arlington, VA, 88-94.

Gatsonis, C., Normand, S-L., Liu, C. and Morris, C. (1993), "Geographic Variation of Procedure Utilization: A Hierarchical Model Approach," Medical Care, 31, YS54-YS59.

Gelfand, A.E., Dey, D.K. and Chang, H. (1991), "Model Determination Using Predictive Distributions with implementations via Sampling-Based Methods". In Bayesian Statistics 4, eds., J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Clarendon Press.

Landwehr, J.M., Pregibon, D., and Shoemaker, A.C. (1984), "Graphical Methods for Assessing Logistic Regression Models," Journal of the American Statistical Association, 79, 61-83.

Malec, D., and Sedransk, J. (1985), "Bayesian Methodology for Predictive Inference for Finite Population Parameters in Multistage Cluster Sampling," Journal of the American Statistical Association, 80, 897-902.

Malec, D., and Sedransk, J. (1993a), "Bayesian Predictive Inference for Units with Small Sample Sizes: The Case of Binary Random Variables," Medical Care, 31, YS66-YS70.

Malec, D., and Sedransk, J. (1993b), "Small Area Inference for Binary Variables in the National Health Interview Survey," Technical Report.

Malec, D., Sedransk, J., and Tompkins, L. (1993), "Bayesian Predictive Inference for Small Areas for Binary Variables in the National Health Interview Survey," In Case Studies in Bayesian Statistics, eds., C. Gatsonis, J.S. Hodges, R.E. Kass and N.D. Singpurwalla. Springer Verlag.

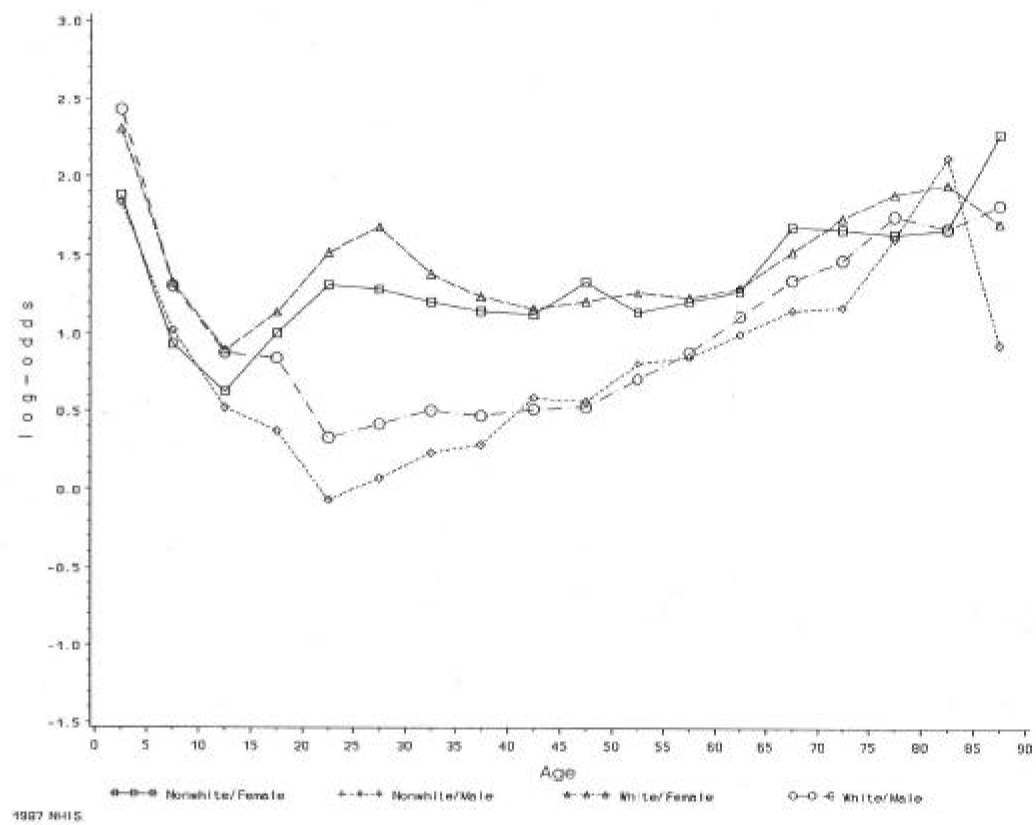
Massey, J.T., Moore, T.F., Parsons V.L., and Tadros, W. (1989), "Design and Estimation for the National Health Interview Survey, 1985-94," National Center for Health Statistics. Vital and Health Statistics, 2:110.

Smith, P.L. (1979), "Splines As a Useful and Convenient Statistical Tool," The American Statistician, 33, 57-62.

U.S. Department of Health and Human Services (1989), "The Area Resource File (ARF) System". ODAM Report No. 7-89.

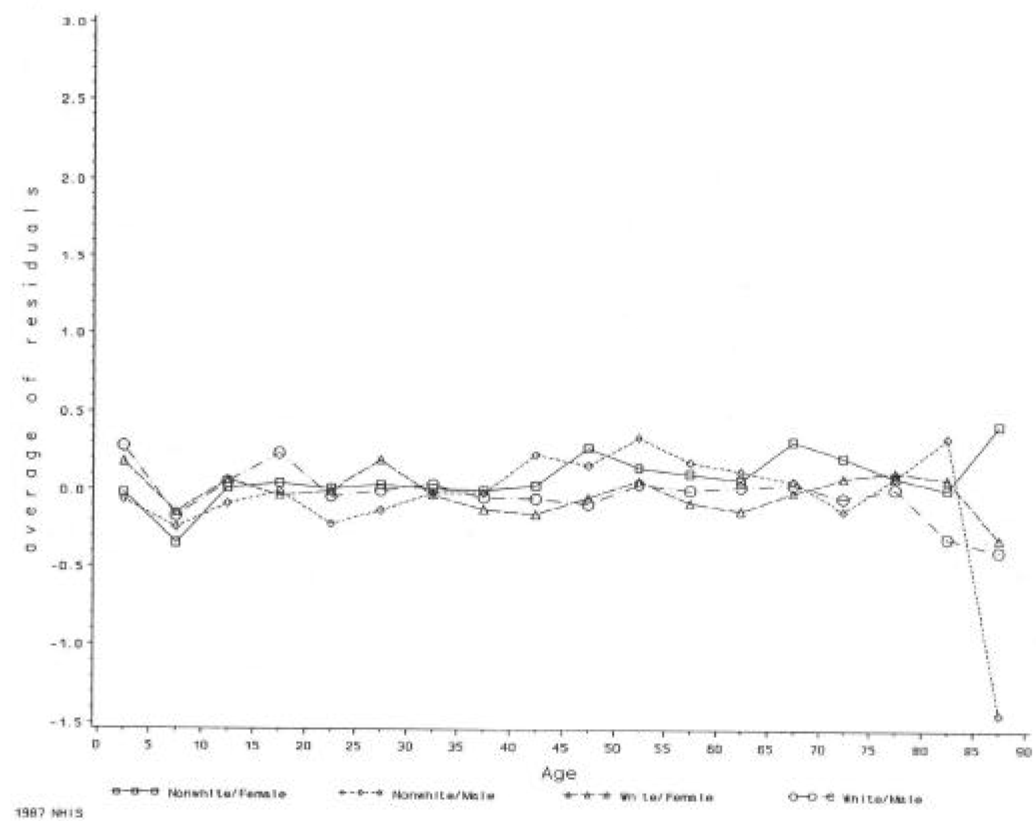
Wong, G.Y., and Mason, W.M. (1985), "The Hierarchical Logistic Regression Model for Multilevel Analysis," Journal of the American Statistical Association, 80, 513-524.

Figure 1: Relationship of log odds of a sample proportion and age for each race by sex group.



NOTE: The binary variable is the presence or absence of at least one visit to a physician within the past year.

Figure 2: Relationship for each sex x race group of average residual and age using the model in section 4.



NOTE: The ordinate of each point is the average residual from the model for all members of a given sex by race by age group.

Figure 3: Estimated percent of population in a state who visited a physician in the past year plotted against state sample size: Hierarchical Bayes, synthetic and design-based estimates.

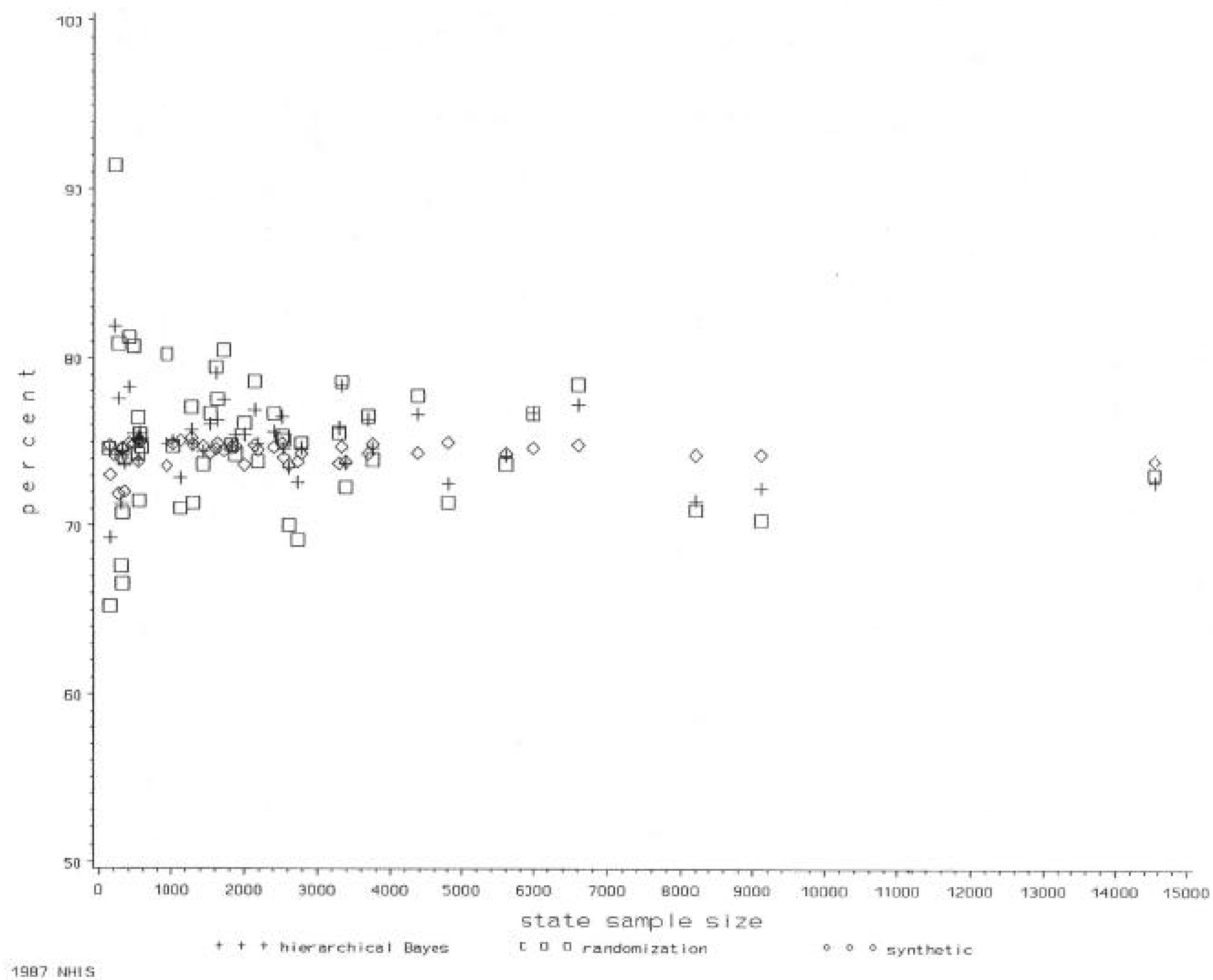


Figure 4: Posterior standard deviations plotted against state sample sizes: Hierarchical Bayes and synthetic models.

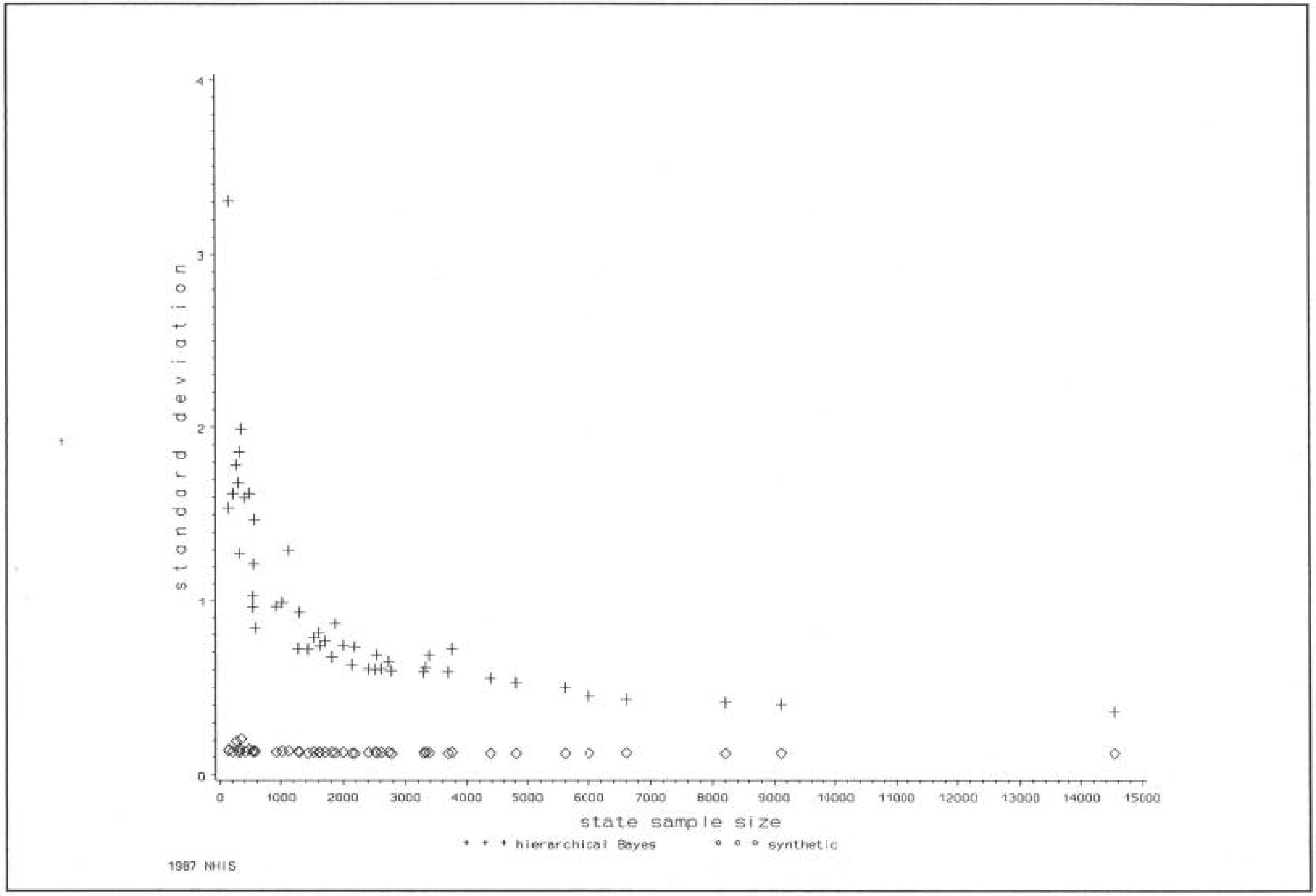
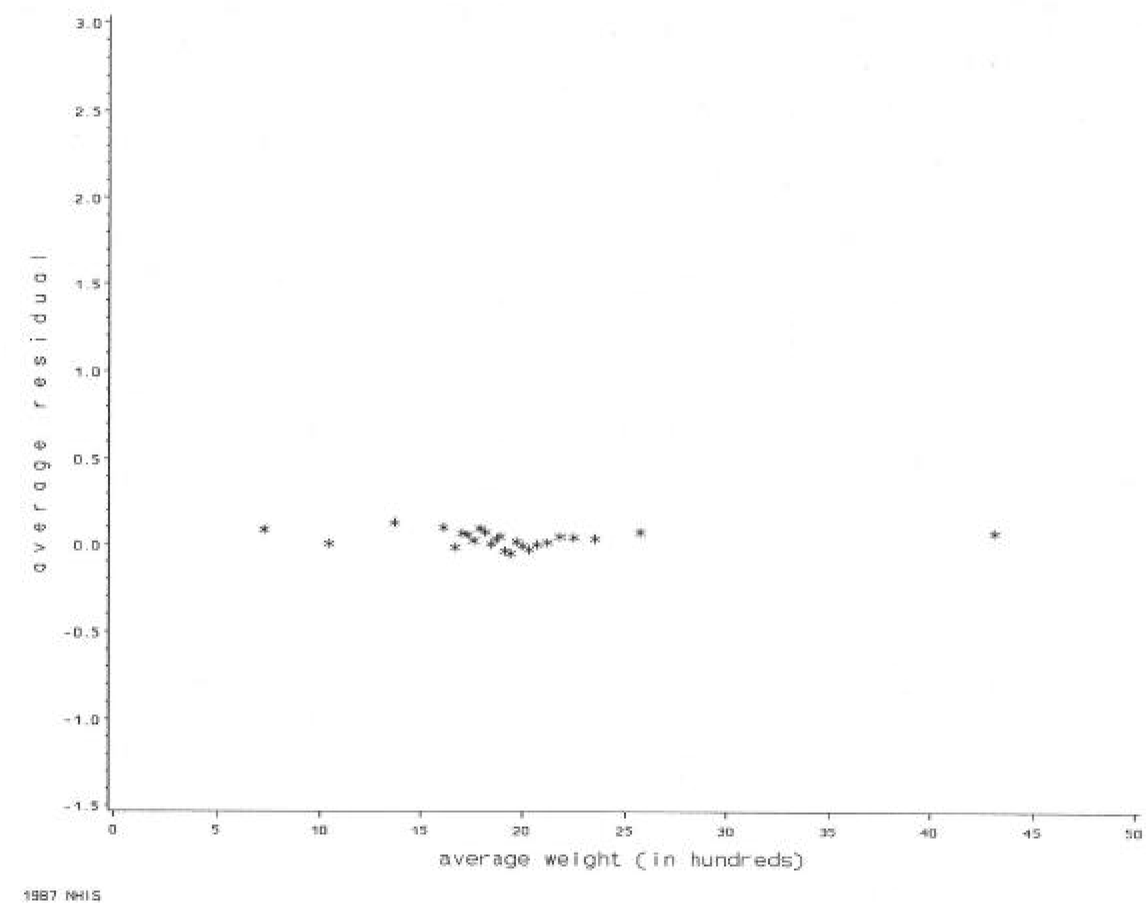


Figure 5: Relationship of average residual and sample weight using the model in section 4.



NOTE: The ordinate of each point is the average residual from the model for all individuals having a sample weight within a narrow range.

THE ROLE OF DESIGN BASED VARIANCES AND COVARIANCES IN SMALL DOMAIN ESTIMATION

Robert E. Fay[1]
U.S. Bureau of the Census

1. Introduction

Two recent reviews provide the context for this paper. The Subcommittee on Small Area Estimation, Federal Committee on Statistical Methodology (Schaible and Gonzalez 1993) surveyed applications of indirect estimators in official U.S. government use. The first chapter of their report noted the predominance of direct estimators in federal statistics. In other words, official estimates are almost always "direct," through exclusive or almost exclusive use of data from both the time and domain of interest. Indirect estimators, relying heavily on data from either other domains or times (or both), are the exception in federal statistics. The report enumerates and discusses indirect estimators in current federal use. (That is, the report considered only those applications published as official estimates, not including methodological tests and discontinued series. Generalizations of survey variance estimates, for example, those often included in source and reliability statements at the end of Census Bureau reports, were also not included.)

Although infrequently employed in federal practice, indirect estimation generally reflects an attempt to address a need for estimates that cannot be reliably produced directly given constraints on resources. The concluding chapter of the report urged caution in the use of indirect methods and eschewed advocacy of them as a general purpose and easily developed solution.

Ghosh and Rao (1994) reviewed the statistical methodology underlying several types of indirect estimators. Their review included demographic and other methods specific to postcensal population estimation; synthetic, composite and related estimators for domain characteristics; and empirical best linear unbiased predictors (EBLUP), empirical Bayes (EB), and hierarchical Bayes (HB). This paper employs their review as a point of departure for comparisons of existing theory to practice.

Several small domain applications that have appeared in the literature share enough common features to be studied as a group. One class of applications, which represents the scope of this paper, combines information from survey estimates at the domain

[1] This article represents results of research undertaken by a staff member of the Census Bureau. The views expressed are attributable to the author and do not necessarily reflect those of the U.S. Bureau of the Census.

level with domain-level characteristics available from independent sources. Examples include estimates of 1970 census income for small places (Fay and Herriot 1979), estimates of 1980 census undercount (Ericksen and Kadane 1985, Cressie 1992), estimates of 1990 census undercount produced by the U.S. Census Bureau in 1991 (subsequently revised under a different methodology), and estimates of median family income by state (Fay, Nelson, and Litow 1993).

In some of these applications, independent data provide a basis for evaluating the methodology. For example, in estimating median family income by state, the decennial census figures serve as a gold standard by which to judge the performance of the resulting small domain estimates. Although this comparison is available only every 10 years, the empirical results support the application. As a second example, the relatively small number of available special censuses taken after the 1970 census also corroborated the application to 1970 census income for small places. In other cases, however, including the analysis of census undercount, there is no gold standard by which to evaluate the resulting estimates. Consequently, the validity of the application of the underlying theory for both the properties of the resulting estimates and the measurement of their reliability is of considerable importance.

Comparison of these and other applications to the available theory generally shows that the explicit theoretical conditions are not completely satisfied, although to varying degrees. Consequently, each application implicitly requires that the departures from the theory do not pose serious consequences. As the title of this paper suggests, the theoretical results typically assume that the sampling errors of the small domain estimates are known, whereas in practice they are frequently estimated from the data, either directly or through a model to generalize the variances.

Section 2 reviews much of the existing theory for the class of estimators under discussion. Section 3 then compares the applications just mentioned to the requirements of the theoretical formulations to note implicit extensions of the theory that, for the most part, still lack a theoretical foundation. Section 4 reports the results of simple Monte Carlo studies to assess evidence in some of these areas. Although mathematical proof is preferable to computer demonstration, the empirical results present useful evidence on the significance of various issues arising from the practical application of these procedures and suggest directions for new research.

2. Theoretical Results for a Class of Small Area Estimators

As noted in the previous section, Ghosh and Rao (1994) reviewed several general small area approaches. The class of models of interest to this paper employs auxiliary data $\boldsymbol{x}_i$ = (x_{i1} , ..., x_{ip}), which are assumed measured without sampling error. In their notation, the parameters of interest, θ_i , are assumed to be related to the $\boldsymbol{x}_i$ by

$$\theta_i = \boldsymbol{x}_i \beta + \nu_i z_i, \qquad i = 1, \ldots, m, \tag{2.1}$$

Frequently, the model takes the simpler form:

$$\theta_i = \boldsymbol{x}_i \beta + \nu_i, \qquad i = 1, \ldots, m, \tag{2.2}$$

where β is a vector of regression parameters, and the ν_i are independent, identically distributed (iid) random variables with:

$$E(\nu_i) = 0, \qquad V(\nu_i) = \sigma_v^2. \tag{2.3}$$

In (2.1), the z_i 's are known positive constants. Ghosh and Rao (1994) develop the theory in the general form (2.1). Results specific to the simpler model (2.2) are offered here because the formulas are more accessible.

The θ_i in (2.1) and (2.2) represent the parameters of interest for the small areas, such as local area per capita income, the ratio of correction population to census population, the number of employed, etc. The model reflects a possible lack of fit between the regression $\boldsymbol{x}_i \beta$ and the actual value through random effect terms, ν_i .

In this class of models, direct estimates, $\hat{\theta}_i$ are available at the domain level with

$$\hat{\theta}_i = \theta_i + e_i, \tag{2.4}$$

where the e_i represent sampling errors with

$$E(e_i | \theta) = 0, \qquad V(e_i | \theta) = \psi_i. \tag{2.5}$$

(In this section, the sampling errors are also assumed independent, but extensions have reflected correlated sampling errors.) In other words, the $\hat{\theta}_i$ are design-unbiased estimators. Ghosh and Rao comment that these conditions may be quite restrictive. For example, the estimators may not be unbiased, as in the case of undercount adjustment. In addition, the sampling variances ψ_i may not be known.

The combined model, using (2.2) and (2.4), is

$$\hat{\theta}_i = \boldsymbol{x}_i\beta + \nu_i + e_i . \qquad (2.6)$$

As Ghosh and Rao note, (2.6), which is a linear combination of fixed and random effects, is a special case of the general mixed linear model.

Ghosh and Rao (1994) discuss the estimation of (2.6) from the perspectives of EBLUP, EB, and HB; this paper will primarily focus on the EBLUP formulation. They cite Henderson (1950) as the originator of best linear unbiased predictors (BLUP) for models such as (2.6), when the variance components are known. Ghosh and Rao express the BLUP of θ_i as

$$\tilde{\theta}_i^H = \gamma_i\hat{\theta}_i + (1-\gamma_i)\boldsymbol{x}_i\tilde{\beta} , \qquad (2.7)$$

where

$$\tilde{\beta} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\hat{\theta} \qquad (2.8)$$

is the BLUE of β , $\boldsymbol{V}$ is the diagonal matrix with elements $\sigma_\nu^2+\psi_i$, and

$$\gamma_i = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \psi_i} . \qquad (2.9)$$

When the variance components are known, the mean square error of (2.7) under model (2.6) is

$$E(\tilde{\theta}_i^H - \theta_i)^2 = g_{1i}(\sigma_\nu^2) + g_{2i}(\sigma_\nu^2) , \qquad (2.10)$$

where

$$g_{1i} = \gamma_i \psi_i , \tag{2.11}$$

and

$$g_{2i}(\sigma_v^2) = (1-\gamma_i)^2 \, \boldsymbol{x}_i (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{x}_i^T . \tag{2.12}$$

Because σ_v^2 is typically unknown, a two-stage estimator, $\hat{\theta}_i^H$, arises by first estimating σ_v^2 from the data and then using it to obtain (2.7). Ghosh and Rao reference several options for estimating σ_v^2. A simple moment estimator $\sigma^2_{v(1)} = \max(\sigma^2_{v(1)}, 0)$, where

$$\sigma^2_{v(1)} = (m - p)^{-1}\left[(\hat{\theta} - \boldsymbol{X}\beta^*)^T(\hat{\theta} - \boldsymbol{X}\beta^*) - \sum_i \psi_i \{1 - \boldsymbol{x}_i\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i^T\}\right], \tag{2.13}$$

and

$$\beta^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1} \boldsymbol{X}^T \hat{\theta}$$

is the ordinary least squares estimator of β, has the advantage of not requiring iteration.

The remaining methods to be considered here each require iteration, unless the sampling errors ψ_i are equal. For a given trial estimate of σ_v^2, β is estimated through (2.8) at each cycle for each of the methods.

Fay and Herriot (1979) used an estimator $\sigma_{v(2)}$, based on the method of moments as the solution to the equation

$$(\hat{\theta}_i - \boldsymbol{X}\beta)^T \boldsymbol{V}^{-1} (\hat{\theta}_i - \boldsymbol{X}\beta) = m - p \tag{2.14}$$

or 0 in the case that no solution exists. If the sampling errors ψ_i are equal, then (2.13) and (2.14) have the same solution.

A third alternative, maximum likelihood, $\sigma_{v(3)}$, maximizes

$$L(\beta, \sigma_v^2) = -1/2 \log(|V|) - 1/2 (\theta - X\beta)' V^{-1} (\theta - X\beta), \qquad (2.15)$$

which is the log-likelihood up to a constant.

Cressie (1992) suggested the application of another estimator, $\sigma_{v(4)}$, based on restricted maximum-likelihood (REML), which maximizes the adjusted likelihood,

$$\begin{aligned} L^*(\beta, \sigma_v^2) = &-1/2 \log(|V|) - 1/2 (\log(|X'V^{-1}X|) \\ &- 1/2 (\theta - X\beta)' V^{-1} (\theta - X\beta), \end{aligned} \qquad (2.16)$$

where terms not involving the parameters have been dropped. Cressie (1992) further describes this procedure, which was originally developed by Patterson and Thompson (1971, 1974). In short, however, the procedure examines the likelihood of the residuals from the regression. When all the sampling errors ψ_i are equal, then (2.13), (2.14), and (2.16) have the same solution, while (2.15) yields a generally smaller estimate of σ_v^2.

Under normality of the error terms, Kackar and Harville (1984) showed that

$$E(\hat{\theta}_i^H - \theta_i)^2 = E(\tilde{\theta}_i^H - \theta_i)^2 + E(\hat{\theta}_i^H - \tilde{\theta}_i^H)^2,$$

where, for large m, the second term may be approximated by

$$g_{3i}(\sigma_v) = \psi_i^2 (\sigma_v^2 + \psi_i)^{-3} \bar{V}(\hat{\sigma}_v^2),$$

where $\bar{V}(\hat{\sigma}_v^2)$ is the asymptotic variance of $\hat{\sigma}_v^2$.

Prasad and Rao showed that an approximately unbiased estimator of the mean square error of the EBLUP estimator is

$$mse(\theta_i^H) = g_{1i}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2 g_{3i}(\hat{\sigma}_v^2) \qquad (2.17)$$

with bias of order lower than m^{-1}.

The MSE estimators studied in Section 4 share (2.17) but differ in the approach to estimate $\overline{V}(\hat{\sigma}_v^2)$. Section 4 describes these differences.

3. Assumptions Made by Some Previous Applications of EBLUP

Section 2, and the more complete review in Ghosh and Rao (1994), detail the assumptions of the available theory for EBLUP. This section briefly reviews potential discrepancies between the theory and some previously published applications.

One feature is common to all of the applications discussed here and can be assumed to occur almost universally for sample surveys, namely, that the sampling variances, ψ_i, are estimates rather than known values. The following discussion notes the consequent adaptations, which range from direct use of estimated variances to variance generalization.

Fay and Herriot (1979) described a large-scale implementation of EBLUP/empirical Bayes estimation to estimate per capita income in 1969 for small places and minor civil divisions with population below 1000 persons. The sample estimates, $\hat{\theta}_i$, were based on the long form sample of the 1980 census. Predictors included the county average PCI, 100% data from the census on housing value, and reported income from IRS returns. Because of computing constraints at the time, the authors refrained from any recalculation of the census sampling variances but instead employed the available variance generalization. The generalization was a simple national model without any allowance for geographic variation. Since the generalization yielded a linear relationship between ψ_i and θ_i^2, a logarithmic transformation of $\hat{\theta}_i$ gave a closer fit of the application to the theory. They employed (2.14) to estimate σ_v^2. Generally, the compositing, (2.7), drew on both the sample estimates and the regression in approximately equal amounts, rather than relying almost exclusively on one of the two. The authors employed (2.2) but observed some evidence of variation in σ_v^2 by size of place. The evidence suggested (2.1) with z_i decreasing with increasing size, n_i, although at a rate closer to $n_i^{-1/4}$ than $n_i^{-1/2}$. The authors did not attempt MSE estimation, but presented some limited empirical evidence from special censuses favoring the EBLUP approach.

Application of EBLUP to sample estimates of decennial census undercounts has been controversial, and the review here will simply focus on assumptions incorporated in the implementations rather than systematically evaluating the merits of the work on this subject. As Ghosh and Rao (1994) comment in passing, survey estimates of undercount in both 1980 and 1990 have been subject to substantial sources of bias, and the existing theory does not provide a clear measure of how EBLUP behaves under such conditions. Furthermore, gains from EBLUP and estimators of MSE have figured prominently in the undercount debate, since the 1980 PES estimates at the state level and the 1990 PEP estimates based on the original 1392 strata have such high sampling variability as to preclude adjustment without EBLUP or other smoothing. The estimators placed high weight on the regression and little on the direct estimates.

In both 1980 and 1990, estimates of ψ_i have appeared to depend on $\hat{\theta}_i$. The published 1980 analysis used the estimated variances in spite of this departure from the model. The 1991 analysis of the 1990 PES applied a variance generalization. Although opinions have been offered on the subject, a systematic analysis of the effect of the generalization on the 1991 estimates remains to be done. Furthermore, the 1991 smoothing was multivariate and employed large covariance matrices, formed from the generalized variances and directly estimated correlations. Fay (1992) showed through stratified bootstrap samples that this approach induced substantial additional variability not reflected in the MSE's computed by the Census Bureau.

The 1980 PES estimates were subject to substantial amounts of missing data, yet no estimates of missing data variance are available, and the author is unaware of systematic analysis showing what possible effect this factor might have had on the 1980 analysis.

Ericksen and Kadane (1985) and the 1991 EBLUP for the 1990 PES both employed (2.2), whereas Cressie (1992) reanalyzed 1980 estimates with $z_i = n^{-1/2}$. Although Cressie argued for this choice on intuitive grounds, empirical evidence on this question is limited and virtually impossible to obtain from the undercount estimates themselves. The 1990 application employed (2.2); yet the sample estimates suggested that it failed to hold because σ_v^2 appears much larger in minority poststrata than elsewhere.

The U.S. Census Bureau has employed an EBLUP procedure to estimate median family income for 4-person families by state from the Current Population Survey (Fay, Nelson, and Litow 1993). The model can be calibrated against census values every 10 years. These calibrations have favored continued use of (2.2) at the state level, distinctly rejecting proposals such as $z_i = n^{-1/2}$ (Cressie 1992) in this application. The authors account for different approaches to estimating ψ_i and σ_v^2 over the evolution of the model. Over time, more emphasis has been placed on direct estimates.

In short, 1) each of these applications has rested on implicit extensions of the existing theory, 2) some empirical evidence suggests that these procedures can be useful under some conditions, but 3) a more systematic approach to assessing effects of uncertainty for EBLUP is still needed. The next section does not fully meet this need, but it does suggest the value of large scale Monte Carlo simulation as a productive approach to some of these questions.

4. Monte Carlo Evaluation

4.1 Basic Design of the Study

As noted earlier, the derivation of the estimators of mean square error rest on expectations taken both over repetitions of the sample and over the random effects. The more usual perspective of finite population sampling considers the population as fixed but unknown. In order to bridge the consequences of these two points of view, this study generated several finite populations, θ, and compared the properties of the mean square error estimates for each resulting population. Although the expressions in Section 2 focused on errors for individual components, (2.17), this section studies the accuracy of the estimated sum over domains of squared errors, much as the literature on the James-Stein estimator. The actual MSE's are compared to (2.17) summed over i.

Two values of m, 20 and 50, offer some indication of the effect of number of domains on the estimators. The first offers an approximate lower boundary on the range of usual application, while the second illustrates the effect of somewhat larger m. The primary emphasis will be on 50. Obviously, results for larger numbers of domains, such as 200, would also be desirable.

The section reports results that share the following common elements:

1) A set of population values for the domains, θ, is drawn from some distribution. For normal applications, for example, the domain means are selected.

2) Samples, $\hat{\theta}$, are drawn from the domain population and sampling variances estimated.

3) One or more EBLUP's, $\hat{\theta}^H$, are constructed.

4) Steps 2) and 3) are repeated for a total of 2000 samples from the population defined at step 1).

5) The MSE's of the EBLUP's, estimated as the average value of $(\hat{\theta}_i^H - \theta_i)^2$, calculated from the 2000 samples and summed over the domains, i, become the standard for comparison to the corresponding estimated MSE sums of (2.17) over i.

6) The bias and mean square error of the MSE estimators are then derived, and also key frequencies, such as the percent of samples in which the estimated MSE understates the actual MSE by 25 percent.

Steps 1)-6) generate one point in the Monte Carlo study. In other words, each point represents a specific population realized from the superpopulation, where the performance, over repeated sampling, of each of the EBLUP estimators and MSE estimators is evaluated.

Obviously, the MSE's at step 5) are not entirely free from sampling error themselves, but the relatively large number of samples provides practical justification for this procedure. The results presented in Figures 1-13 show that this procedure produces stable values.

Thus, the perspective is similar to design-based finite population sampling, since the criteria assess the performance for individual over repetitions of the sampling design.

4.2 Results for Normal Populations

For simplicity, four X variables were constructed:

1) The grand mean.

2) An indicator variable dividing the domains into halves according to domain number, $i = 1, \ldots m$. For example, for $m = 20$, the variable distinguishes the first 10 from the second.

3) A linear term, increasing with the domain number.

4) A similar quadratic term.

The sample size, n, for each domain was fixed at either 10 or 20, and the observations were scaled by $n^{1/2}$ in order to give the sample means unit variance. Because the analysis is invariant to the true β, these coefficients were set to 0 in generating the Monte Carlo samples.

Eight combinations were studied:

1) Use of known sampling variances, $\psi_i = 1$, in combination with (2.14) and $\bar{V}(\hat{\sigma}_v^2)$ from Prasad and Rao (1990, p. 167, (5.19)):

$$\bar{V}(\hat{\sigma}_v^2) = 2m^{-1}\left[\hat{\sigma}_v^4 + 2\hat{\sigma}_v^2 \sum \psi_i/m + \sum \psi_i^2/m\right] \qquad (4.1)$$

2) Use of generalized sampling variances assuming that the ψ_i are equal to some unknown constant, which is then estimated as the average of the sample estimates of ψ_i. The remaining estimation is done as in 1).

3) MLE using known sampling variances, $\psi_i = 1$, and:

$$\bar{V}(\hat{\sigma}_v^2) = 2\left[\sum (\hat{\sigma}_v^2 + \psi_i)^{-2}\right]^{-1} \qquad (4.2)$$

4) MLE using the estimated variances and the estimation approach of 3).

5) REML using the estimated variances and $\bar{V}(\hat{\sigma}_v^2)$ from Cressie (1992, p. 82, (3.22) and p. 85, (4.11)).

6) REML using the estimated variances and the more approximate expression (4.2).

7) The method of moments estimator, (2.14), and:

$$\bar{V}(\hat{\sigma}_v^2) = 2m\left[\sum(\hat{\sigma}_v^2 + \psi_i)^{-1}\right]^{-2} \quad (4.3)$$

8) The simple moment estimator, (2.13), and (4.1).

Cressie's (1992) estimator studied included in 5) is the only one of the group to explicitly incorporate the effect of the regression in estimating $\bar{V}(\hat{\sigma}_v^2)$. All others depend on m being large compared to p. In fact, however, differences between 5) and the simpler 6) were extremely modest. Potentially, similar refinements could be incorporated into the other estimators of $\bar{V}(\hat{\sigma}_v^2)$, but their impact is again likely to be small unless p is a substantial proportion of m.

Figures 1-13 results for $m = 50$, $n = 10$, that is a comparatively large number of domains with comparatively few degrees of freedom in each domain to estimate the variance in each. Of course, no one choice of these values is appropriate to represent the usual situation in most small domain estimation. Comments will follow about the results obtained for $m = 20$ and for $n = 20$.

A series of 28 populations are represented: 4 drawn from $N(0, \sigma_v^2)$ with $\sigma_v^2 = .125$, and 8 each from $\sigma_v^2 = .25$, .5, and 1.0. Figures 2, 4, 6, 8, 10, and 12 each omit the points for $\sigma_v^2 = .125$, which are generally far off the scale; further comments on this point follow.

Figure 1 shows the actual MSE for 2) as a function of $\sum\theta_i^2$, which is called the "SS of true deviations" in the figures. Over the entire range studied, the EBLUP improves on the direct sample estimates, but the improvement is most dramatic at the leftmost portion of the range, where $\sigma_v^2 = .125$, and the true values θ almost fit the regression line. The pluses and x's distinguish between different super population values of σ_v^2 used to generate the θ_i, but quite clearly this distinction is unimportant once the results are conditioned on $\sum \theta_i^2$.

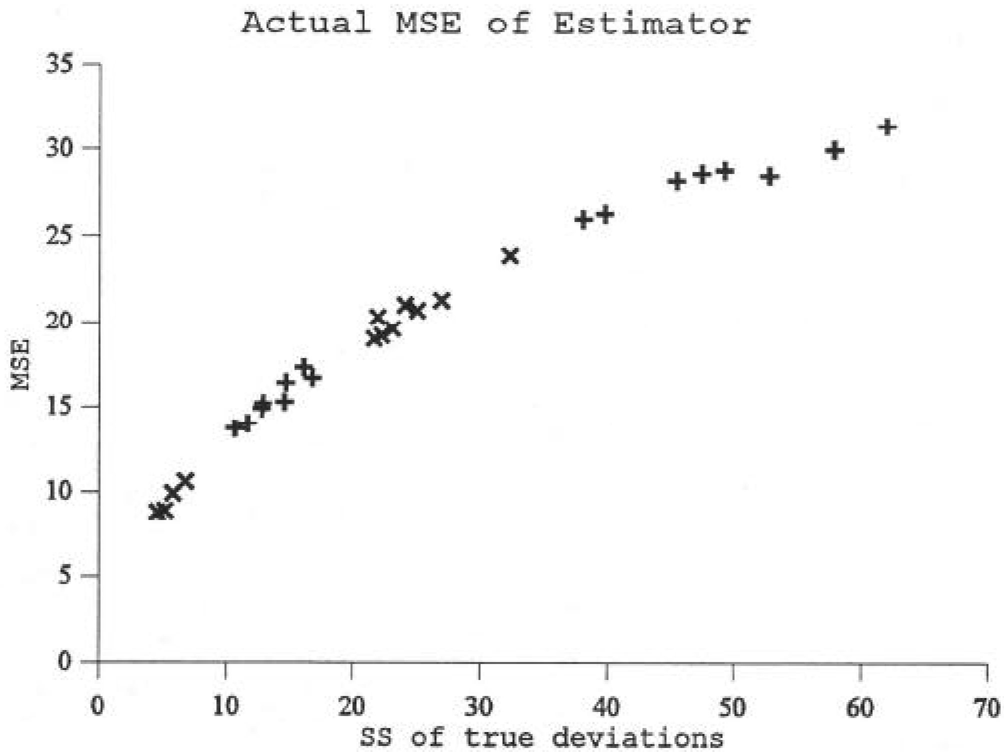


Figure 1 Actual MSE for "unbiased methods," generalized equal variances, m = 50, p = 4, 10 obs. per domain.

If the actual MSE's of any of the alternatives were superimposed on Figure 1, there would be substantial overlap. The MSE's and other performance characteristics of 1), with known variances, are virtually identical to 2). The actual MSE's of 3), MLE with known variances, are also almost identical to those in Figure 1. When sampling errors are instead estimated, the actual MSE's are a bit larger: by about 15-30% for REML and 8-20% for the other alternatives when the actual MSE is below 20, and by lesser amounts over the upper end of the range.

Since the MSE of the sample means is 50, Figure 1 includes a broad range of outcomes. At MSE=30, EBLUP yields distinct gains that, nonetheless, many practitioners might choose to forego in favor of the greater simplicity and interpretability of the direct sample estimates. At MSE=15, the gains from EBLUP may have a substantial impact on the utility of the estimates.

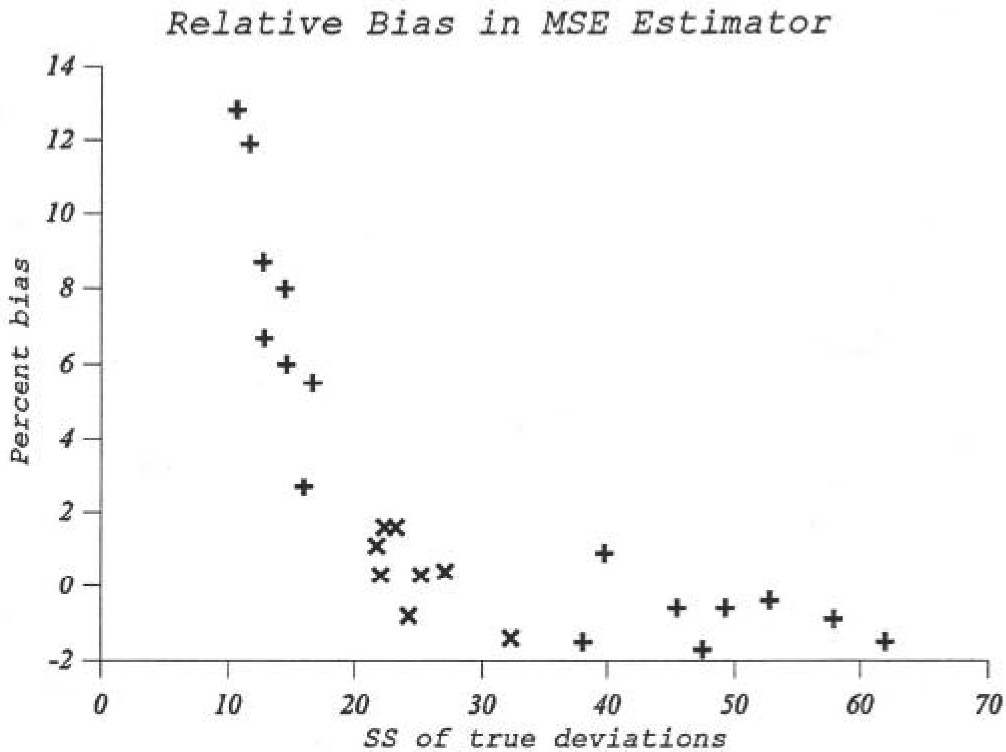


Figure 2 Percent bias in Prasad-Rao MSE estimator, generalized equal variances, "unbiased methods," m = 50, p = 4, 10 obs. per domain. Note: The first 4 points have been omitted.

Figure 2 reports the relative bias of the MSE estimator for 2) over the range of $\sum \theta_i^2$. The leftmost 4 points have been omitted from the graph because the bias increases dramatically, to around 30-40%, in that region. As noted previously, the performance for 1), with known sampling variances, is virtually identical to Figure 2.

From the perspective of bias, the performance of the MSE estimator is quite satisfactory over a large part of the range, but it becomes upwardly biased under conditions where the EBLUP estimator has the most pronounced effect, that is, in the leftmost portion of the range, below MSE = 15 or so.

Figure 3 evaluates the performance of the MSE estimator in a different manner, by showing the proportion of times that the estimated MSE falls below the actual MSE by 25% or more. For example, when the actual MSE=20, the figure reports the percentage of samples in which the estimated MSE is below 15.

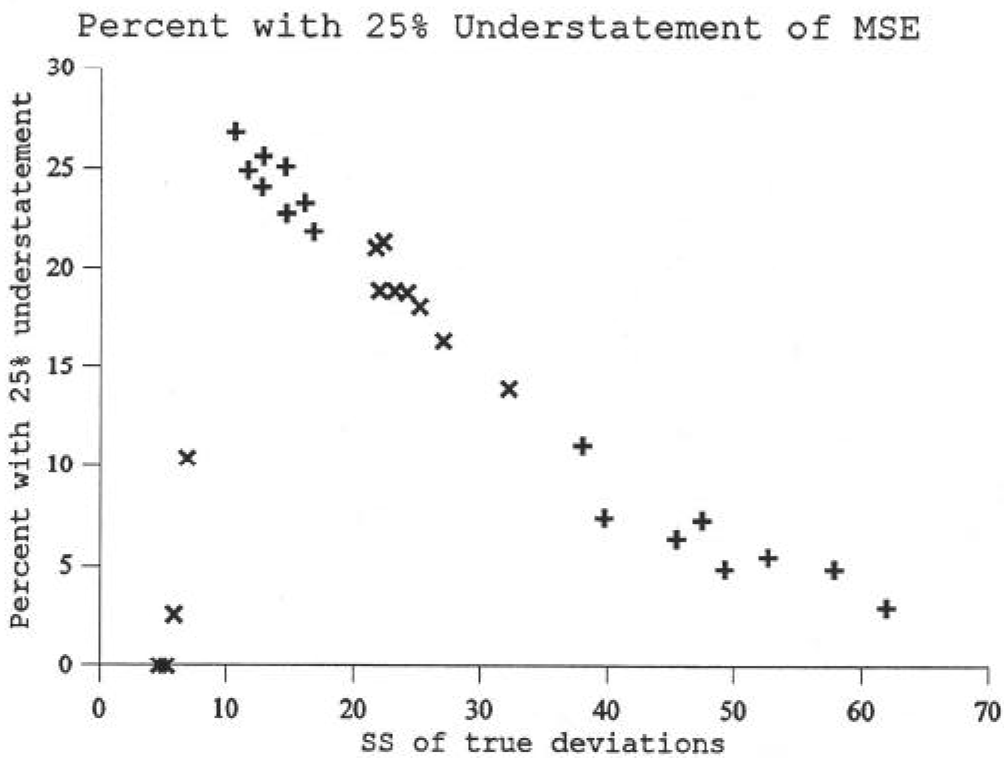


Figure 3 Percent understatement of true MSE by 25 percent or more, generalized equal variances, "unbiased methods," $m = 50$, $p = 4$, 10 obs. per domain.

The findings of Figure 3 are not easily predicted from Figure 2. In spite of the low level of bias in the MSE estimator over the range of MSE=20 and above, the probability that the estimator will substantially understate the actual MSE rises steadily as MSE decreases. Even more striking, however, is the dramatic fall towards 0% at the right of the figure. In fact, in this lower range, the contributions of the more stable components of (2.17), namely its second and third terms, are able to prevent a large understatement regardless of the contribution of the far more erratic first term.

Figure 4 presents comparable results for 3), MLE with known variances. Figure 4 reports a consistent downward bias in the estimated MSE for MLE. This finding agrees with a comparison of REML and MLE by Cressie (1992). Presumably, this downward bias could be even more severe when the ratio of *p* to *m*, which is 4 to 50 in this case, is larger.

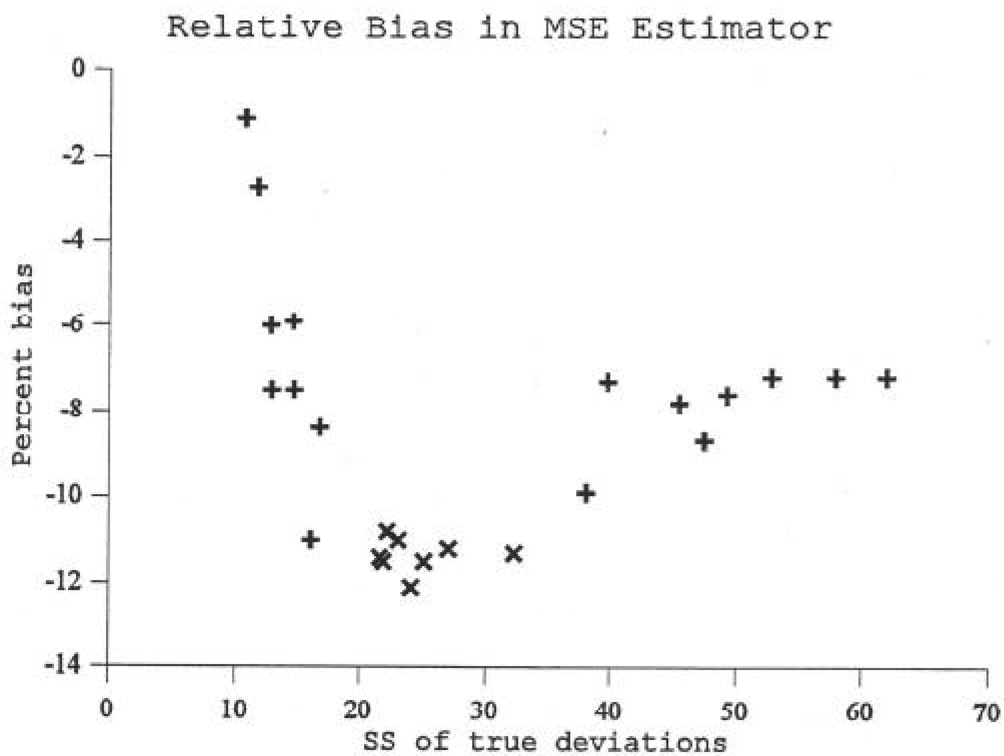


Figure 4 Percent bias in Prasad-Rao MSE estimator, generalized equal variances, MLE, $m = 50$, $p = 4$, 10 obs. per domain. Note: The first 4 points have been omitted.

In spite of the general downward bias in the MSE estimate, the bias changes sign and increases up to about 15-30% for the 4 lowest points included in the study.

Figure 5 presents results for MLE analogous to those in Figure 3. Noting the change in scale between the two figures, Figure 5 shows even higher proportions of significant understatement of the MSE over a large proportion of the range. This finding is consistent with the general downward bias exhibited in Figure 4.

As in Figure 3, however, the probability of significant understatement falls off dramatically near 0.

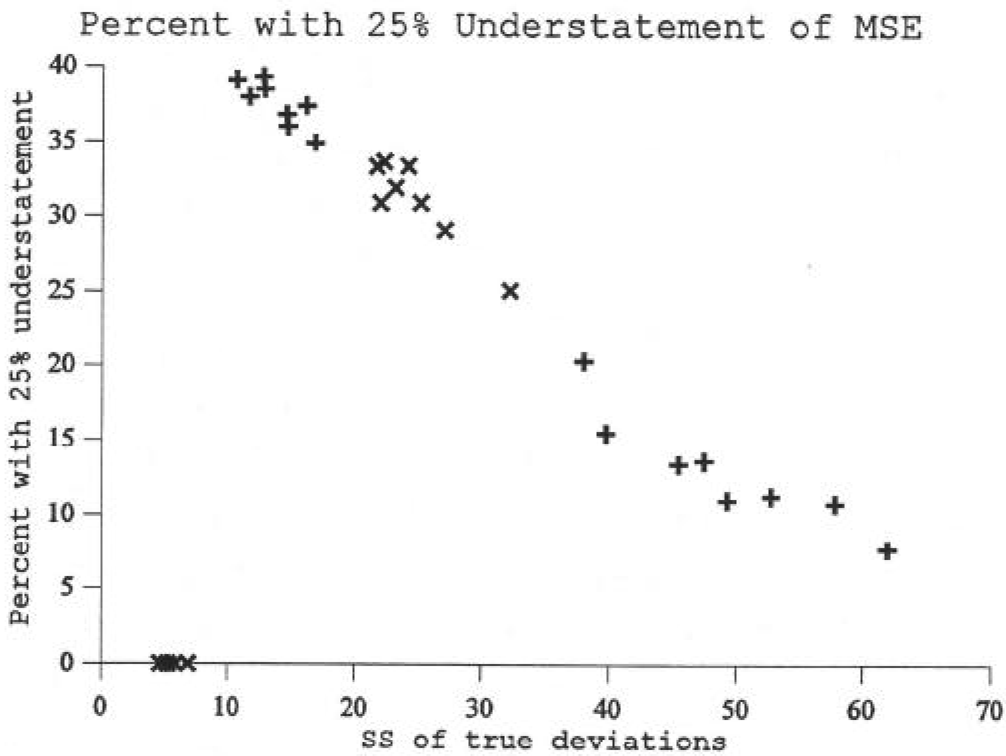


Figure 5 Percent understatement of true MSE by 25 percent or more, generalized equal variances, MLE, m = 50, p = 4, 10 obs. per domain.

As noted earlier, shifting from known variances to estimated variances for each of the domains increases the actual MSE of the MLE by about 8-20% for actual MSE's below 20, and somewhat less for larger actual MSE's. Figure 6 reports the performance of the MSE estimator in this instance, as an estimator of the actual, and now larger, MSE. Comparison of Figures 4 and 6 indicates some common features but considerable differences as well. On the right of Figure 6, the downward bias is even more pronounced than in Figure 4. For decreasing MSE, however, the bias crosses 0 earlier than in Figure 4. The bias for the omitted points rises to approximately the same range, that is, about 15-30%.

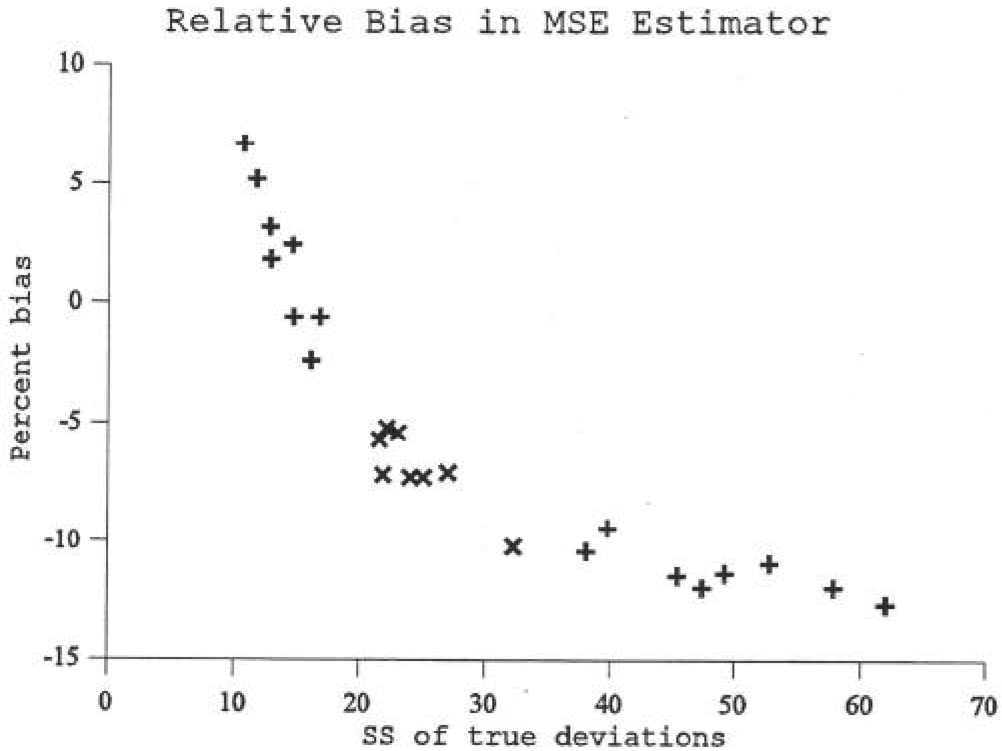


Figure 6 Percent bias in Prasad-Rao MSE estimator, estimated variances, MLE, m = 50, p = 4, 10 obs. per domain. Note: The first 4 points have been omitted.

Figure 7 shows the effect on 25% understatement of the actual MSE when the sampling errors are estimated. Compared to Figure 5, the results are much flatter, in the range of 15-20%, compared to the much more dramatic swings in Figure 5. Unlike Figures 3 and 5, the combination of the extra variability from estimating the sampling variances and the somewhat larger actual MSE eliminates the phenomenon of the dramatic drop towards 0% at the right end of the scale.

It was previously noted that REML applied to the sample data and estimated sampling variances yielded estimates with the largest actual MSE. Specifically, choice 5), with the estimator from Cressie (1992), appears here, although it was previously noted that the alternative 6) produces essentially identical results. Figure 8 shows the bias in the estimated MSE for REML. Figure 8 closely resembles Figure 6 in shape but has estimated biases moved up by roughly 5-10 percentage points. Again, results of this comparison to MLE are consistent with a greater downward bias in the MSE for the latter.

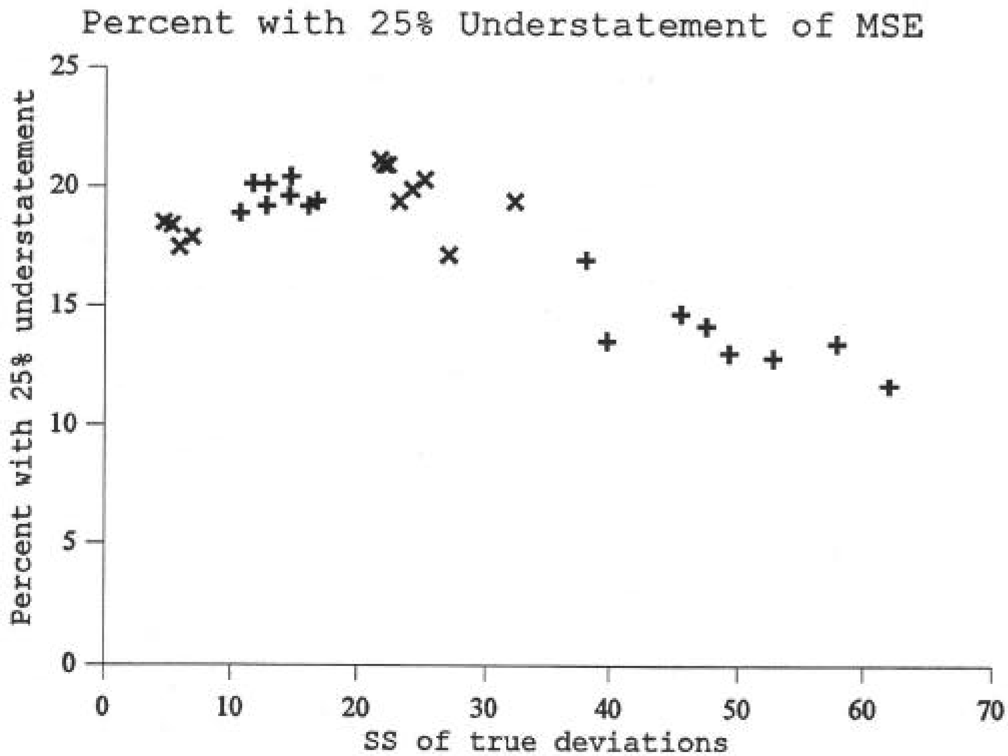


Figure 7 Percent understatement of true MSE by 25 percent or more, estimated variances, MLE, $m = 50$, $p = 4$, 10 obs. per domain.

Figure 9 resembles Figure 7, in showing a flatter performance over the range than Figures 3 and 5. Overall, however, the comparison of Figure 9 to Figure 7 awards a significant advantage to REML compared to MLE in preventing marked understatement of the true MSE. This finding is consistent with the relative shift in bias of the MSE estimators compared in Figures 6 and 8.

As noted earlier, use of sample variances in the method of moments estimator produces an increase in actual MSE comparable to the increase for MLE. Figure 10 shows performance comparable or slightly better than that of REML in Figure 8 under the same circumstances. Again, the MSE estimates exhibit less downward bias than for MLE in Figure 6.

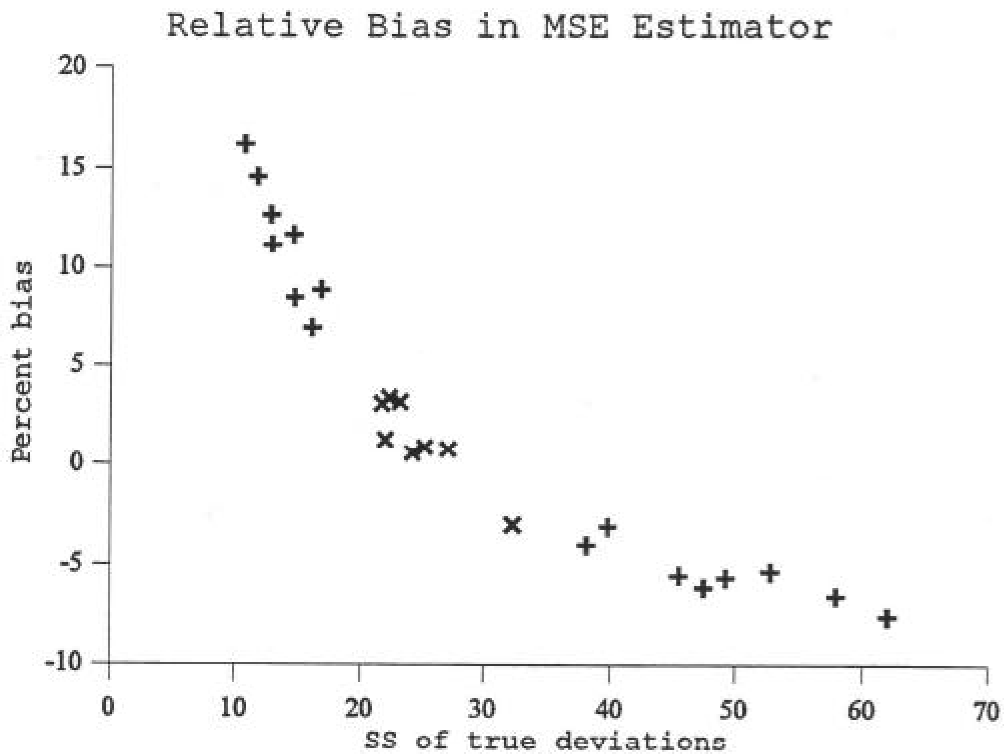


Figure 8 Percent bias in Prasad-Rao MSE estimator, estimated variances, REML, m = 50, p = 4, 10 obs. per domain. Note: The first 4 points have been omitted.

Comparison of Figures 11 and 9 reveals that the slight bias advantage of the method of moments approach compared to REML, shown previously by Figures 10 and 8, is traded against more frequent understatement of the actual MSE by 25% or more. Consequently, there is not a single winner in the contest of these alternatives.

Generally, the method of moments approach does appear to outperform MLE in Figures 6 and 7. The method of moments is subject to less downward bias than MLE at the upper end of the range studied and exhibits less frequent understatement.

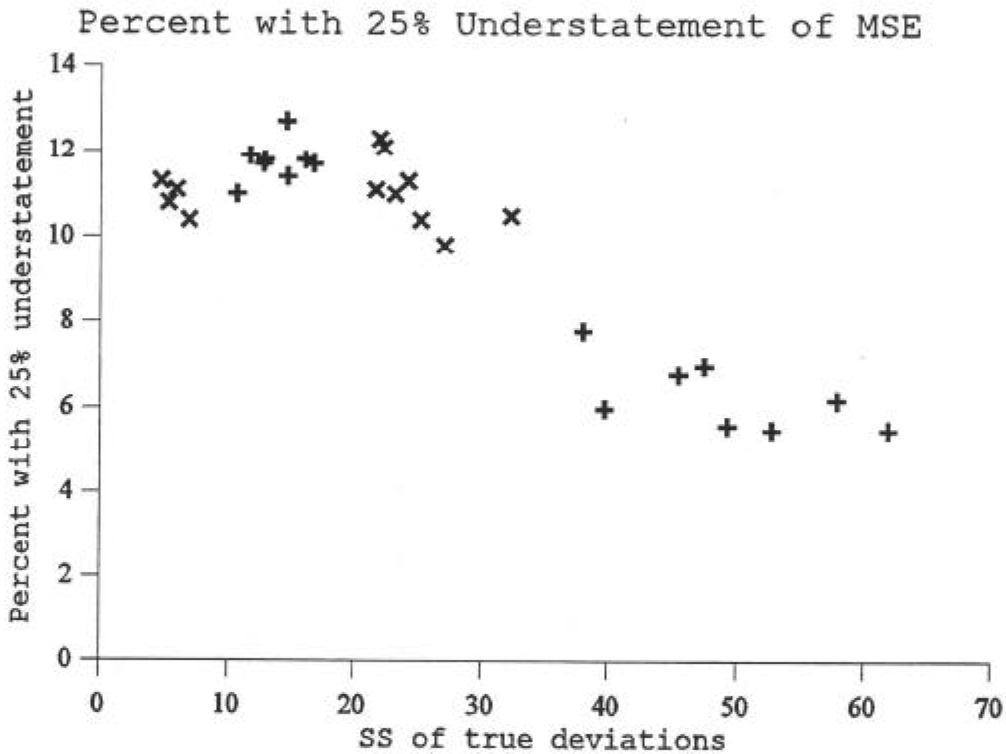


Figure 9 Percent understatement of true MSE by 25 percent or more, estimated variances, REML, m = 50, p = 4, 10 obs. per domain.

Figures 12 and 13 present the results for the last alternative, 8), which weights observations equally in estimating σ_v^2 and which does not require iteration. The findings show a considerable downward bias in MSE estimation under these conditions. For example, comparison of Figure 12 to Figure 10 shows a more consistent downward bias over much of the range studied. In turn, the probability of 25% understatement is higher in Figure 13 than Figure 11.

Generally, the findings show that the properties of the MSE estimators are affected to a significant degree as a result of estimating sampling variances when there are relatively few observations or degrees of freedom in each of the domains. These empirical findings do not appear to be a straightforward consequence of the available theoretical results.

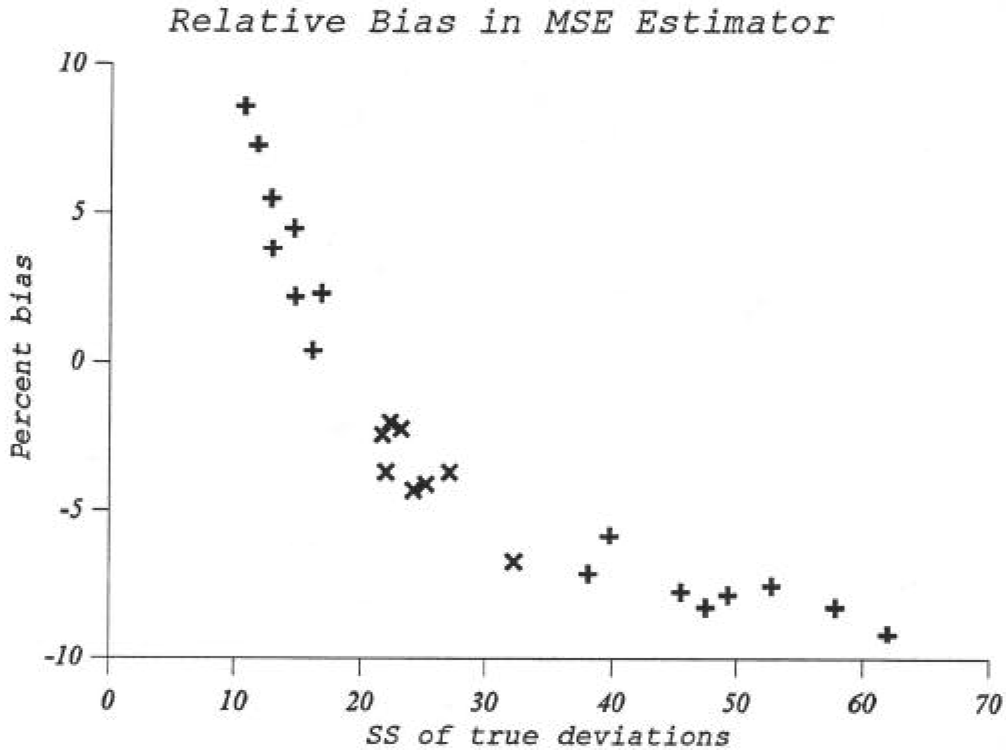


Figure 10 Percent bias in Prasad-Rao MSE estimator, estimated variances, method of moments, $m = 50$, $p = 4$, 10 obs. per domain. Note: The first 4 points have been omitted.

When $n = 20$ observations are instead available for variance estimation within each cluster, the effects of estimating the variances becomes less pronounced. In other words, the corresponding Figures 6 and 7 for $m = 20$ become more like Figures 4 and 5, and the pairs of Figures 8 and 9, 10 and 11, and 12 and 13 each resemble Figures 2 and 3 more closely. Consequently, and not surprisingly, the effect on MSE estimation depends on the degree of precision of the sampling variances in the domain, and not simply on the fact that the sampling variances have been estimated.

Translation of the implication of these results to application will, in the author's opinion, not be simple. Compared to the estimation of variance for standard estimates, such as the sample mean, the issue of the variance of the variance, that is, the design-based variance of a variance estimator, is a fairly arcane subject that has consequently received relatively little attention.

A simple count of the algebraic degrees of freedom will not typically provide an adequate indication of the expected performance of the variance estimator, except in the sense that a variance estimator based on a small number of observations or clusters is certain to be highly variable. Generally, non-normality of the individual or clustered observations may increase the variance of the variance substantially compared to its behavior under normality.

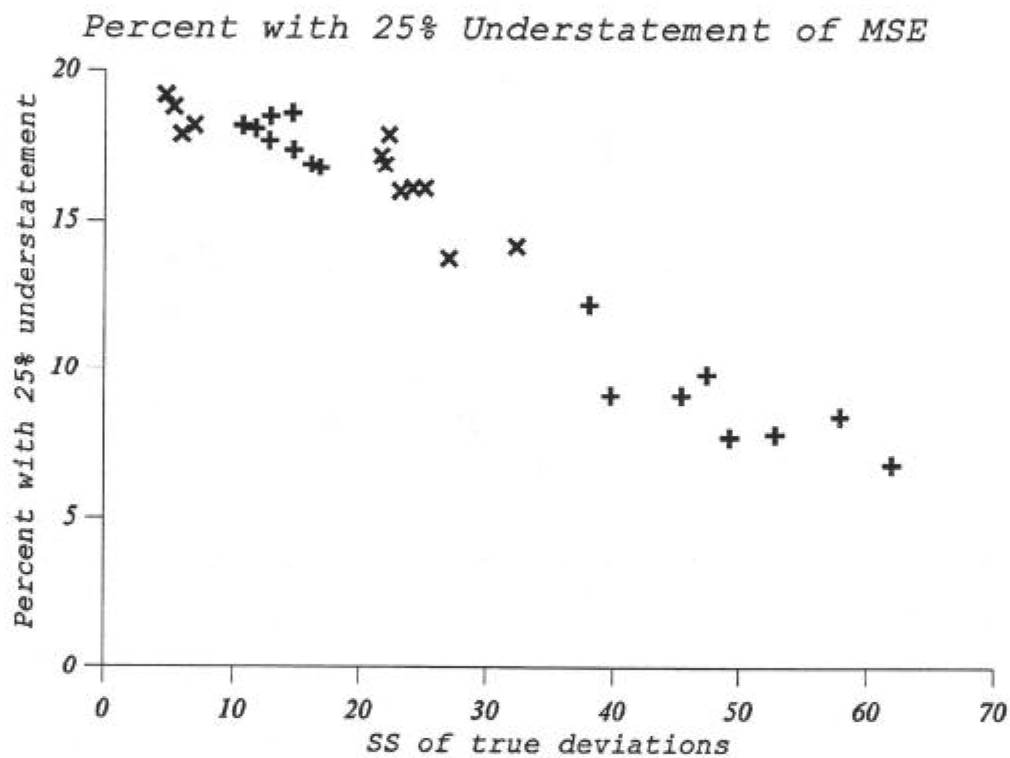


Figure 11 Percent understatement of true MSE by 25 percent or more, estimated variances, method of moments, $m = 50$, $p = 4$, 10 obs. per domain.

Results for $m = 20$ domains follow many of the same patterns as $m = 50$. Overall, however, there is substantially less evidence to evaluate whether the EBLUP has yielded substantial declines in MSE. When $m = 50$, the MSE estimators begin to exhibit relatively extreme behavior, including their upward bias, when the actual reductions are 75% or more. For $m = 20$, the same patterns appear much earlier, at around 50% actual reduction. Similarly, the phenomenon in Figure 3 and others where the MSE estimator suddenly stops overestimating the true MSE by 25% or more shows up much earlier for $m = 20$. Thus, effective MSE estimation in situations where the gains from EBLUP are substantial requires numbers of domains on the order of $m = 50$. Specific findings are available from the author.

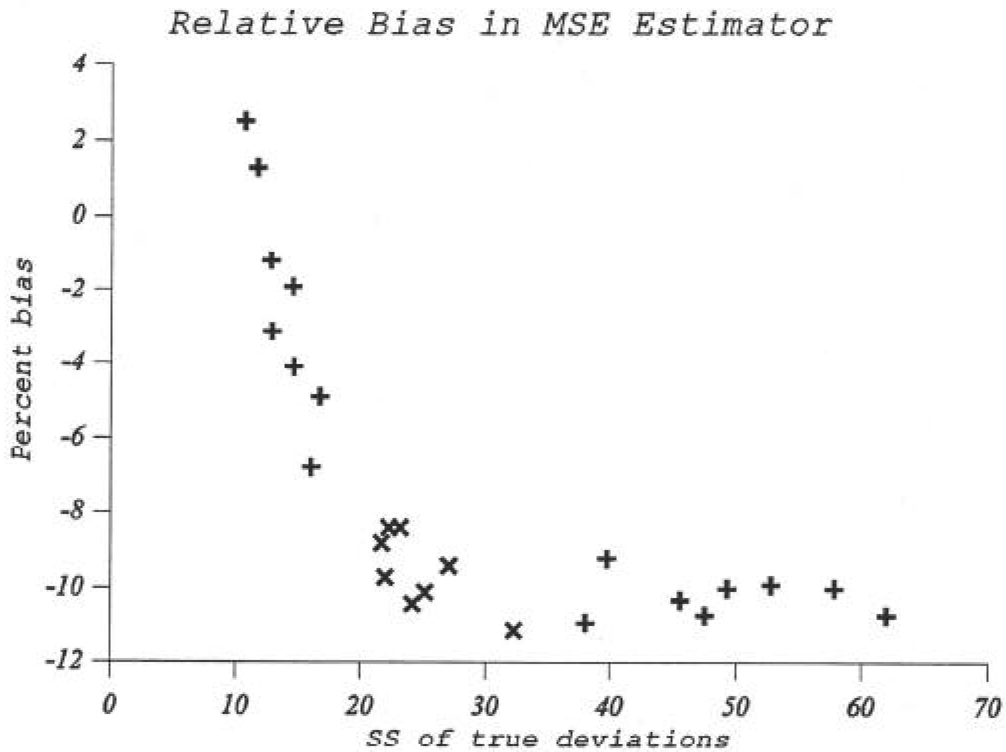


Figure 12 Percent bias in Prasad-Rao MSE estimator, estimated variances, equally weighted method, $m = 50$, $p = 4$, 10 obs. per domain. Note: The first 4 points have been omitted.

Except for separate FORTRAN programs to generate the sample data used in the Monte Carlo study, the variance program VPLX calculated the EBLUP estimators and summarized the results. PC's with 486-class processors performed the calculations for $m = 20$, and a Sun SPARC 10 for $m = 50$, although selective problems were checked against each other to verify independence of results on the choice of platform.

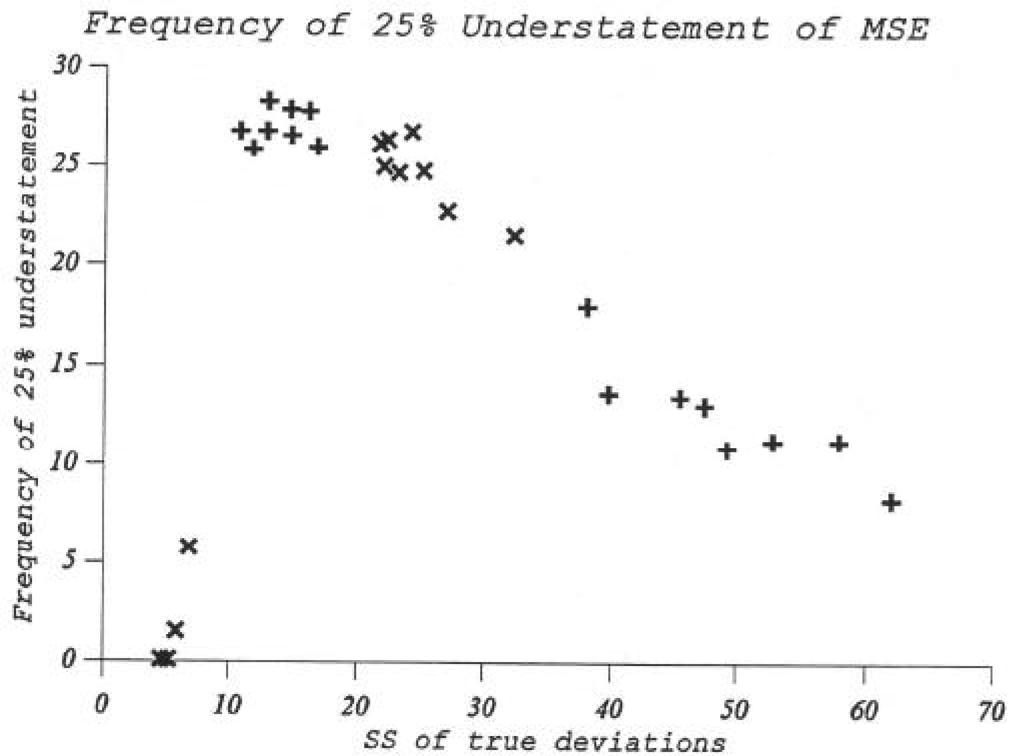


Figure 13 Percent understatement of true MSE by 25 percent or more, estimated variances, equally weighted method, $m = 50$, $p = 4$, 10 obs. per domain.

5. Concluding Remarks

Continued advances in computer technology is certain to have a continued impact on the practice of statistics. Figures 1-13 summarize empirical results that the author would not have had the resources to undertake even a few years ago. Even so, such answers are not yet easily obtained -- for example, each set of points appearing in Figures 1-13 represents about 5 1/2 hrs. of calculation.

The findings, although not generally remarkable, illustrate the subtleties of applying complex estimation methods to practical problems. Features appear that are difficult to anticipate from knowledge of the theoretical results alone. Over time, Monte Carlo assessment should become even more of a standard to complement theoretical findings.

Substantially more work can and should be done. Section 4.1 outlines a general strategy for useful additional study. As examples, the effect of linkage between θ_i and ψ_i can and should be studied in this manner. Variance generalization has appeared in applications, but what are the consequences of applying a deficient model, i.e., a variance generalization that overpredicts some sampling variances and underpredicts others? What are the consequences of misspecifying (2.1)? How should the variance effects of missing data be taken into account? Issues such as these may have a substantial effect on the behavior of EBLUP procedures, and further Monte Carlo work offers an effective approach.

REFERENCES

Cressie, N. (1992), "REML Estimation in Empirical Bayes Smoothing of Census Undercount," *Survey Methodology,* 18, 75-94.

Ericksen, E.P., and Kadane, J.B. (1985), "Estimating the Population in a Census Year (with discussion)," *Journal of the American Statistical Association,* 80, 98-131.

Fay, R.E. (1992), "Inference for Small Domain Estimates from the 1990 Post Enumeration Survey," unpublished manuscript.

Fay, R.E., and Herriot, R.A. (1979), "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," *Journal of the American Statistical Association,* 74, 269-277.

Fay, R.E., Nelson, C.T., and Litow, L. (1993), "Estimation of Median Income for 4-Person Families by State," in "Indirect Estimators in Federal Programs," W. L. Schaible and M. A. Gonzalez (eds.), *Statistical Policy Working Paper 21*, Statistical Policy Office, Office of Management and Budget, pp. 9-1 - 9-17.

Ghosh, M., and Rao, J.N.K. (1994), "Small Area Estimation: An Appraisal (with discussion)," *Statistical Science,* 9, 55-93.

Henderson, C.R. (1950), "Estimation of Genetic Parameters (abstract)," *Annals of Mathematical Statistics,* 21, 309-310.

Kackar, R.N. and Harville, D.A (1984), "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models," *Journal of the American Statistical Association,* 78, 47-59.

Patterson, H.D., and Thompson, R. (1971), "Recovery of Interblock Information When Block Sizes are Unequal," *Biometrika,* **58**, 545-554.

______ (1974), "Maximum Likelihood Estimation of Components of Variance," *Proceedings of the 8th International Biometric Conference,* Washington, D.C., Biometric Society, pp. 197-207.

Prasad, N.G.N. and Rao, J.N.K. (1990), "The Estimation of the Mean Squared Error of Small-Area Estimators," *Journal of the American Statistical Association,* **85**, 163-171.

Schaible, W.L. and Gonzalez, M.A., eds. (1993), "Indirect Estimators in Federal Programs," *Statistical Policy Working Paper 21,* Statistical Policy Office, Office of Management and Budget.

DISCUSSION

Phillip S. Kott
National Agricultural Statistics Service

For a number of years now, many of us in the survey sampling community have been grappling with the following question:

"What is the proper role of models in survey sampling?"

The answer for survey sampling purists can be found in Hansen, Madow, and Tepping (1983). Their Guiding Principle No. 4 states:

"Models are appropriately used to guide and evaluate the design of probability samples [including the choice of estimators], but with large samples the inference should not depend on the model."

This principle clearly justifies the use of model-assisted methods within a randomization-based framework, which is the basis for Särndal, Swensson, and Wretman's celebrated new textbook (1992). It is in sharp contrast, however, to the approach that Hansen and his colleagues label "model-dependent."

Unfortunately, it is not at all clear how Guiding Principle No. 4 applies to the issue of estimation in small domains. In fact, in Guiding Principle No. 7, Hansen, Madow, and Tepping concede:

"... model-dependent methods may have an advantage with quite small samples, for which probability-sampling many not be appropriate"

This suggests that our original question needs to be turned around:

"What is the proper role of randomization-based inference when estimating small domains?"

To Bayesians like Don Malec and Joe Sedransk, the answer to this question is simple: "none." Others, like Bob Fay and myself, would like to estimate the value θ_i in a small domain i with an estimator t_i that has the following property: as the sample size within domain i grows arbitrarily large (but the sampling fraction stays fixed), t_i approaches θ_i in probability irrespective of the validity of the model used in choosing t_i.

We realize, of course, that the sample size within domain i is not arbitrarily large. In fact, in small domain estimation, the sample size within i is usual so small that a conventional model-assisted, randomization-based estimator, $t_{i(rb)}$, has an unacceptably large standard error, hence the need for a more creative, small domain estimator in the first place! Still, we would not be happy using an estimator that did not work well when it should; that is, when the sample size within domain i was large.

One can write

$$t_{i(rb)} = \theta_i + s_i, \tag{1}$$

where s_i is the sampling error of estimator $t_{i(rb)}$. Let us assume that the model-assisted randomization-based estimator $t_{i(rb)}$ is (at least) nearly randomization unbiased so that $E_p(s_i) \approx 0$, where the subscript p denotes that the expectation is with respect to the probability sampling process ("nearly unbiased" means that the bias is small because the sample size <u>across all domains</u> is large). Let us also assume that $\theta_{i(rb)}$ is nearly unbiased under a model governing the elements of the population; that is to say, $E_M(s_i) \approx 0$, where the subscript M denotes that the expectation is with respect to the model. Finally, let us assume that the $t_{i(rb)}$ is randomization consistent; i.e. $\text{plim}_{n(i)\to\infty}(t_i/\theta_i) = 1$, where n(i) is the sample size in domain i.

In small domain estimation, it is common to model the behavior of the domain values θ_i as well as of the population's elements. For convenience, let us restrict our attention to the following domain-level "random effects" model:

$$\theta_i = \mu(\mathbf{x}_i) + \epsilon_i, \tag{2}$$

where $\mathbf{x}_i$ is a vector of characteristics for domain i, μ has a known functional form (e.g., linear or logistic) but unknown parameters, and ϵ_i, the random effect, is a random variable with mean zero and positive variance.

Let m_i be a nearly unbiased estimator for $\mu(\mathbf{x}_i)$. An estimator for t_i of the form:

$$t_i^{(g)} = (1 - g)t_{i(rb)} + gm_i$$

is nearly model unbiased. Its mean squared error is (approximately) minimized when

$$g = \frac{\text{Var}(s_i)}{\text{Var}(s_i) + E[(\theta_i - m_i)^2]} . \tag{3}$$

Whether $\text{Var}(s_i)$ is the model or randomization variance of s_i depends on whether one's goal is to minimize the model or design variance of $t_i^{(g)}$. The same holds true for the interpretation of $E[(t_i - m_i)^2]$.

From both a model and randomization-based perspective, $Var(s_i)$ and $E[(\theta_i - m_i)^2]$ are unknown. Särndal, Swensson, and Wretman's estimator for $Var_p(t_{i(rb)}) \approx Var_p(s_i)$ is also a reasonable estimator for the model variance of s_i. A reasonable estimator for $E_p[(\theta_i - m_i)^2]$ is illusive, but a good estimator for $E_M[(\theta_i - m_i)^2] \approx Var(\epsilon_i)$ is not difficult to develop.

Suppose one estimates $Var_M(s_i)$ and $Var(\epsilon_i)$ from the sample, plugs those estimates into equation (3), and then computes $t_i^{(g)}$. Call the result t_i^*. As the sample size in domain i increases, $Var_M(s_i)$ decreases, while $Var(\epsilon_i)$ remains a positive constant. Thus, as n(i) grows arbitrarily large t_i^* converges to $t_{i(rb)}$, making it randomization consistent just like $t_{i(rb)}$. In fact, t_i^* is fully in the spirit with Hansen, Madow, and Tepping's Guiding Principle No. 4: models have been used in the choice of the estimator, but the estimator itself, while biased, is randomization consistent.

Let us now turn to the primary question addressed in the Malec & Sedransk and Fay papers: how should the variance of a small domain estimator like t_i^* be estimated? Both papers take a model-dependent approach. The problem with this approach, of course, is that models can fail. Since Fay's paper deals with simulations, he avoids the problem. Malec & Sedransk do not.

Malec & Sedransk are to be commended for their thoughtful and thorough work in developing complex models at both the element and domain levels that are appropriate for the survey data they are examining. I have absolutely no problems with the determined parts of these models. What bothers me are the random parts. In particular, the authors build in random effects at the county level only. They allow no additional clustering effects within area segments or households. Moreover, they assume county effects are uncorrelated both across adjacent counties and within states. An example of counties in a state likely to be correlated are Kings, Queens, New York, and Bronx Counties -- the four big boroughs of New York City. I suspect that more than one of these counties are represented in the authors' sample.

It should be noted that the goal of the Malec & Sedransk paper is to produce state not county estimators. Their domain-level model is on the county level, however. Thus, they estimate $\theta_{(state)} = \Sigma_{i \epsilon state} \theta_i$ with $\Sigma_{i \epsilon state} t_{i(MS)}$, where $t_{i(MS)} = m_i$ for counties not represented in the sample. For counties represented in the sample, $t_{i(MS)}$ is similar to the t_i^* discussed above. Nevertheless, because of how the other counties are handled, there is no easy way of modifying a Malec/Sedransk state estimator to make it randomization consistent.

If g were determined from an outside source, the model variance of $t_i^{(g)}$ would be

$$\mathrm{Var}_M(t_i^{(g)}) \approx (1-g)^2\mathrm{Var}_M(s_i) + g^2\mathrm{Var}(\epsilon_i). \qquad (4)$$

Once estimators for $\mathrm{Var}_M(s_i)$ and $\mathrm{Var}(\epsilon_i)$ are computed, an estimator for $\mathrm{Var}_M(t_i^{(g)})$ quickly presents itself.

When a g (approximately) satisfying equation (3) is determined from the sample so that $t_i^{(g)} = t_i^*$, it is tempting to simply plug that value into equation (4) along with estimates of $\mathrm{Var}_M(s_i)$ and $\mathrm{Var}(\epsilon_i)$. A good deal of high powered statistical work has gone into showing why such a practice can be mistaken. I have a more prosaic problem with this approach to variance estimation: it relies entirely on the truth of the model; in particular, on the model for the ϵ_i. It is true that we modeled the ϵ_i in developing the estimator t_i^* in the first place, but to my mind this fact only reinforces a need to be able to evaluate the accuracy of t_i^* in a way that does not require the same model assumptions.

The randomization mean squared error of $t_i^{(g)}$ is

$$\mathrm{MSE}_p(t_i^{(g)}) \approx (1-g)^2\mathrm{Var}_p(s_i) + g^2E_p[(\theta_i - m_i)^2].$$

Let $v(s_i)$ be a randomization-based estimator for $\mathrm{Var}_p(s_i)$. One can estimate $E_p[(\theta_i - m_i)^2]$ with $(t_{i(rb)} - m_i)^2 - v(s_i)$. Unfortunately, this estimator is dreadfully unstable. It has, at most, 1 degree of freedom. For many domains, $v(s_i)$ will also be very unstable, since it has, at most, n(i) - 1 degrees of freedom.

It may come as a shock, but few users of our statistics are all that concerned with variances. With this is mind, perhaps we should abandon the search for a near perfect variance estimator for t_i^*. We do need to be assured that t_i^* has some minimum degree of accuracy. One possibility is to model $v_p(s_i)$ and $(t_{i(rb)} - m_i)^2 - v(s_i)$ across all the domains and to use the results to derive a conservative indication about the accuracy of t_i^* for each i.

REFERENCES

Hansen, M., Madow, W. and Tepping, B. (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inference in Sample Surveys (with discussion)," Journal of the American Statistical Association, 776-807.

Särndal, C. Swensson, B. and Wretman, J. (1992), Model Assisted Survey Sampling. New York: Springer-Verlag.

DISCUSSION OF SMALL AREA ESTIMATION PAPERS
COPAFS CONFERENCE, MAY 26, 1994

David A. Marker
Westat, Inc.

Both of these papers are important for their general approach to the problem of small area estimation: they attempt to understand the application of new methods through the explicit use of models. Ideally, one would always design surveys to allow for the production of accurate, direct, design-based estimates. However, when such estimates cannot be produced, one is left with only two choices: either don't produce estimates or use models.

Malec and Sedransk present the use of hierarchical Bayes procedures for small area estimation. I find this approach to be more satisfactory than empirical Bayes procedures for at least three reasons. First, hierarchical Bayes procedures do not assume a particular model to be true. To quote George Box, "All models are wrong, but some are useful." Second, by assuming instead that the truth comes from within a class of prior distributions, it is possible to examine the robustness of the estimates; although this is limited to the range of priors contained in the class. Third, hierarchical Bayes allows for the use of informative priors. While Malec and Sedransk do not make use of informative priors, this is a possible area for extending their results. Particularly for repeated surveys such as the NHIS, there is a wealth of historical data that can be used. These data can be incorporated for model selection, as variables in the actual model, or to construct informative priors.

Many authors, including Malec and Sedransk, use the Gibbs sampler to produce hierarchical Bayes estimates. The advantage of the Gibbs sampler is that it allows for computations from complex distributions. However, the experiences relayed by Malec and Sedransk and others indicate that this approach is extremely time intensive, in some cases taking months to produce stable estimates. This raises questions about the actual utility of this approach to produce timely small area estimates.

One additional point is worth making regarding the Gibbs sampler. As mentioned earlier there is an abundance of historical NHIS data from which informative priors could be developed. It would be very interesting to see the results of using the Gibbs sampler when beginning with informative, rather than uninformative, priors.

Malec and Sedransk develop their model using forward stepwise regression. While this is a reasonable approach, it can lead to suboptimal results under complex situations. Therefore, it might be worthwhile to examine alternative model-selection methods. In selecting their model, they disregarded the sampling weights. They reported that the weights would not have had significant impact based on analyses at the national level. My concern is that given state-to-state differences this might not imply that nothing is lost by disregarding weights when producing state estimates.

Fay uses simulation to examine the real situation of computing the accuracy of small area estimates when the variances are unknown. The Prasad and Rao approach that he evaluates is limited to situations in which the mean and variance are independent. Unfortunately in many situations, including the binomial variable used by Malec and Sedransk, this is not true. Prasad and Rao developed a procedure for producing approximately unbiased mean square errors (MSEs) for model-dependent small area estimates. These MSE estimates are, however, conditional on the model.

For government agencies there is a strong interest in producing design-based measures of accuracy, not ones conditional on models. A method for producing design-based small area specific MSEs was introduced by Marker (1993). This approach replaces the average MSE of Gonzalez and Waksberg with a small area specific MSE, where the variance of the model-dependent estimator is computed for each small area i using replicated methods (jackknife or balanced repeated replication). The bias is computed by averaging across small areas.

$$MSE\,(\hat{y}_i) \quad = \quad var\,(\hat{y}_i) \; + \; avebias^2\,(\hat{y}_i)$$

where

$$avebias^2\,(\hat{y}_i) \quad = \quad aveMSE\,(\hat{y}_i) \; - \; avevar\,(\hat{y}_i)$$

This estimator is not completely small area specific, but if the variance term dominates the bias, the root mean square error will provide a useful substitute for the traditional standard error. If the bias term dominates, the small areas can be grouped by expected similar biases. The average bias can then be computed separately for each group of areas so that the MSE more accurately reflects small area differences. It would be very useful if both Fay and Malec and Sedransk could examine the utility of this approach.

Reference:

Marker, David A., "*Small Area Estimation for the National Health Interview Survey*," Proceedings of the American Statistical Association Section on Survey Research Methods, 1993.

Session 12

NONRESPONSE IN SURVEYS

EXPLORING NONRESPONSE IN U.S.FEDERAL SURVEYS

Maria Gonzalez, OMB; Dan Kasprzyk, NCES; and
Fritz Scheuren, IRS

Section 1: Introduction

This paper is intended to provide a broad summary of nonresponse rate trends in U.S. federal government surveys. We have built directly on the work of a Subcommittee on Survey Nonresponse, commissioned in 1991, by the Office of Management and Budget's Federal Committee on Statistical Methodology (FCSM). A particular debt of gratitude needs to be acknowledged for the role played by Bob Groves (Subcommittee Chair), Mick Couper and the other members of that Subcommittee for their input into what follows (see acknowledgements for a full list of the members).

Highlights of the Subcommittee's efforts have already appeared in the April AMSTAT NEWS (Gonzalez, Kasprzyk, and Scheuren,1994). A more extended treatment will be given in this paper. Still other papers based on the Subcommittee's work will appear in the Proceedings of the 1994 meetings of the ASA.

The present material is organized into four main sections, along with supporting figures, references, acknowledgements and an afterword. First, there is this short Introduction (Section 1); some background considerations come next. These considerations led to the establishment of the FCSM Nonresponse Subcommittee (Section 2).

In Section 3, an overview of the work of the Subcommittee is given, including the principal findings on nonresponse rate trends in federal surveys. Naturally, a discussion is given of limitations as well.

Finally, the recommendations of the Subcommittee are revisited in Section 4 and comments made on the future steps that we, as federal government statisticians, should take -- both as individuals working on our own surveys and by acting collectively to improve practice as a whole.

Section 2: Some Background on Nonresponse in Federal Surveys

Like the poor, nonresponse in surveys may always be with us. In the days of "representative" samples drawn purposively, nonresponse was present but not visible. (Quota sampling, even today, makes measuring the extent of the actual nonresponse difficult -- maybe impossible). With the ascent of the random sampling paradigm (Bellhouse,1988), nonresponse became a problem that needed to be "solved."

In so far as U.S. Federal surveys are concerned, the turning point in government practice for the randomization paradigm came when Deming invited Neyman to lecture at the U.S.D.A. Graduate School in 1937. Morris Hansen, using Neyman's ideas and his own, and with many collaborators, did the rest (e.g., Hansen, Hurwitz, and Madow,1953).

It seems clear that Hansen and the other early pioneers understood quite well that randomization-based inference was directly challenged by nonresponse. Concerns about bias, for example, were evident from the beginning. In Cochran(1977) there is an example of an early treatment that simply widens the confidence intervals directly to account for the nonresponse bias. This conservative approach was consistent with the main focus of the random samplers of that era who were busy inventing ways to reduce nonresponse to the bare minimum. The U.S.Census Bureau in its Current Population Survey (e.g., Hanson,1978) still continues successfully in that tradition.

Hansen and his collaborators, in addition to a primary emphasis on "prevention," developed designs which called for the subsampling of nonrespondents (e.g., Hansen and

Hurwitz,1946). These were a natural extension of the basic randomization paradigm and called for more thorough fieldwork on a random subsample of nonrespondents. One of the results of this work was to introduce the idea of a weighted response rate. Such samples naturally also had their own nonresponse problems; so this approach too was seen from the beginning as only a partial one. Post-survey adjustment techniques to compensate for flaws in the randomization due to nonresponse were also attacked as well.

For those interested in more information, a special September 1975 issue of the Journal of the American Statistical Association is a recommended reference (Gonzalez, Ogus, Shapiro, and Tepping, 1975). This article provides a useful summary of federal goverment (largely Census Bureau) practices on the reporting of sampling and nonsampling errors, including nonresponse (see also Duncan and Shelton (1978) for still more on the history of sampling in U.S. Federal surveys).

While nonresponse in federal surveys has always been said to be an indicator of the quality of survey data, interest and concern has grown during the last two decades:

- The Panel on Incomplete Data, established by the Committee on National Statistics in 1977, produced three volumes focussing on incomplete data in sample surveys (Madow, Nisselson, Olkin, and Rubin, 1983).

- The Council of American Survey Organizations (CASRO) reviewed response rate definitions with the intent of trying to establish uniformity of definitions across surveys (CASRO, 1982).

- Steeh (1981) and Groves (1989) reviewed trends in the response rates in nongovernment surveys, indicating a decline in response rates over time.

- During the last ten years, the tight federal budget climate has prompted questions about the ability of federal statistical agencies to maintain high response rates with a constant budget.

Theoretical developments in the handling of nonresponse have grown enormously since the mid-1970's. Indeed, the problem has drawn the attention of some of the best statisticians now working on surveys. The National Academy Panel's report on Incomplete Data(1983) was a culmination of sorts. A review of nonresponse adjustment techniques was done by Kalton (1983). Even so, in the ten years since the Panel's report, there has been a lot more done and no end is in sight. The book on nonresponse by Little and Rubin (1986) and a separate book by Rubin (1987) on multiple imputation are perhaps the two most prominent examples of the important work that continues. The treatment of Sarndal et al (1992) and Lessler and Kalsbeek (1992) also are valuable for the way, among other things, they place nonresponse in context of total survey error.

Within this general environment of greater interest in nonresponse, the FCSM decided to sponsor an effort to learn what was known about nonresponse as a source of bias in federal survey estimates. Prominent factors in making this decision were --

- The lack of a systematic review of the topic since the 1983 Committee on National Statistics report.

- A growing perception among the members of the federal statistical community that nonresponse in federal surveys had been increasing over time.

In any event, in 1991 a Subcommittee of the FCSM was formed to study nonresponse in federal surveys. The initial charge of the Subcommittee was, simply stated, to "begin an effort to better understand unit nonresponse in surveys." The proposed approach was to conduct a broad-based review of the level of unit nonresponse rates, currently and over time, in federal surveys. The details of the Subcommittee's work are covered in the next Section.

Section 3: Work of FCSM Subcommittee on Survey Nonresponse

The Subcommittee was specifically charged with the mission to investigate for Federal surveys the levels of response rates, the measures used to compute these response rates, response trends from 1982-1991, perceived correlates of nonresponse, and other related information.

In carrying out its mission, the Subcommittee obtained information from 26 demographic and 21 establishment surveys. These surveys were not selected by probability methods, because no machine-readable listing of Federal surveys with sufficient auxiliary information for appropriate stratification was available. The 47 surveys were chosen, however, to include Federal surveys that differed on a number of key design parameters: those conducted on an ongoing or an intermittent basis, those conducted by Federal agencies, and those carried out by contractors under Federal auspices.

Because of the large differences in the design of surveys to collect establishment versus demographic data, separate questionnaires were constructed for each type and sent to respective survey sponsor or data collection agency. The intent of both questionnaires was to elicit information on a variety of survey features that earlier literature has shown to affect nonresponse. In addition, information was sought on strategies for post-survey adjustment for nonresponse.

The Study itself incurred no unit nonresponse but did incur a small amount of item nonresponse in its data collection activities. Indeed, it was difficult to get the agencies to respond to the nonresponse questionnaire.

The findings of the Subcommittee span the range from the expected to the surprising. As in any research undertaking, of course, the conclusions drawn from an analysis of the questionnaires should be treated with caution. This point is particularly well taken here given the purposive nature of the sample, the small number of surveys included in the data collection, and the wide variety of design differences that characterize these surveys. Some highlights follow.

Trends in Nonresponse Rates. Despite the prior beliefs of many in the Federal survey community, there was little evidence of declining response rates over time for either the establishment or demographic surveys included in the Subcommittee's study:

- Establishment Surveys. To analyze the response rates of establishment surveys over time, it is more meaningful to limit the analysis to those surveys which reported response rates for several years. For this reason, the analysis of time trends in the response rates for establishment surveys cover only the nine surveys for which both weighted and unweighted data were available for at least six reporting periods between 1981-1991.

Figure 1 shows the average weighted response rate for the nine selected surveys. Figures 1-5 are based on work of the FCSM Subcommittee on Survey Nonresponse. As may be seen, the weighted response rate was only slightly decreasing over the period covered by the data. The average decrease was about 1/4 percent per year. Figure 1 also shows the mean unweighted response rate for the selected nine surveys from 1984-1990. The unweighted rate was slightly increasing, but stable over the period. The average increase was about 1/2 percent per year.

Figure 2 shows weighted response rates for the nine establishment surveys. Five of these weighted response rates are 90 percent or above. Two series have a weighted response rate between 70-90 percent and two series are around 50 percent. More about establishment trends in response will be said in Osmint, McMahon, and Martin.

- Demographic Surveys. Most demographic surveys used unweighted response rates rather than weighted rates for routine monitoring of the data collection process and so we have followed this convention here as well. The analysis of trends over time for demographic surveys was restricted to those surveys with at least 4 data points in the period 1982 to 1991. Only 8 of the 26 demographic surveys included in our data collection met this criterion.

  The mean nonresponse rate by year was calculated for these eight surveys from the data provided, along with refusal rates and noncontact rates where available. Although the stimulus for the creation of the Subcommittee was the belief that response rates were declining over time in demographic surveys, Figure 3 does not support that belief. The mean nonresponse rates for the surveys included in the sample are minimally lower in 1991 than in 1982. This figure shows that refusal rates, a major component of nonresponse, have remained about the same. More about demographic trends in response will be said in Johnson, Botman, and Basiotis.

For the Current Population Survey a longer time series of data is available. Figure 4 shows that the level of nonresponse has been stable for some years. Since the refusal rates seem to have increased, a possible implication is that more effort may have been made to reduce other nonresponse components -- so as to achieve relatively constant overall response rates.

An examination of response rates for the more-frequently fielded demographic surveys reveals large variations across surveys. This variation can be partially understood by separating the studies into two groups (see Figures 5). One group has response rates in the 95 percent range, while a second cluster lies about 10 percentage points lower. The studies in the 95 percent range consist of ongoing studies conducted by the same interviewer corps. The studies in the lower group tend to be less frequently conducted. Neither group exhibits strong trends over time.

In summary, despite the prior beliefs of many in the survey community, there was little evidence of declining response rates over time among either the establishment or demographic surveys included in the study. This could be due to a greater effort in data collection but technological and other survey context changes make this hard to verify. One final point about these results may be worth making again: There were only a limited set of surveys on which time trends can be measured--just nine establishment surveys and eight demographic surveys.

<u>Other Findings</u>.-- There are other findings from the Subcommittee's work; but only three are highlighted here. These involved issues in the definition of nonresponse, response rate documentation, and post-survey adjustment methods:

- Definitions for Nonresponse. Despite the study's focus on nonresponse rates and despite having contacts in the agencies, major difficulties arose in obtaining consistent information. Just as was found in an early (albeit more general) study,..."rates have different names and different definitions in different places and times." (Bailar and Lanphier, 1978) This issue led, in part, to one of the study's major recommendations (see figure 6, Subcommittee Recommendation 3).

- Response Rate Documentation. Reporting practices for documenting response rate components varied widely across the surveys in the study. Common practice in establishment surveys is in contrast to common practice in demographic surveys. Sponsors of demographic surveys not only were more likely to maintain records regarding a wider variety of nonresponse components but also tended to maintain more historical information. For example, all of the demographic surveys in our data collection included some information about response/nonresponse components. In contrast, for the establishment surveys analyzed, 10 out of 21 did not track any nonresponse components.

- Post-survey Nonresponse Adjustment. Respondents were asked about a number of post-survey adjustment techniques designed to reduce the effects of nonresponse: post-stratification (e.g., simple ratio or raking ratio adjustment), regression modelling of the propensity to respond, and imputation. All surveys in the Subcommittee study used some degree of post-survey nonresponse adjustment. Some of the approaches were very traditional, while others reflected more recent research on estimation strategies.

In the remaining section of this paper we cover the Subcommittee's recommendations and a few ideas on future steps.

<u>Section 4: Some Next Steps for Practice</u>

The Subcommittee made four recommendations that are given in detail in figure 6. Stated briefly the subcommittee recommended:

- Survey practitioners should compute nonresponse rates in a uniform fashion over time.

- In repeated surveys, response rate components should be monitored in conjunction with cost and design changes.

- Agencies that sponsor surveys should publish how they compute response rate and their components in survey reports and their relevance to the quality of the survey results discussed.

- Ongoing research should be conducted on nonresponse adjustment variables, costs and benefits of converting refusals, and similar nonresponse management concerns.

All of these recommendations seem rather obvious. They address some very basic survey management and reporting requirements; furthermore, the suggestions are close in spirit and substance to those made by other groups over the last two decades.

It is true that every survey program examined by the Subcommittee calculated nonresponse rates in some fashion and had some auxiliary information about aspects of nonresponse. It is also true that most survey programs did not have **readily** available what the Subcommittee viewed as "basic" data on nonresponse; nor did repeated surveys have a time series easily available of nonresponse rates and nonresponse components.

What can we expect for the future based on the results of this small exploratory study? Some conjectures follow:

- First, it is unrealistic to assume that the recommendations by yet another subcommittee will be adopted uniformly by the agencies of the Federal statistical system.
- Second, unless mandated, individual survey program managers are likely to remain individualistic and independent with respect to their acceptance and adoption of recommendations concerning their surveys.
- Third, it is important to recognize the diversity of the management of individual survey programs and build on each survey programs' strengths. In other words, these recommendations should be no more than guidelines in any case.
- Fourth, the survey data collection manager and the agency that sponsors the survey need to work together as a team with the interests of the ultimate customer paramount -- recognizing that an information system producing data on nonresponse and its components is mutually beneficial.

One need not assume the points made above are necessarily pessimistic. The underlying theme is the development of a fully professional partnership among data collection managers, agencies that sponsor survey programs, and ultimate customers. Mutual respect and understanding for each other's requirements (given budget constraints) is essential for improving the reporting of nonresponse and nonresponse components. The current theme of "reinventing government" speaks well to the prospect of improving these professional relationships through its team building and customer orientation emphases.

Finally, we can expect incremental improvements in the issues discussed here through the continuing work of the Federal Committee on Statistical Methodology (Gonzalez, 1994) and the National Science Foundation initiated "Program in Survey Methodology" offered by the consortium of the University of Maryland, University of Michigan, and Westat. Both programs are dedicated to the improvement of the quality of Federal survey data. Through these efforts and the individuals involved in Federal data collection programs, progress will be made.

Most of this paper looks inward at the federal statistical system. Obviously, much can be learned by examining private sector experiences and through international comparisons. The companion paper at this session by David Binder and his colleagues from Statistics Canada is an example of what we have in mind. Clearly, too, the Statistics Canada approach to nonresponse rates is worthy of further study by those interested in this area (see Statistics Canada, 1993 and Hidiroglou, Drew, and Gray, 1993)

In the spirit of "reinvention," a systematic benchmarking approach is needed. Some important beginning efforts that bear mention in this regard include the papers by Lyberg and Dean (1992) and Christianson and Tortora (1993).

Acknowledgements and Afterwords

This paper is based on the work of the Subcommittee on Survey Nonresponse whose members included: Robert M. Groves, Chair, BOC, Susan Ahmed, NCES, J. Donald Allen, NASS, David Belli, BEA, Peter Basiotis, HNIS, Steve Botman, NCHS, Eileen Collins, NSF, Mick Couper, BOC, Patricia Guenther, HNIS, Paul Hsen, BLS, Ayah Johnson, AHCPR, Arthur Kennickell, FRB, Paul McMahon, IRS, Jeffrey Osmint, BoM, Antoinette Martin, EIA, Pamela Powell-Hill, BOC, Maria Reed, BOC, Carolyn Shettle, NSF.

One final point may be made about the exploratory nonresponse Study focussed on in this paper. We would like to engage in a dialogue about this study and its recommendations with those who conduct both Federal and non-Federal surveys. Please contact Maria E. Gonzalez (OMB, Statistical Policy Office, NEOB, Room 10201, Washington, D.C. 20503) with your comments or experiences in this field.

References

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

Bellhouse, David (1988). "A Brief History of Random Sampling Methods," Handbook of Statistics, Vol 6, North-Holland, Amsterdam, pp. 1-14.

Christianson, Anders and Robert D. Tortora (In Press). "Issues in Surveying Establishments: An International Survey," in BG Cox et al (Editors), Survey Methods for Businesses, Farms, and Institutions, John Wiley & Sons. Presented at the International Conference on Establishment Surveys, 1993 Buffalo, NY.

Council of American Survey Organizations (CASRO) (1982). On the Definitions of Response Rates, Port Jefferson, New York.

Duncan, J.W. and Shelton, W.C. (1978). Revolution in United States Government Statistics 1926-1976, U.S. Department of Commerce, Office of Federal Statistical Policy and Standards.

Gonzalez, M.E. (1994). "Improving Data Awareness in the United States Statistical Agencies," The American Statistician. American Statistical Association. Vol. 48, No. 1.

Gonzalez, M.E., Kasprzyk, D., and Scheuren, F. (1994). Nonresponse in Federal Surveys, AMSTAT News, April.

Gonzalez, M.E., Ogus, J.L., Shapiro, G., and Tepping, B.J. (1975). Standards for the Discussion and Presentation of Errors in Census and Survey Data, Journal of the American Statistical Association, 70, 351, Part II.

Groves, Robert M. (1989). Survey Errors and Survey Costs, John Wiley and Sons, New York.

Hansen, M., Hurwitz, W., and Madow, W. (1953). Sample Survey Methods and Theory (2 Volumes), John Wiley: New York.

Hansen, M. and Hurwitz, W. (1943). On the Theory of sampling from finite populations. Ann. Math. Statist. 14, 333-362.

Hanson, R. (1978). The Current Population Survey--Design and Methodology, Technical Paper 40, U.S. Bureau of the Census.

Hidiroglou, M.A., Drew, J.D., and G.B. Gray (1993). "A Framework for Measuring and Reducing Nonresponse in Surveys," Survey Methodology, Vol. 19, No. 1, pp. 81-94.

Johnson, Ayah, Steve Botman, and Peter Basiotis (forthcoming). "Nonresponse in Federal Demographic Surveys: 1981-1991." To be presented at the 1994 annual meetings of the American Statistical Association, Section on Survey Research Methods.

Kalton, G. (1983). Compensating for Missing Survey Data, Research Report Series, Ann Arbor, Michigan: Institute for Survey Research.

Lessler, J. and Kalsbeek, W. (1992). Nonsampling Errors in Surveys, John Wiley & Sons: New York.

Little, R. and Rubin, D. (1986). Statistical Analysis and Missing Data, John Wiley & Sons: New York.

Lyberg, L. and Dean P. (1992). "Methods for Reducing Nonresponse Rates: A Review," presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg, FL.

Madow, W.G., Nisselson, H., Olkin I., and Rubin D.B. (Editors) (1983). Incomplete Data in Sample Surveys, 3 Volumes, Academic Press, NY.

Osmint, Jeffrey, Paul B. McMahon, and Antoinette Ware Martin (forthcoming). "Response in Federally Sponsored Establishment Surveys." To be presented at the 1994 annual meetings of the American Statistical Association, Section on Survey Research Methods.

Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons: New York.

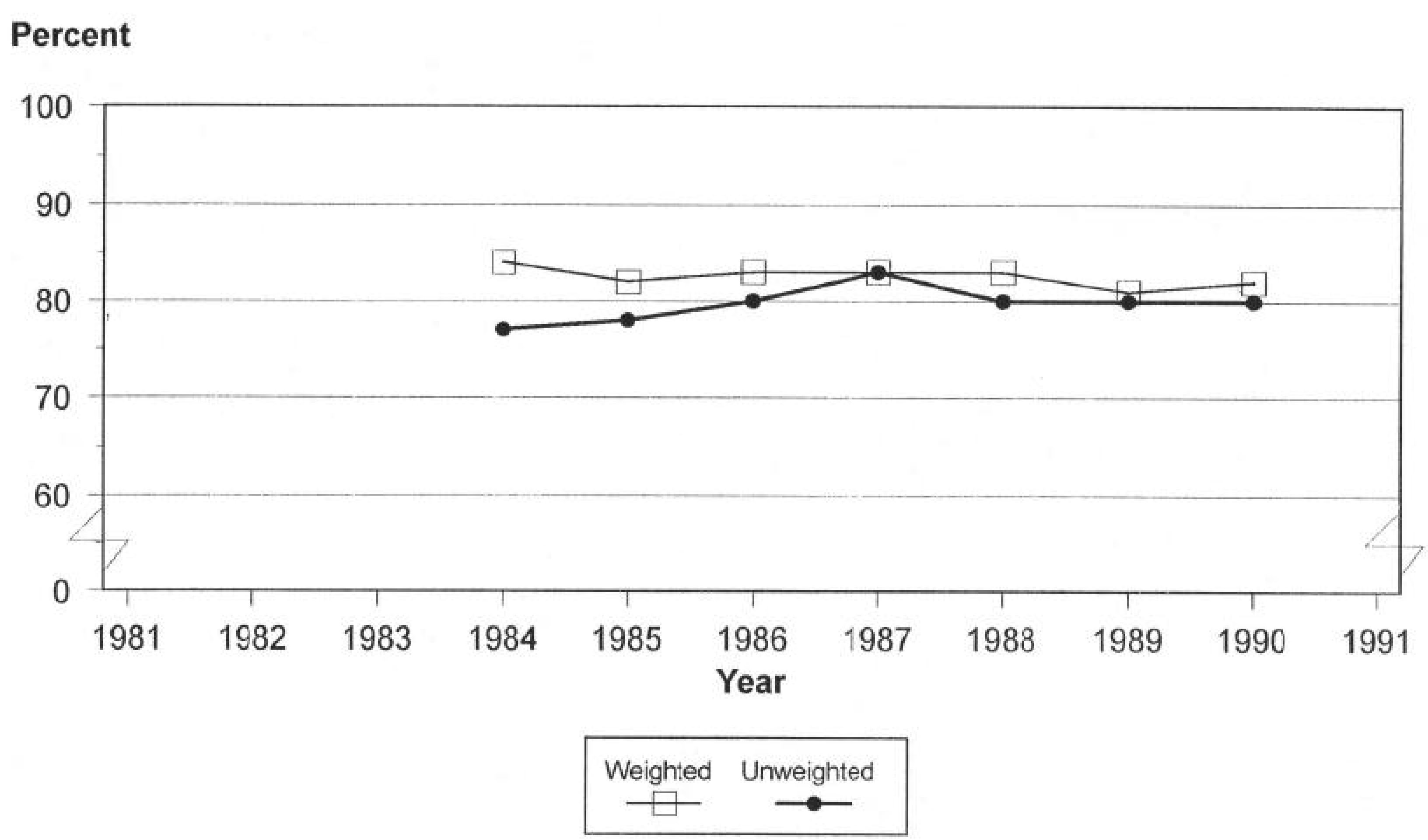
Sarndal, C.E., Swenssen, B., and Wretman, J. (1992). Model Assisted Survey Sampling, Springer-Verlag: New York.

Shettle, Carolyn F., Patricia M. Guenther, Daniel Kasprzyk, and Maria Elena Gonzalez (forthcoming). "Investigating Nonresponse in Federal Surveys." To be presented at the 1994 annual meetings of the American Statistical Association, Section on Survey Research Methods.

Statistics Canada (1993). Standards and Guidelines for Reporting of Nonresponse Rates, Ottawa, Canada.

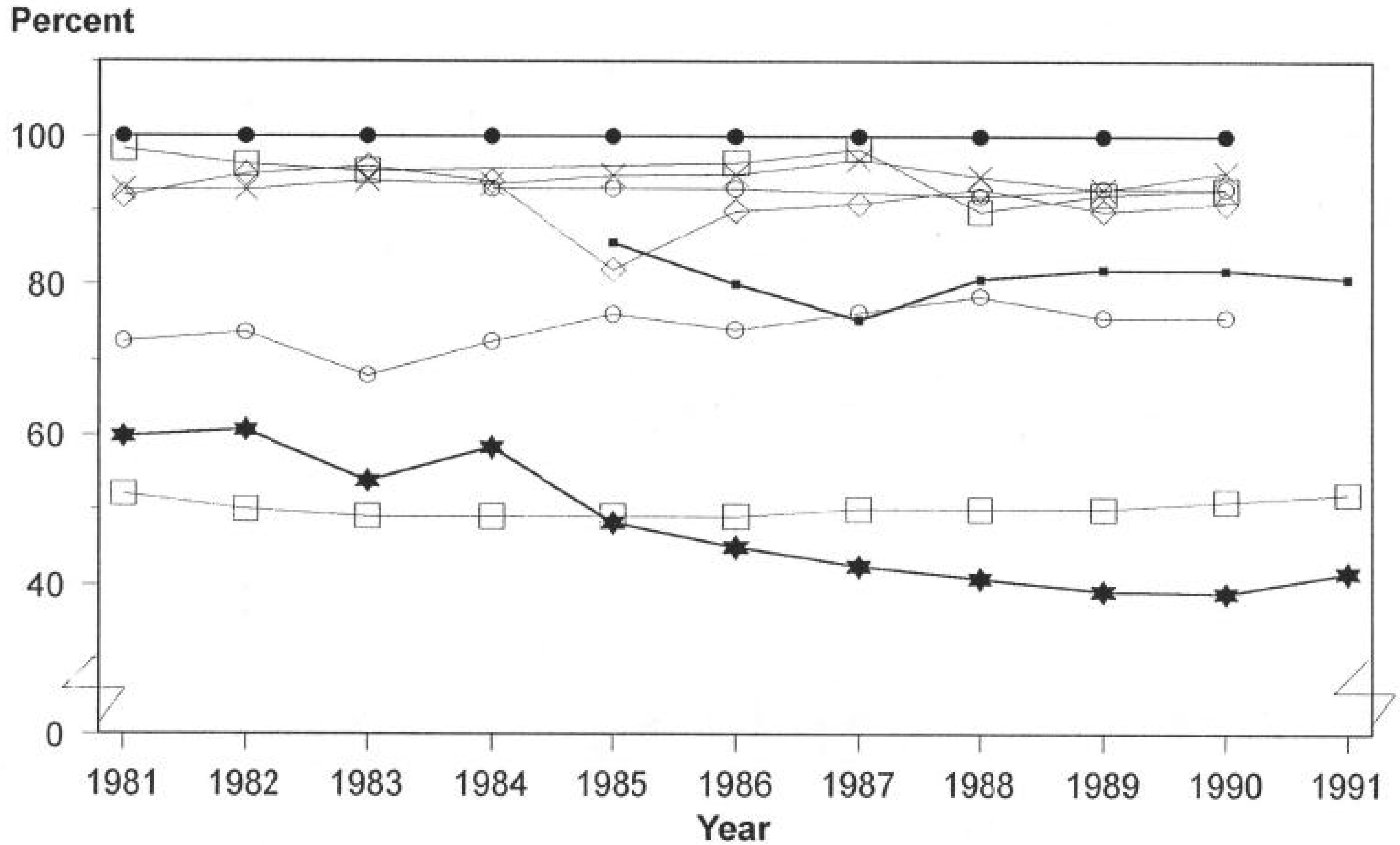
Steeh, C. (1981). "Trends in Nonresponse Rates," Public Opinion Quarterly, Vol. 45, pp.40-57.

Figure 1.
Mean Response Rates, Economic Surveys



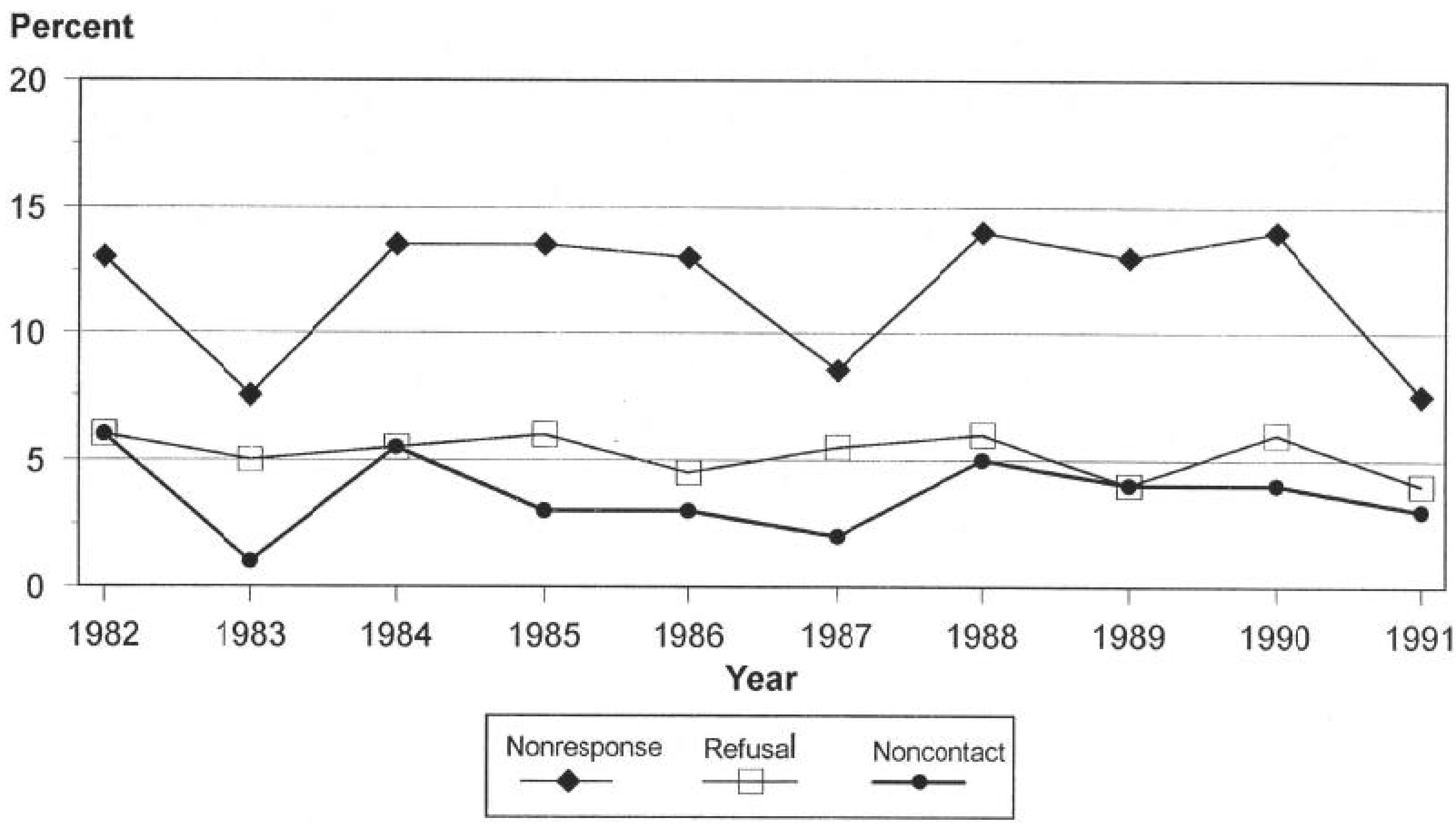
SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

Figure 2.
Weighted Response Rates, Economic Surveys



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

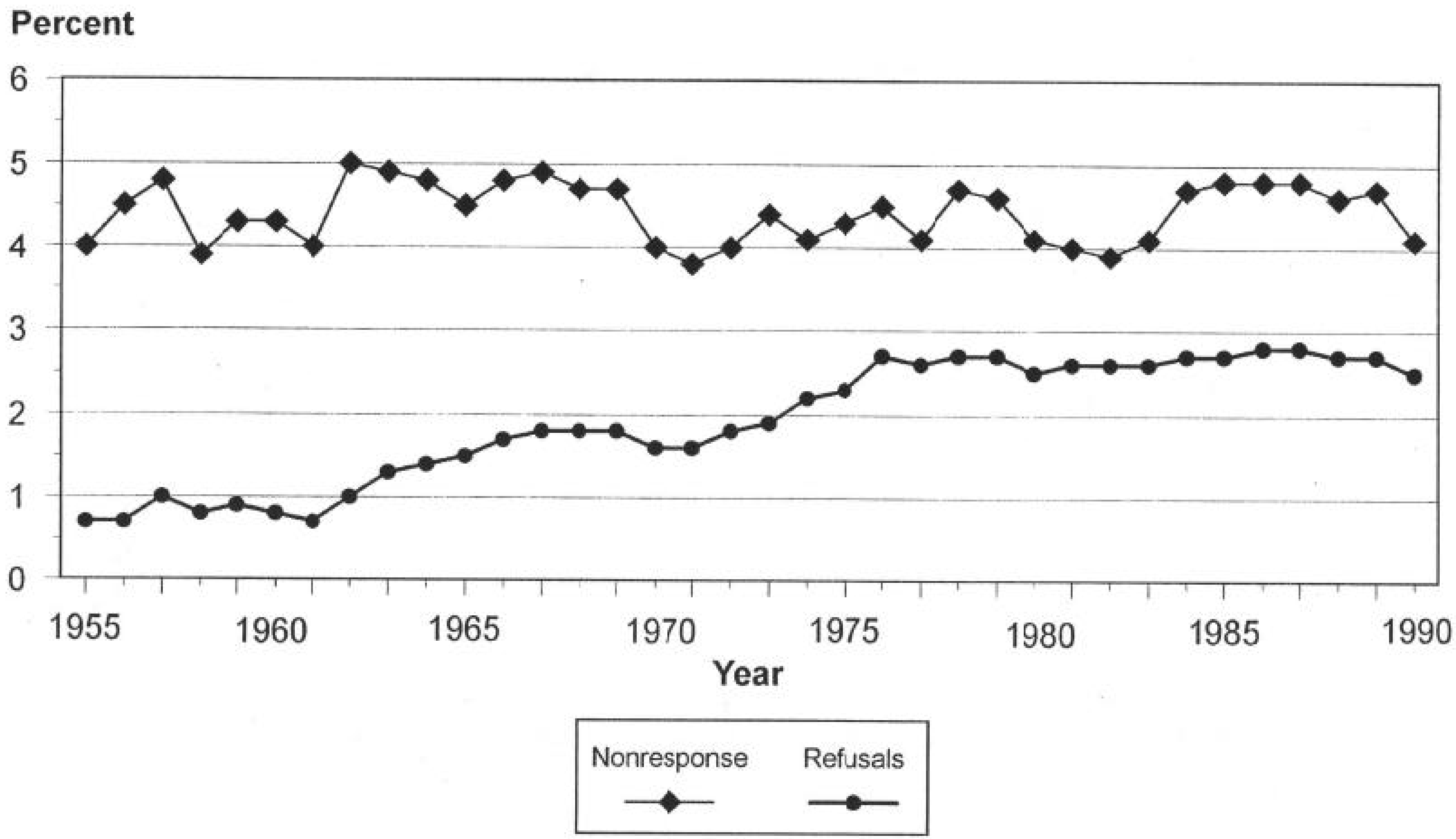
Figure 3.
Mean Unweighted Nonresponse Rates, Demographic Surveys



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

Figure 4.

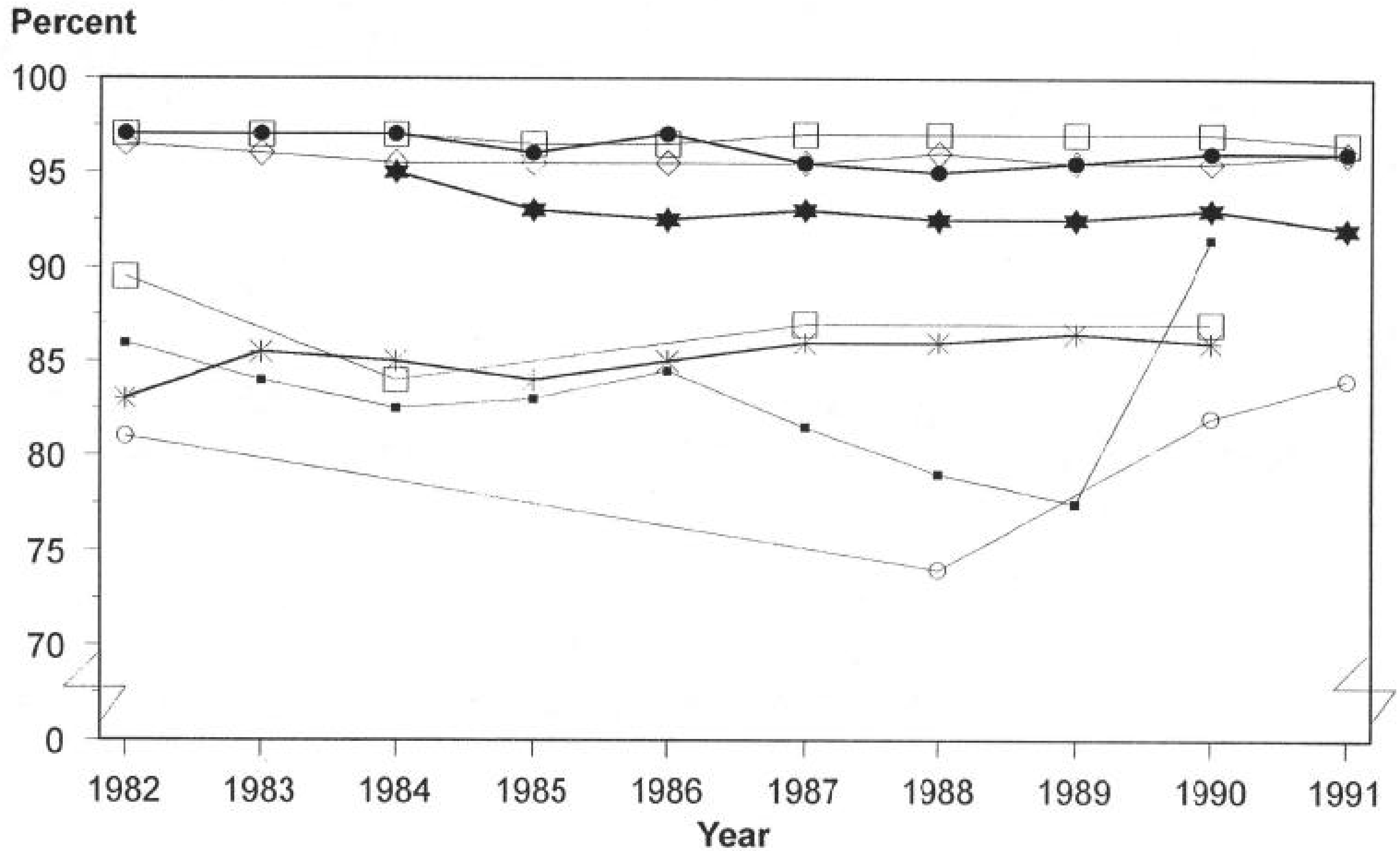
Unweighted Nonresponse Rates, Current Population Survey



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

Figure 5.

Unweighted Response Rates, Demographic Surveys



SOURCE: Federal Committee on Statistical Methodology, Subcommittee on Survey Nonresponse.

Figure 6.--Summary of FCSM Recommendations on Survey Nonresponse

Recommendation 1. **Survey staffs should compute response rates in a uniform fashion over time and document response rate components on each edition of a survey.**

The subcommittee chose not to recommend that every survey use the same response rate computations. Other groups have recommended such uniformity (see CASRO, 1982). In the Subcommittee's view, every definition of response rate components offers some useful information. Some response rate definitions inform the designers about the rate of success of measurement of the average sample unit; others focus on different causes of nonresponse. One can distinguish between measures useful as management tools and measures that should be reported to data users so that they can assess the quality of the survey data.

Recommendation 2. **Survey staffs for repeated surveys should monitor response rate components (e.g., refusals, not-at-homes, out-of-scopes, address not locatable, postmaster returns, etc.) over time, in conjunction with routine documentation of cost and design changes.**

The Subcommittee believes that response rate components are useful tools to monitor changes in the quality of survey statistics. Response rates should be easily accessible and timely. By themselves, they are not error measures; however, for repeated surveys, changes in response rate components may signal the need for supplementary study of nonresponse error properties. Such changes can alert the survey designers to changes in the "survey-taking climate" that affect completion of measurement, point to changes in the administrative controls over response rates that may need adjustment, and help measure the effects of any design changes made.

For ongoing surveys, graphs of time series of response rate components, juxtaposed with costs for each collection cycle, and indicators of design changes introduced in that cycle, can be valuable management tools. Survey managers need better tools to diagnose the causes of cost changes in data collection activities. Falling response rates, especially those associated with cases requiring much effort prior to the ultimate nonresponse, magnify cost pressures on surveys. The subcommittee's study did not collect data on survey costs, because comparable cost information across surveys was not believed to be available.

Recommendation 3. **Agencies that sponsor surveys should be empowered to report the response rates of their surveys. The sponsoring agency should explain how response rates are computed for each survey it sponsors. Response rates for any one survey should be reported using the same measures over time, so that**

users may compare the response rates. Response rate components should also be published in survey reports.

An agency that sponsors surveys should compute and explain in its survey publications the response rates for each of the surveys it sponsors. Surveys sponsored over time should report the same measure of response for all data collection periods so that users can compare these measures over time. The actual method used to compute response rates should be described in all publications issued.

The results of recommendations 1 and 2 should be shared routinely with the users of survey data, along with discussions of the relevance of response rates to evaluating the quality of the survey data. An analysis of the characteristics of the nonrespondents should be implemented routinely as part of each cycle of data collection.

Recommendation 4. **Some research on nonresponse can have real payoffs. It should be encouraged by survey administrators as a way to improve the effectiveness of data collection operations.** The Subcommittee believes that areas of research most likely to yield payoffs include:

- Studies of the relative costs of final efforts to raise response rates, through persuasion, repeated callbacks, and other measures. When these costs are compared to number of cases added to the respondent pool, the relative cost per case can be computed. Studies of the effects of these final cases can be made in an effort to assess the cost effectiveness in terms of mean square error of the final efforts.

- Studies of the measurement error properties of information provided by the reluctant respondent cases, relative to the nonresponse bias in statistics that would omit them from computations. This would address a key question in survey design: When data collectors exert great effort to persuade the reluctant to respond, is one type of error, nonresponse, merely exchanged for another type, measurement error? Perhaps, those persuaded to respond may exert less effort at providing accurate data?

- Studies on what variables should be collected to improve post-survey adjustment for unit nonresponse (see Madow et al, 1983: Recommendation 10(2)). When observable or inferred characteristics of nonrespondent units are related to the survey variables and to the likelihood of participation, then collecting and using these variables in post-survey adjustment models might be a cost effective method of reducing overall mean square errors.

MODEL-BASED REWEIGHTING FOR NONRESPONSE ADJUSTMENT

David A. Binder, Sylvie Michaud and Claude Poirier
Statistics Canada, Ottawa, Ontario, Canada K1A 0T6

ABSTRACT

Nonresponse in surveys is inevitable. Much has appeared in the literature on methods of compensating for this source of nonsampling error. There is a growing interest in attempting to understand the causes of nonresponse and studying the differences in characteristics between respondents and nonrespondents. In this paper, we briefly review some related literature, discuss modelling approaches for adjusting for nonresponse and present the research findings for two surveys conducted at Statistics Canada. In both the Survey on Labour Income Dynamics and the Farm Financial Survey, we examine differences in characteristics between respondents and nonrespondents and the suitability of adopting a modelling approach for compensating for nonresponse.

KEY WORDS: Generalized regression estimators; Logistic regression; Response propensity.

1. INTRODUCTION

In virtually every survey, no matter how carefully it is designed, we must accept the fact that some data will be missing. Other than data that is missing by design, such as data from nonsampled units, data can be missing for many reasons; for example, non-contact with the respondent, refusals, late reporting, collection and processing errors, data deletion due to edit failure, undercoverage, etc.

Some measures must be taken to deal with such nonresponse. Over the years, a host of techniques has been developed. The actual choice of technique should depend on a number of factors. These include the method of estimation to be used, the amount of information about the nonrespondents that is available, the extent of other sources of error such as sampling error and response error, the relative importance of the variables to be estimated, the resources available for exploring the problem, the nature of the analyses to be performed and the statistical inferences to be made from the survey, etc.

However, even with all of these criteria, there must necessarily be some subjective judgments on the nature of the nonresponse. As we shall see, many of the methods for coping with nonresponse make use of models, either explicitly or implicitly. Even the most ardent advocates of the pure design-based school will resort to some model assumptions when it comes to adjusting for nonresponse. This presents a new set of problems associated with the statistical inferences, since the randomization distributions on which the inferences are based are no longer purely design-based, unless the nonresponse mechanism can be considered to be part of that design.

In this paper, we shall focus on the implications of the estimation method to be used and the amount of information about the nonrespondents that is available. It will be assumed that the prime focus of the survey is to obtain estimates of descriptive statistics, such as means, totals, differences and ratios. Often nonresponse is broadly categorized into *unit-level* nonresponse and *item-level* nonresponse. This categorization is often extended, in the case of longitudinal or follow-up surveys, to *wave-level* nonresponse, where wave nonresponse is usually unit nonresponse on a particular survey occasion. In fact, it is for the case of wave nonresponse where we have the richest source of data for the nonrespondents who reported in previous waves.

Unit nonresponse is usually defined as cases where only the frame information is available for the respondents. In practice, this definition is extended to other cases where there is insufficient usable data from the respondents. The usual method for dealing with unit nonresponse is to use an "appropriate" weighting procedure to compensate for the nonresponse. (We define weighting procedures here broadly to include weight adjustments implied by regression, ratio or similar estimation techniques using auxiliary data.)

On the other hand, item nonresponse is handled either through imputation at the item level, or by ignoring the usable information and treating the respondent like unit-level nonrespondents. For wave-level nonresponse to longitudinal surveys, either reweighting or item imputation may be suitable. In this paper, we focus on the methods that use weighting techniques.

In Section 2, we discuss the basic theory underlying many of the adjustment methods and give a brief literature review. In Sections 3 and 4, we give examples of two surveys at Statistics Canada where some of these models have been studied recently. We summarize our findings in Section 5.

2. SOME GENERALITIES

2.1 Estimation

In general, we are interested in means, totals, ratios, etc. of survey variables. We denote the value of the i-th survey variable for the k-th respondent as y_{ik}. In cases where the occasion, t, is relevant, we can use y_{ikt} instead. A sample is selected according to some well-defined sampling plan. The sampling plan is usually based on frame information such as geography, and other classification and size variables. We use s to refer to the selected sample. Unfortunately, in practice, after the k-th respondent is selected, a number of things can go wrong in the process of obtaining and recording the y-values. Some of these are in the general category of response errors, where we obtain data, but they are not the y-values we were seeking. In this paper, we ignore these types of errors, except to point out that if these errors lead to large biases, the resources for nonresponse concerns may need to be trimmed in order to address the larger problem. The problem that we are addressing here is the case where the y-values are unobtainable. We denote by $s' \subseteq s$ the set of units for which we obtain usable y-values. (The subscript t is implied, where appropriate, for longitudinal surveys.)

First, we describe the estimators in the case of no nonresponse. Associated with each sampled unit, k, we have a survey weight given by

$$w_k(s) = g_k(s)\,\pi_k^{-1},$$

where π_k is $\Pr(k \in s)$, the usual first-order inclusion probability, and $g_k(s)$ is a weight adjustment that makes use of auxiliary frame data, such as poststratification, regression and ratio adjustments, etc.; see, for example, Särndal, Swensson, and Wretman (1992). We assume that the estimator of a total for a y-variable on the t-th occasion is given by

$$\hat{Y}_{it} = \sum_{k \in s_t} w_k(s_t)\, y_{ikt}. \qquad (2.1)$$

Note that this estimator could be made more general, if necessary, to allow for composite estimators and multiphase samples which can depend on y-values that are observed on other occasions, but we do not introduce this complexity here. Sufficient conditions for (2.1) to be asymptotically design consistent are:

1) the probability distribution of s depends only on the auxiliary data but not directly on the y-values for the current occasion, (2.2.1)
2) the limiting expectation of $g_k(s)$ is unity, (2.2.2)
3) the variance of $\hat{Y}$ is asymptotically zero. (2.2.3)

We now consider the implications of nonresponse. Formally we assume that, given the sample, s, the set of responding units, s', follow a probability distribution $p(s'|s)$. This is completely general, allowing for correlated response patterns. It also allows for the classical case, where it is assumed that the response behaviour is nonrandom and is an inherent attribute of the selected respondents, just like the survey variables. We now consider methods of nonresponse adjustment which we refer to as *generalized reweighting methods*. Associated with each responding unit, k, we have an adjusted weight given by

$$w'_k(s', s) = g'_k(s', s)\, w_k(s),$$

where $g'_k(s', s)$ is a weight adjustment that makes use of auxiliary frame data, as well as other information that may be available for the nonresponding units. This allows the weight adjustment to depend on survey values that were observed on previous occasions from a longitudinal survey. We assume that the estimator of a total for a y-variable on the t-th occasion is given by

$$\hat{Y}_{it}^{(GR)} = \sum_{k \in s'_t} w'_k(s'_t, s_t)\, y_{ikt}. \qquad (2.3)$$

We let $\rho_k(s)$ be $\Pr(k \in s' | s)$. In addition to (2.2.1) to (2.2.3) above, sufficient conditions for (2.3) to be asymptotically consistent with respect to the original design and the response probabilities are:

1) the probability distribution of s' given s depends only on the auxiliary data and the survey data from previous occasions, but not directly on the y-values for the current occasion, (2.4.1)

2) the limiting expectation of $g_k'(s', s)$ is $\{E[\rho_k(s)]\}^{-1}$, (2.4.2)

3) the variance of $\hat{Y}^{(GR)}$ is asymptotically zero. (2.4.3)

If (2.4.2) is violated, then the expectation of $\hat{Y}^{(GR)}$ is

$$\sum E[g_k'(s', s)] E[\rho_k(s)] y_k . \qquad (2.4.4)$$

The form of this bias is important, because if one were to impose model assumptions on the y-variables, it is possible that the model-bias becomes small. However, for those who wish to make the fewest model assumptions, it is clear that one should restrict attention to adjustment methods which yield condition (2.4.2) as closely as possible. This implies that the weight adjustment should reflect the propensity to respond as nearly as possible. Of course, the probability mechanism generating these response probabilities are generally unknown, so the weight adjustment must necessarily be model-based.

Another important feature of (2.4.2) is that if there are some "hard-core" nonrespondents -- that is, units where $\rho_k = 0$ -- there would be no consistent estimates.

2.2 Examples from the Literature

The most basic form of reweighting for nonresponse that may lead to acceptable results is to simply use $g_k(s')$ instead of $g_k(s)$ in (2.1). This implies that

$$g_k'(s', s) = g_k(s') / g_k(s) .$$

This was suggested by Bethlehem (1988) for the case of the generalized regression estimators. In this case we have that the bias of the estimator is $X\beta^* - Y$, where β^* is the expected value of the estimated π-weighted regression coefficient with no nonresponse adjustment. We see then that even though this estimator is generally biased, if the regression model is reasonable, the bias can be small.

Oh and Scheuren (1983) discussed weighting class adjustment methods, which is a poststratified estimator using weighting classes as poststrata. We see that this is consistent under the assumption that the response propensities are equal within weighting classes. In practice, this technique is in widespread use; see, for example Chapman, Bailey, and Kasprzyk (1986). It was extended to generalized regression estimators by Särndal and Swensson (1987).

One of the difficulties with weighting class adjustment methods is that there may be too many weighting classes to control. Binder and Théberge (1988) showed that with a multiplicative model for response propensities, raking ratio estimators will yield unbiased estimates. This is consistent with (2.4.2). More complex weighting schemes are proposed by Alexander (1987) and Deville, Särndal, and Sautory (1993). These could be justified under various model assumptions for the response propensities.

Many authors have proposed the use of logistic regression models to explain the nonresponse mechanism. This is a commonly used model for binary dependent variables. Examples of this can be found in Ekholm and Laaksonen (1991), Folsom (1991), and Lepkowski, Graham, and Kasprzyk (1989). In the latter paper, the logistic regression model is compared to weighting class adjustment methods, where the weighting classes are determined through some data analytic searching methods.

In Iannacchione, Milne, and Folsom (1991), after weights are included to reflect the estimated propensity to respond, the weights are fine-tuned so that certain estimates correspond to the estimate obtainable with the nonrespondents included. This is possible for wave nonresponse where certain estimates can be made for a previous wave using either the previous wave respondents or the current wave respondents. This technique should generally improve the estimates. The differences in the estimates can also be used as an diagnostic tool for the model.

Judkins and Lo (1993) and Eltinge and Yansaneh (1993) used logistic regression to model the nonresponse propensities, but then created weighting classes based on the fitted values and used weighting class adjustment methods to reweight. One of the drawbacks of the weighting class adjustment methods is that the appropriate weighting classes are not always obvious, so that such data modelling is used to help define the classes. It is expected that this method should yield results that are similar to the weights based on the logistic regression. However, if the logistic model is correct, the method will tend to introduce a small bias since (2.4.2) will be violated. In practice, though, the logistic regression model is only an approximation to the true probability mechanism.

As we can see, reweighting methods have a strong base in the literature. The theory we have given in Section 2.1 indicates that the validity of these methods are model-based. Therefore it can be important to study the characteristics of the nonrespondents to develop the most suitable model. In Sections 3 and 4, we perform such studies on each of two surveys. We see that the models help our understanding of the factors that contribute to nonresponse.

An important side benefit of such studies is to help the survey manager pinpoint areas for improvement in the data collection phase.

3. SURVEYS OF LABOUR AND INCOME DYNAMICS AND LABOUR MARKET ACTIVITY

3.1 Introduction

Statistics Canada launched a major panel survey of households in 1994 called the Survey of Labour and Income Dynamics (SLID). The survey follows individuals and families for six years, collecting information on their labour market experiences, income and family circumstances. Its origins are in several surveys, including the Labour Market Activity Survey (LMAS). The LMAS served both as a longitudinal and as a cross-sectional survey. Two panels have been conducted to date, a two-year panel during 1986-1987 and a three-year panel during 1988-1990. For each longitudinal panel, respondents who participated in the first wave were interviewed and traced. All persons living with them in the following waves were also interviewed but not traced. Different studies are currently being conducted on nonresponse to the LMAS in hopes of finding approaches that will minimize the impact of nonresponse on the SLID data. Here we discuss our study on model-based reweighting.

Similarly to its predecessor (LMAS), the longitudinal sample for SLID is selected from the sample of dwellings that participated in the Labour Force Survey (LFS) in January 1993. The LFS has a response rate of 95%. Out of those respondents close to 90% agreed to participate in SLID. This sub-sample of respondents, comprising 15,000 households, is defined as the longitudinal sample, representative of the Canadian population as of January 1993. The longitudinal sample will be interviewed for six years, with two interviews carried out each year. Note that a sub-sample of LFS respondents who had refused to participate to SLID has been selected for evaluation purposes. If they respond in subsequent years, we may be able to determine how different they are from the rest of the sample. Preliminary analysis could not find systematic differences in the LFS characteristics between the nonrespondents and the respondents. More studies will be done by linking the full sample to administrative files to be able to evaluate if there are differences in terms of income characteristics.

Attritional nonresponse will be compensated with a weighting adjustment. Imputation will be used to compensate for some nonresponse; for example, nonresponse that is non-attritional. The weighting will include the following steps:

i) calculation of the initial weight based on the sample design,
ii) nonresponse adjustment,
iii) post-stratification by province, age groups, and sex to the 1993 population estimates.

The longitudinal panel of LMAS has been used as the research vehicle for the nonresponse modelling and weighting adjustments.

3.2 LMAS Survey Design and Nonresponse

For the first interview of the panel, LMAS is conducted as a supplement to the January Labour Force Survey (LFS). All eligible respondents from the LFS are included in the LMAS sample. In the subsequent waves, for the longitudinal component of LMAS, all respondents to the first wave are interviewed in January of the following year(s). People are traced if they have moved.

LFS uses a multiple stage sample design. A stratum is defined based on geographic variables. At least two distinct PSU's (primary sampling units) are selected within each stratum. LFS initial weights go through a series of adjustment factors at the stratum level to produce a sub-weight. This sub-weight is then adjusted to population estimates by province/age-group/sex groups, plus an adjustment by Economic Region and Census Metropolitan Area, to produce a final weight. More details may be found in Singh, Drew, Gambino, and Mayda (1990).

For the LMAS longitudinal sample, nonresponse adjustment is done at the stratum- component level, corresponding to a PSU or a group of PSU's, as defined for the LFS. A poststratification is then done to adjust the nonresponse adjusted weights to population estimates at the province/age-group/sex level.

When the LMAS file was evaluated, it was found that nonresponse was quite different among certain groups:

- movers, including people that could not be traced, had a nonresponse rate of close to 20% while nonresponse for non-movers was about 2%. This was by far the characteristic that presented the most differences,
- based on characteristics from Wave 1, persons that were employed in Wave 1 had higher response rates after three years than those who were unemployed in Wave 1,
- similarly, persons that were married in Wave 1 had higher response rates in Year 3, compared to those who were single in Year 1,
- persons who lived in non-urban areas in Year 1 had higher response rates after three years.

The different characteristics between respondents and nonrespondents suggested that nonresponse adjustments should be done at some level different than stratum-component. Logistic regression was used to model the nonresponse behaviour. The multiple logistic response function is

$$\text{logit}(p) = \log[p/(1-p)] = \boldsymbol{\beta}'\boldsymbol{x},$$

where p is the probability of response to the 1987 survey for a 1986 survey respondent, $\boldsymbol{\beta}$ is the column vector of regression parameters, and x is the vector of independent variables.

3.3 Modelling the Response Probabilities

The dataset for the 1986/87 panel of LMAS consisted of 66,817 individuals, of which 3,385 (5%) were nonrespondents to the 1987 interview. Demographic variables that were likely to be related to nonresponse were chosen from the 1986 LMAS master file as possible independent variables for the model.

The variables examined for inclusion in the nonresponse model were:

Province at 1986 interview
Urban/Rural area indicator at 1986 interview

Household size at 1986 interview
Type of dwelling (house; other) at 1986 interview
Status of dwelling (owned; rented) at 1986 interview

Sex
Age at 1986 interview
Marital status at 1986 interview
School attendance (full time; part time; none) in 1986
Highest level of education at 1986 interview

Any employment in 1986
Any unemployment in 1986
Any out-of-labour-force in 1986
Number of jobs in 1986
Any short tenure jobs (< 2 years) held in 1986
Any long tenure jobs (2 years or more) held in 1986
Any absences from work in 1986
Industry of job(s) in 1986

Average weekly income (over all jobs) in 1986
Received any unemployment insurance in 1986
Received any welfare in 1986
Moved (changed address between 1986 interview and 1987 interview)

All the categorical variables were converted to groups of dichotomous variables. The differences between respondents and nonrespondents with respect to the independent variables were analyzed. The correlations between all pairs of these variables were examined to find any potential multicollinearity.

First, a stepwise linear regression procedure was used to identify potentially useful variables for the modelling. This reduction in the choice of variables resulted in fewer variables to be entered into the logistic procedures saving considerable computer resources. The variables given in the STEPWISE procedure were entered into the SAS procedure PROC LOGISTIC with the BACKWARD and FAST options. These options allowed LOGISTIC to use an approximate backward elimination method to eliminate nonsignificant variables. Different logistic regression models were fitted to the full dataset using combinations of the most significant variables identified from the sample file. A consideration in choosing the model was the number of variables. It was desired to have a model with a small number of variables so that utilizing the model would be simple.

The model is used to make adjustments to the weights of the respondents in the second year (1987). For this model, the dependent variable was total nonresponse, and the independent variables were characteristics observed the previous year (1986) plus the current year's information (1987) on whether or not the person moved.

The BACKWARD option of PROC LOGISTIC was used with the sample file to identify eight variables related to nonresponse.

Male	(MALE)
Single	(SINGLE)
Rented dwelling	(RENT)

Any employment	(ANYEMP)
Highest education=secondary	(EDUCSEC)
Moved since 1986 interview	(MOVED)
Household size, to a maximum of 8	(HHS)
Age	(AGE)

Before fitting the models on the full dataset, the two continuous variables (household size and age) were examined for linearity on the logit scale. As with the prediction model, the age variable was replaced by two dichotomous variables for age: AGE1 for persons aged 25-54, AGE2 for persons aged 55-69 - the survey was conducted for persons aged 16-69 - and a transformation was applied to household size (HHSTRANS=¦HHS-4.5¦).

Four models were fitted to the full dataset: (1) using all eight variables; (2) using all except RENT; (3) using all except EDUCSEC; (4) using all except EDUCSEC and AGE. Although all eight variables were significant using the sample file, when the models were fitted to the full data file, certain ones no longer appeared important. However, it was decided to retain them in the models anyway. The statistics for evaluating the fit of the models indicated few differences between the four models. The Pearson residuals were plotted against the fitted values and the residual plots were examined. The residuals from Model (3) indicated a slightly better fit with fewer extreme values. Again using the sample file, the data were examined for the presence of two-way interactions between the variables in the model. Two sets of interactions were added to the model: the (AGE1 AGE2)*HHSTRANS and (AGE1 AGE2)*SINGLE. A summary of the fitted values for this model is given below. Note that the age and single variables as well as their interactions are not statistically significant. Nevertheless, when a model was fitted with these variables removed, it was found that there were more extreme values in the residuals.

Table 1

Parameter Estimates for Weighting Final Model.

Variable	β	s.e.	χ^2
INTERCEPT	-3.81	0.14	702.59
HHSTRANS	0.13	0.06	4.97
MALE	0.25	0.04	41.98
RENT	0.23	0.04	29.14
SINGLE	0.11	0.16	0.43
MOVED	2.31	0.04	3065.95
AGE1	-0.15	0.17	0.75
AGE2	-0.19	0.15	1.65
AGE1*HHSTRANS	0.02	0.07	0.07
AGE2*HHSTRANS	0.05	0.06	0.55
AGE1*SINGLE	0.13	0.18	0.52
AGE2*SINGLE	0.11	0.17	0.40

Using the estimated parameters from the final model, predicted probabilities of nonresponse were calculated for all respondents to the 1987 interview and a nonresponse adjustment was made. Finally, a poststratification adjustment to population control totals at the province-sex-agegroup level, yielded the 1987 final weight.

3.4 Evaluation of the Weights

If the nonresponse weighting adjustment is adequate, there should be no difference in estimates obtained from the 1986 respondents and estimates obtained from the 1987 respondents when tabulating on 1986 characteristics. A number of demographic and labour-related characteristics were evaluated. Estimates were calculated using the 1986 weights, the 1987 model-adjusted weights, and the 1987 regular weights, including a ratio-adjustment at low geographic levels for nonresponse adjustment. For each characteristic a 95% confidence interval was calculated for the estimate based on the 1986 weights. The two 1987 estimates were compared for differences to the 1986 estimates as well as differences to each other. Tables 2 and 3 below show some of the results. Table 3 incorporates the poststratification adjustment, which, in general, improves the estimates.

Table 2

Comparison of the estimates with the two non-response adjustments, before the post-stratification, tabulated on 1986 characteristics.

	1986 estimate	95% c.i. for 1986 estimate	1987 model-based estimate	1987 regular estimate
Marital Status Married	64.6%	(64.1,65.1)	65.1%	65.7%
Single	26.7%	(26.3,27.0)	26.3%	25.7%
Widowed	3.1%	(2.9,3.3)	3.0%	3.0%
Divorced	5.7%	(5.4,6.0)	5.6%	5.5%
Highest Education Grade 0-8	14.7%	(14.2,15.2)	14.6%	14.6%
Secondary	50.3%	(49.7,50.9)	50.0%	50.0%
Some Post-Secondary	10.1%	(9.8,10.4)	10.2%	10.1%
Post-Sec. Cert./Dip.	12.9%	(12.5,13.3)	13.1%	13.1%
University Degree	12.0%	(11.6,12.4)	12.2%	12.2%
Weeks Employed in 1986 0 weeks	22.8%	(22.4,23.2)	22.6%	22.5%
1-26 weeks	12.0%	(11.7,12.3)	11.7%	11.6%
27-48 weeks	12.2%	(11.9,12.5)	12.1%	12.0%
49-52 weeks	53.0%	(52.4,53.6)	53.6%	54.0%

Of all the characteristics compared, only one 1987 estimate was outside the 1986 confidence interval: weeks employed=49-52 using the regular weighting. One pattern was clear, however. The estimates using the model-based weights were consistently closer to the 1986 estimates than those using the regular method of weighting.

4. FARM FINANCIAL SURVEY

4.1 Introduction

The Farm Financial Survey (FFS) has been a regular agricultural survey since 1980. The objective of the survey is to gather financial information on Canadian farmers. The survey collects information on revenues, expenses, assets and liabilities. Crop and livestock information are also collected to measure physical characteristics of the farms. Due to the collection of sensitive data, a low response rate has always been observed for the survey. A study was initiated on the 1992 survey data to identify the causes of nonresponse and possible solutions to reduce its impacts on the estimates.

The population of interest consists of all Canadian farms active for the reference year, excluding the multi-holding companies, the institutional farms, the community pastures, the farms on Indian Reserves and the farms with less than $2,000 in sales. The survey population is represented by a list frame and an area frame. The 1992 list frame was a register of all of the 1986 Census farms without the farms defined by the above exclusion rules. The list frame was stratified within each province by farm type and by farm size. The farm size was defined by the total farm assets derived on the Census.

The area frame was used to compensate for the undercoverage due to the Census itself or caused by new farms which started their activities since 1986. Basically, the area frame was a list of land segments outlined on topographic maps. Stratified replicates of segments were selected from the area frame. All farmers operating some land in the sampled segments were enumerated, and a register was created. There were 1,153 area frame farms that did not appear on the list frame. They were all contacted for the FFS as for other agricultural surveys. In addition to the area frame farms, a stratified sample was selected from the list frame to obtain a overall sample of about 12,000 farms. See Britney and Poirier (1992) for more details on the 1992 FFS sample design.

Domain estimation within each stratum was performed to obtain estimates of level from both the list and area samples. The simple expansion estimator was used on the 1992 list sample. The initial weighting was done by stratum using the population size over the observed sample size, so that a nonresponse adjustment is made at the stratum level. For the area frame, the estimation was done separately by replicate. For a given replicate, the data were aggregated at the segment level by applying to the farm data, factors corresponding to the proportion of the farms within the segment. Then, the segment totals received expansion weights (π) to represent the population. When nonresponse occurred for an area farm, the respondents within the same segment were reweighted on an area basis to compensate the farm land for which data were unavailable. For both the list and area units, partial nonresponses were donor imputed and used the same way the regular respondent were. Details are given by Maranda (1989).

The nonresponse observed in the 1992 Farm Financial Survey was relatively important. The FFS questionnaire was relatively long with many sensitive questions related to the financial balance sheet. The resulting total unit-level refusal rate of about 15% across the country was the highest of our agricultural surveys. In addition to the total refusals, the no-contacts represented another 5% of the sample. Some provinces presented higher nonresponse rate than others. In Saskatchewan, data were unavailable for almost 30% of the sampled farms. Table 4 shows the nonresponse distribution across the country.

Table 4

1992 Nonresponse Distribution

Province	Sample Size	Total Refusal	No-Contact
Newfoundland	211	20	16
P.E.I.	528	64	14
Nova Scotia	668	74	11
New Brunswick	537	48	18
Quebec	1311	124	51
Ontario	1513	250	84
Manitoba	1756	321	109
Saskatchewan	1880	424	126
Alberta	1868	312	109
B.C.	1448	175	138
Total	11720	1812	676

4.2 Nonresponse Models

A part of our study was first to identify the causes of nonresponse. This could help taking decisions related to the collection methods to increase the response rate. It also allowed the identification of factors that may be considered in any nonresponse reweighting models. Since, the no-contacts and the refusals were possibly caused by different factors, they were kept separate in all of the hypotheses we made. The potential causes that were studied on the 1992 FFS data are:

1 - **The frame origin:** This corresponded to whether or not the farm was selected from the list frame. Since the area frame farms were conceptually missed by the Census, they probably showed characteristics that trend to generate nonresponse. Also, since the area sample is being used by many agricultural surveys, the area frame farms might refuse because of their response burden.

2 - **The farm size:** The capability or the will to respond could depend on the farm organisation and on its size.

The size was evaluated using the farm assets and sales obtained from the 1986 Census of Agriculture. This size was available only for the list units.

3 - **Geography:** The geographic location aimed to identify the interviewer effect and the impact of farmer associations which could boycott government surveys because they were not benefited by their programs. Census divisions were used to verify this hypothesis.

4 - **Farm type:** The farmer's availability depends on the type of his farm. Seven categories of farm type were used to differentiate the farms.

5 - **Response burden:** Because the large number of agricultural surveys held in a short period of time, the response burden became important for some farmers. The overlaps with the December Stock Survey and the January Livestock Survey (JLS) were both studied to verify its impact on the response rates. These surveys were both conducted less than two months before the FFS. The effect of the overlap with the previous FFS, held in 1990, was also investigated.

6 - **Age of operator:** The age of an operator could affect its will to cooperate, but the data available to verify this hypotheses were not reliable enough to do any studies.

Tests of independence were conducted to verify if any of the above factors could affect the response status: 'completed', 'no-contact' and 'refusal'. The partial refusals were included with the completed questionnaires. The statistic used to conduct the independence tests was the weighted Pearson statistic χ^2 with the Fellegi (1980) correction to take into account the design effect. This test is known to be conservative.

The farm assets and sales, which were both indicators of the farm size, were replaced by categorical variables defined using the estimated quartiles. The census divisions representing the geographic location were grouped into a maximum of 9 classes within each province. This ensured a minimum number of observations within each cell of the cross classification with the response status.

In some cases, where dependence was detected between the factors and the response status, additional tests were conducted to identify the nature of the dependence. This was done through statistical tests on proportions. The most important conclusions are described here but more details can be found in Poirier (1994).

The independence tests, conducted with a confidence level of 5%, identified certain causes of nonresponse. First, within each province except Ontario, the farm type had a high impact on nonresponse. Also, the farm size measured in term of assets affected the response rates in most of the provinces, but no significant impact was due to the sales variable. The geographic location and the response burden generated by the previous FFS survey significatively affected the probability to respond in three provinces. Finally, the frame origin and the overlap with the January Livestock Survey or the December Crops Survey seemed to not affect the response status at all.

As in Section 3, we modelled the nonrespondent behaviour with logistic regression modelling using the SAS procedure LOGISTIC. We performed the analysis separately by province. Using frame origin as an independent variable, the results confirmed the previous conclusions of no frame effect. Since some variables were not available for the area sample and since the frame origin did not seem to affect the response, the remaining analyses were performed only on the list units, which represented more than 90% of the whole sample. In the rest on this paper, the results applied for the list units only.

For the purposes of this study, the following variables were included in the model:

- i) Assets (1 if assets are smaller than the median, 0 otherwise),
- ii) Sales (1 if sales are smaller than the median, 0 otherwise),
- iii) $Type_i$ (1 if in the i^{th} farm type, 0 otherwise),
- iv) $Area_i$ (1 if in the i^{th} geographic area, 0 otherwise),
- v) FFS (1 if in the 1990 FFS sample, 0 otherwise),
- vi) JLS (1 if in the 1992 JLS sample, 0 otherwise),

The farm types are (1) crop farms, (2) dairy farms, (3) cattle farms, (4) hog farms, (5) poultry farms, (6) sheep farms, and (9) unknown type of farm.

The variables that were found more significant by the BACKWARD option within the provinces were kept in the model. The most commonly selected variables were the farm types and the FFS variables. Table 5 shows the resulting estimated parameters corresponding to the variables kept in the model. It also provides the attached χ^2 values.

Table 5

Nonresponse Logistic Parameters with their χ^2 Values

Prov.		Intercept	FFS	Assets	Type4	Type5	Type6	Type7	Area1	Area2	Area3	Area5	Area6	Area7
NFLD	B	1.37	1.05	.	.	-1.78	.	.	.	.	.	.	.	.
	χ^2	37.23	5.75			6.76								
PEI	B	1.75	.	.	.	.	.	.	.	.	.	.	.	.
	χ^2	204.18												
NS	B	1.93	.	.	.	.	.	.	.	.	.	.	.	.
	χ^2	275.05												
NB	B	1.65	0.71	.	.	.	.	.	.	.	.	.	.	.
	χ^2	98.87	6.79											
QUE	B	2.05	.	.	.	-0.58	.	.	.	-0.45	.	.	.	.
	χ^2	392.98				5.11				4.03				
ONT	B	0.93	0.53	.	.	.	1.07	.	.	.	.	.	.	.
	χ^2	71.00	14.18				4.09							
MAN	B	0.94	0.44	.	.	-0.45	.	-1.16	0.59	.	.	.	.	.
	χ^2	113.35	13.34			4.21		38.90	14.05					
SASK	B	0.94	0.51	0.31	.	.	.	-2.40	.	.	.	.	-0.61	-0.32
	χ^2	78.44	18.67	6.85				134.31					12.20	4.22
ALB	B	1.04	0.69	0.26	-0.33	.	.	-1.12	.	.	.	.	.	.
	χ^2	79.25	30.12	4.30	5.51			40.48						
BC	B	0.88	.	.	.	.	.	.	0.93	0.86	0.47	0.43	.	.
	χ^2	74.82							15.29	21.01	5.85	5.13		

Variables that were not significant for any of the provinces have been removed from this table. From the χ^2 results, it appears that the FFS overlap and the farm Type7 (representing the sheep farms) have the most important impacts on nonresponse. The positive FFS parameters mean that farms overlapping the previous FFS tended to have higher response rates, whereas the negative sheep farm parameters (Type7) imply they tended to respond less often.

Weighted regressions were also fitted to the data using the WEIGHT statement of the LOGISTIC procedure. The weighting variable was defined at the stratum level as the design weight adjusted to the overall sample size. Stratum level adjustments were not performed. The resulting estimated parameters were very close to the first set of estimates which, as we explained in Section 3, is highly desirable.

4.3 Evaluation of the Weights

To evaluate the nonresponse adjustment, the 1992 frame values representing farm assets were estimated from the sample. Assets levels were estimated for each province with the corresponding coefficient of variation (CV), including the nonresponding units. Then, estimates based only on respondents were produced, using the original weight, adjusted for nonresponse at the stratum level only. By comparing both set of estimates we could derive the nonresponse bias introduced by the current method. Finally, regression adjusted estimates were produced from the above logistic model.

If we denote by $\hat{Y}_0$ and CV_0 estimated level and coefficient of variation, respectively, from the full sample, and $\hat{Y}_{adj}$ the corresponding adjusted estimates based only on respondents, we can estimate the bias associated with the adjustment model. In Table 6 we show the results.

Table 6

Comparison of the Adjustment Models for 1992 Frame Value of Farm Assets

Prov.	Y_o	CV_o (%)	Stratum Adjusted Weight	Logistic Adjusted Weight
			BIAS (%)	BIAS (%)
NFLD	7.7 E07	2.91	2.34	1.80
PEI	6.7 E08	0.76	-0.35	0.19
NS	7.9 E08	0.68	-1.15	-0.27
NB	6.1 E08	0.82	-0.16	0.46
QUE	8.5 E09	0.53	-0.28	-0.04
ONT	2.1 E10	0.56	-1.16	-0.85
MAN	9.0 E09	0.57	0.21	0.45
SASK	2.7 E10	0.52	-0.40	-0.89
ALB	2.7 E10	0.57	0.53	-0.12
BC	5.6 E09	0.62	-2.32	-2.24
TOTAL	1.0 E11	0.25	-0.35	-0.54

We see that the logistic adjusted weight generally performs better, but not consistently so. In fact the bias increases for NB, MAN, SASK, and the Total. To improve the model, inclusion of some interaction factors like size and farm type, or size and geography was tried but they were rarely kept in the model and when they were, the resulting effects were small and their impact was negligible.

4.4 Conclusion

The selected model did not consistently provide the expected bias adjustment. This may be caused by a low number of factors included in the model or by the fact that significant factors were used in the frame stratification. Future work might include looking for more interactions using the Automated Interaction Detection method used in Section 3. Also, now that the 1993 data are available, the study could be extended.

5. SUMMARY

Nonresponse adjustment through reweighting is now in common use. We have shown that the success of this technique generally depends on having available variables that can be used as good predictors of the nonresponse behaviour. Having such variables, various models can be used to adjust the estimates based on the predicted response propensities. This seems to be the best general approach. Other approaches include using estimation methods such as regression estimators to compensate for the deficiencies of the sample. We have seen that if the regression models are valid, the nonresponse bias vanishes.

We have concentrated here on asymptotic biases. However, there are still many unresolved issues for estimation of variances and construction of confidence intervals. As well, we have not properly addressed the issue of whether or not to use the sampling weights when fitting the nonresponse models. In our examples, the weighted and unweighted versions of the estimated response models gave similar results. This is highly desirable since it confirms the validity of the model.

Nonresponse problems will not go away. A better understanding of the response mechanisms will lead to better survey practices in the long run.

ACKNOWLEDGEMENTS

We are especially grateful to Graham Kalton, who most kindly provided us with a comprehensive list of references on this topic. We would also like to thank Ralph Folsom and Steven B. Cohen for sending us some of their recent work.

REFERENCES

Alexander, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology* 13, 183-198.

Bethlehem, J.G. (1988). Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics* 4, 251-260.

Binder, D.A. and Théberge, A. (1988). Estimating the variance of raking-ratio estimators. *Canadian Journal of Statistics* 16, 47-55.

Britney, H. and Poirier, C., (1992). 1992 Farm Credit Corporation: Design documentation, Internal Paper, Agriculture Section, Business Survey Methods Division, Statistics Canada.

Chapman, D.W., Bailey, L., and Kasprzyk, D. (1986). Nonresponse adjustment procedures at the U.S. Bureau of the Census. *Survey Methodology* 12, 161-180.

Deville, J-C., Särndal, C-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* 88, 1013-1020.

Ekholm, A. and Laaksonen, S. (1991). Weighting via response modeling in the Finnish Household Budget Survey. *Journal of Official Statistics* 7, 325-337.

Eltinge, J.L. and Yansaneh, I.S. (1993). *Weighting adjustments for income nonresponse in the U.S. Consumer Expenditure Survey*. Technical Report No. 202, Department of Statistics, Texas A&M University, College Station, Texas.

Fellegi, I.P. (1980). On adjusting the Pearson chi-square statistic for cluster sampling. *Journal of the American Statistical Association* 71, 665-670.

Folsom, R.E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. *Proceedings of the Social Statistic Section*, American Statistical Association, 197-202.

Iannacchione, V.G., Milne, J.G. and Folsom, R.E. (1991). Response probability weight adjustments using logistic regression. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 637-642.

Judkins, D. and Lo, A. (1993). Components of variance and nonresponse adjustment for the Medicare Current Beneficiary Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association.

Lepkowski, J., Graham, G., and Kasprzyk D. (1989). Weighting adjustments for partial nonresponse in the 1984 SIPP Panel. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 296-301.

Maranda, F. (1989). Proposal for NFS estimation, Internal Paper, Agriculture Section, Business Survey Methods Division, Statistics Canada.

Oh, H.L. and Scheuren, F. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies* (W.G. Madow, I. Olkin and D. Rubin, eds.), 143-184. Academic Press, New York.

Poirier, C. (1994). The causes of non-response in the context of agricultural surveys. Internal Paper, Agriculture Section, Business Survey Methods Division, Statistics Canada.

Särndal, C-E. and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review* 55, 279-294.

Särndal, C-E., Swensson, and Wretman, J. (1992), *Model Assisted Survey Sampling*. Springer-Verlag, New-York.

Singh, M.P., Drew, J.D., Gambino, J.G., and Mayda, F. (1990). *Methodology of the Canadian Labour Force Survey 1984-1990*. Statistics Canada, Catalogue 71-256, Ottawa.

NONRESPONSE DISCUSSION

J. Michael Brick
Westat, Inc.

I would like to thank the authors of both papers for presenting insights on very different aspects of the nonresponse problem. The first paper, "Exploring Nonresponse in U.S. Federal Surveys" by Gonzalez, Kasprzyk, and Scheuren, reports on the activities of a Subcommittee on Nonresponse for the Federal Committee on Statistical Methodology (FCSM). The main focus of this paper is on trends in response rates for federal surveys over the last 10 years. The second paper, "Model-based Reweighting for Nonresponse Adjustment" by Binder, Michaud and Poirier, reports on research in a particular approach to reduce the bias due to nonresponse.

In "Exploring Nonresponse in U.S. Federal Surveys," the authors describe recent investigations of the levels of response rates and trends in nonresponse and refusal rates in federal surveys conducted from 1982-1991. The other parts of the mission of the subcommittee include reporting on correlates of response rates and other related matters. We look forward to reports on those activities later this summer.

The findings on trends in nonresponse rates are especially welcome. In general, these data show that there has been no increase in nonresponse or refusal rates over the last decade. This finding contradicts the conventional wisdom that nonresponse rates, and refusal rates in particular, have been rising due to the increases in the burden on respondents. It clearly shows that it is risky to make general conclusions about this type of phenomenon based on limited data.

Given that response rates seem to be fairly stable over the last decade, the question still remains about the value of judging the quality of survey data over time from this type of data. After all, we are really interested in the quality of the data, not the response rates themselves. However, response rates are the only measure of quality typically produced from surveys. While having something that is measurable is useful, response rates are frequently not very informative of the quality of the data.

Using what Deming calls the "modern" approach to quality control, it is important to recognize that the response rates are really a product characteristic of a survey. Survey methodologists, on the other hand, must concentrate on the process of collecting data. The danger is that examining product variables alone can mask the process related to the quality of the data. For example, if the population surveyed changes so that a group that has generally high response rates is now sampled, then we might expect the response rates to increase. Thus, response rates might change even if respondents willingness to participate remained the same. Similarly, unweighted response rates may be affected by changes in the sample design, such as changes in the sampling rates for different segments of the population.

For survey methodologists, the process must be the most important part of their job. They are responsible for producing survey data of the highest possible quality within the constraints of the survey. In many ways, this paper only peripherally addresses this set of core users, because it does not discuss the process itself. To do this requires more process data, much of which is not comparable across surveys.

The focus on response rates is also limited for other reasons. The important relationship between response rates and costs is not explored in this paper. Cost data are crucial, since the cost of obtaining the same level of response rates may have increased or decreased. Unfortunately, cost data are extremely difficult to obtain and formulate in a manner that is useful for comparison purposes. Developing a general method for analyzing cost data remains, in my opinion, one of the most important unresolved problems in survey research.

As the authors of this paper indicate, improving response rates does not always improve the quality of the survey estimates. This has been misinterpreted by some researchers to mean that response rates are not important. They are important and reasonable efforts should be made to eliminate nonresponse. The real question is related to how much effort should be placed on improving response rates versus reducing other sources of error in a survey. The answer to this question is not always clear, but there are guidelines for reasonable practice. In general, the resources devoted to a source of error (be it sampling or nonsampling) should be proportionate to the size of that error relative to the sum of sampling and nonsampling errors.

In one of their recommendations, Gonzalez, Kasprzyk, and Scheuren state that research should be encouraged by survey administrators as a way to improve the effectiveness of data collection operations. They include a specific recommendation that studies of collecting items to improve post-survey adjustments should be encouraged. I think this is a very interesting choice for the recommendation. It supports collecting items rather than exploring new methodologies. I applaud this direction. Research on collecting additional items is a sound direction to improve our ability to decrease nonresponse bias.

The paper by Binder, Michaud and Poirier, "Model-based Reweighting for Nonresponse Adjustment" is very consistent with the recommendations of the FCSM Subcommittee. Since the type of nonresponse discussed in this paper is not planned, randomization does not apply and a modeling approach must be used to reduce the bias. The authors discuss the results from applying a specific form of modeling nonresponse for two different surveys.

Their efforts to reduce the bias in the estimates for the Survey of Labour and Income Dynamics (SLID), closely parallel work done on the U.S. Survey of Income and Program Participation (SIPP). The Census Bureau sponsored several efforts to examine methods to reduce attrition bias in the SIPP and many of the papers from these efforts will be presented this summer. In the SIPP research conducted at Westat, we used many of the same techniques to model the nonresponse as Binder, Michaud and Poirier, with much the same result.

I would like to make two specific comments on the paper prior to returning to a discussion of the importance of collecting items for reducing nonresponse bias. The first deals with the use of the logistic regression model used to predict response propensity. This work, and our own efforts in the SIPP research, suggest that these types of models are not the most conducive to uncovering relationships between variables when faced with a large number of potential predictor variables. Categorical search algorithms seem better equipped at identifying nested relationships that might be important for reducing nonresponse bias. In this paper, these relationship were uncovered with the search algorithms and added to the logistic regression model. We have used the output from the search algorithms to form cells for nonresponse adjustment directly. Both methods seem to work reasonably well.

The second comment is really a question. The item most related to nonresponse was mover status. People who moved were far more likely to be nonrespondents. This characteristic has obvious appeal and could be of great importance in nonresponse bias reduction. The question is how nonrespondents were defined as movers or nonmovers based on the data available. In most surveys, we do not have these data for nonrespondents. I think it would be useful for other survey methodologists to see how the authors made this classification for SLID.

If an item such as whether or not the person has moved is available, then it can be used to estimate response propensity. Having items that are good estimators of response propensity is critical for reducing bias. I have found that different nonresponse adjustment methods, when properly applied, often have approximately the same performance. This may happen because almost all the methods estimate individual response propensities conditional on auxiliary data. Estimating a response propensity for each unit is very difficult since only one observation is available (the unit either responded or did not respond). Items related to the likelihood of responding at the individual level are, therefore, extremely valuable to improve this estimation.

The authors take an approach that seems most sensible. They include as large a number of items as possible in the model without allowing the adjustments to become so variable as to unduly increase the variance of the estimates. Since nonresponse bias is a function of the covariance between the estimate and the response propensity, items related to both should be entertained. As suggested in this research, the first items included should be those related to response propensity. If it were possible to model these response propensities precisely, then the nonresponse bias for all the estimates would be eliminated. Since the modeling is imperfect, additional items highly related to key statistics from the survey should also be included, where possible.

While longitudinal surveys and studies in which there are frames with substantial data on the sampled units may be able to use this method to reduce the bias due to nonresponse, the more typical situation may be that faced in the Farm Financial Survey. In this case, the data on the frame are limited and do not appear to be predictive of either response propensity or the key estimates. The methods used are not very effective on reducing the nonresponse bias, as might be expected. If the nonresponse is large enough in a survey of this nature, the recommendation of the FCSM to develop a mechanism for collecting items useful for nonresponse adjustment should be seriously considered. This common situation reinforces the importance of keeping nonresponse to a minimum.

DISCUSSION

Joseph L. Schafer
Pennsylvania State University

1. Comments on "Exploring nonresponse in U.S. Federal Surveys"

This paper by Gonzalez, Kasprzyk, and Scheuren (GKS) is worthwhile reading. Section 2 gives a nice historical overview of the development of statistical methods for survey nonresponse. As this section clearly demonstrates, many of the great strides in the practice of nonresponse adjustment, and survey sampling in general, have come about as a direct result of personal interaction between statisticians in federal agencies and in those in academic circles. It is our hope that this type of fruitful interaction will continue in the years ahead.

In the world of surveys, there are many different types of nonresponse. Unit nonresponse arises when the entire vector of survey variables for a sample unit is missing. Item nonresponse arises when individual elements of the vector are missing. As pointed out in the other paper in this session, panel surveys often suffer from wave nonresponse, which occurs when the entire vector of survey variables is missing for a unit at a particular occasion or wave. The GKS paper has introduced a new type of missingness: nonresponse nonresponse, which occurs when members of the federal statistical community fail to report their nonresponse rates to members of the FCSM subcommittee. An even worse type of nonresponse nonresponse occurs when producers of federal surveys do not report basic information on nonresponse--such as nonresponse rates and methods of adjustment--to the users of their data.

It is worthwhile to ask whether nonresponse nonresponse is ignorable, in the sense defined by Rubin (1987). In the FCSM subcommittee's study, the nonresponse nonresponse would be ignorable if the nonresponding statisticians' surveys were representative of all federal surveys in terms of missingness rates, methods of adjustment, etc. Chances are, the nonresponding statisticians had desks that look like mine--paperwork stacked up a foot high all around--and they didn't find the time to return the questionnaire. Or, perhaps they didn't respond because they hadn't been tracking basic information such as nonresponse rates, and compiling the information would have required an unusual amount of effort. Whether those characteristics of the nonrespondents are systematically related to the basic study variables is anyone's guess. In the common, non-technical sense of the word, however, this nonresponse to the FCSM subcommittee's study is probably not ignorable; we shouldn't ignore it. If members of the subcommittee found it so difficult to obtain even basic information on nonresponse from the very people who produce the survey, imagine how hard it must be for the average data user to do the same thing.

It was astonishing to see that almost half of the establishment surveys in the study had no tracking at all of the various components of nonresponse. We hope that in the future, producers of federal surveys will take to heart the recommendations of this report and devote a little more time and effort to studying and documenting the levels and causes of nonresponse. At any rate, it is comforting to learn that the basic perception that many of us had--that nonresponse rates have been rising over the past few years--may be only a perception, and the sum total of the various nonresponse components may not have changed very much.

Reading this paper made me think a lot about the future, and what directions we should take in our research on nonresponse and nonresponse adjustment. The suggestion near the end of the paper--that data collectors and statisticians work together as a team, sharing information in a way that is mutually beneficial--is especially thought-provoking. If this were done, it would open up entirely new avenues for developing improved methods of nonresponse adjustment. As shrinking budgets force us to re-allocate our resources, it may not be possible to continue to spend so much money chasing after nonrespondents, trying to get them to hand their data vectors over to us. It's becoming increasingly likely, for example, that the Census Bureau will not have the resources in the year 2000 to follow up on every housing unit that doesn't mail back its census form with a personal interview. Statisticians and data collectors should start to think long and hard about which nonrespondents they should attempt to follow up, and how persistently they should attempt to do so. They need to weigh the costs and benefits involved in converting refusals and not-at-homes to responses, so that they may decide whether the money might be better spent elsewhere.

In regard to this point, I would like to make two basic observations. The first observation is that not all nonrespondents are equally important. In industrial settings, statisticians have to carefully design their experiments. Trial runs are typically expensive, and they have to decide under what values of X they should collect their next value of Y. They know that certain, carefully chosen combinations of Xs will help them to "nail down" the quantities of interest much better than other combinations of Xs, so they choose their values of X carefully before spending additional money to collect another Y.

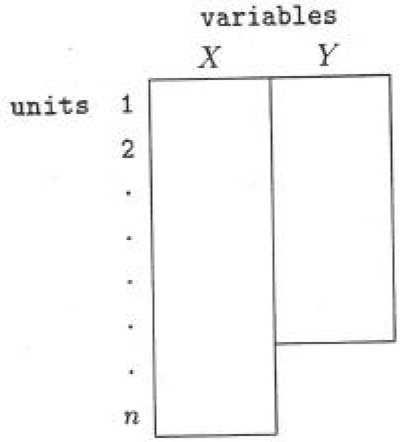


Figure 1: Data after the initial phase of collection

In a survey setting, there is typically an initial phase of data collection in which every unit in the sample has been contacted once, or at least an attempt has been made to contact them once. After this initial phase there will be a pool of nonrespondents, and there will be (a limited amount of) money to spend on attempts to convert some of them to respondents. Which nonrespondents should we go after first? To answer this question, we should look at their Xs. At this point, the data will resemble the diagram in Figure 1. Variables available from the sample frame, denoted by X, will known for all sample units. Survey variables, denoted by Y, will be available for the respondents only. Let us partition Y as $Y = (Y_{obs}, Y_{mis})$ where Y_{obs} denotes the values of the survey variables for the respondents, and Y_{mis} the values of the survey variables for the nonrespondents. It should be possible to build a regression model for the distribution of Y_{obs} given X, and then use this model to guess or predict what the values of Y_{mis} will be. These predictions may then form the basis for ranking the nonrespondents according to the priority with which they should be followed up.

Let $P(Y_{obs}|X, \theta)$ denote the form of the model fit to Y_{obs} given X, where θ represents some unknown parameters. Let y_i^* denote an element of Y_{mis}. To the classical survey statistician, y_i^* is a fixed, unknown constant. To the Bayesian, however, it is unknown and therefore a random variable. The uncertainty about y_i^* can be expressed Bayesianly as

$$V(y_i^*|X, Y_{obs}) = EV(y_i^*|X, Y_{obs}, \theta) + VE(y_i^*|X, Y_{obs}, \theta),$$

where the outer moments are taken with respect to $P(\theta|X, Y_{obs})$, the posterior distribution of the unknown model parameters. The first component on the right-hand side, $EV(y_i^*|X, Y_{obs}, \theta)$, represents the residual variation of y_i^* about its predicted value from the regression model. The second component,

$VE(y_i^*|X, Y_{obs}, \theta)$, represents the uncertainty in the regression prediction itself. This decomposition suggests that if we want to minimize our uncertainty after a limited amount of followup, we should target as high priority those units that (i) have a high amount of residual variance, and (ii) have high leverage for estimation of θ. In other words, we should try to follow up the units (i) whose values of Y cannot be predicted well by our model, and (ii) whose values of X are unusual and thus, if they were converted to respondents, could greatly improve our ability to predict the missing Y values for the other nonrespondents.

In addition to the predictive variance of y_i^*, we also need to consider the probability that a nonresponding unit can successfully be converted to a respondent. Even if the predictive variance for a particular unit is high, it may not make sense to attempt followup if the followup operation is likely to be unsuccessful. This suggests construction of another regression model to predict the probability of successful followup--or, perhaps, the cost of successful followup in terms of number of attempts, field worker time, etc. Data for fitting this model might come from similar survey operations of the past, perhaps updated by data from the current survey as they become available.

As data collectors begin to share information with statisticians on an ongoing basis, one can imagine the development of a continuous-loop feedback system in which the field-operations unit provides data on respondents as they become available, and the statistical unit processes the information, updates the parameters of its regression models, and decides which of the remaining nonrespondents should be designated for followup.

The second general observation that I would like to make is that not all nonresponse mechanisms are the same. From a theoretical standpoint, it is useful to classify nonresponse mechanisms into two categories: mechanisms that are ignorable and mechanisms that are nonignorable. Using the notation developed above, an ignorable mechanism is one in which the probabilities of response do not depend on Y_{mis} after accounting for dependence on X and Y_{obs}. Ignorable nonresponse mechanisms tend to be easier to deal with than nonignorable ones, and virtually all methods of nonresponse adjustment in use today make some implicit assumptions of ignorability.

From a practical standpoint, however, nonresponse mechanisms should probably be classified into a slightly different dichotomy: mechanisms that are known to be ignorable, versus mechanisms that are not known to be ignorable. Mechanisms that are known to be ignorable include those in which the missing data are missing by design. Surveys that employ double sampling, matrix sampling, etc. result in rectangular datasets with patches of missing data that are missing by design; the data are unrecorded because the data collector never intended to collect them. When data are missing by design, ignorable missing-data techniques may be applied without fear of introducing bias. The more insidious type of missingness mechanism is the unknown type. When the nonrespondents are a self-selecting subsample, we do not really know how strongly the selection process may be related to the missing data Y_{mis}. When faced with missingness of this type, the only thing that a practitioner can usually do is to apply some ignorable missing-data technique and hope for the best--i.e. pray that any biases incurred by nonignorability will not be severe.

As long as resources for data collection are finite, a certain amount of missing data will be inevitable. But the point that I want to make is this: By intelligent allocation of resources in the followup operation, we may be able to convert a substantial amount of the missing data that would ordinarily be of the type "unknown" to the type "ignorable." In a typical followup operation of today, attempts are made to follow up nonrespondents in a rather haphazard (i.e. unplanned) fashion until the resources run out, at which time the data collectors close out their operation and get on with their lives. Decisions about which nonrespondents are to be followed up are not made by a central decision-making unit, but are made in the field by supervisors or by the interviewers themselves. It may be that the field staff is placing high priority on the nonrespondent units that appear to be easy to get, thereby attempting to minimize the number of nonrespondents that remain after closeout. Although minimizing the nonresponse rate is a laudable goal, the end result is that all of the nonresponse that remains after closeout is of the type "unknown." From a statistician's point of view, a better strategy may be to concentrate one's resources on obtaining data for a probability sample of nonrespondents, a sample that is guaranteed to be representative of the nonrespondent pool. Even if data for these units are expensive to obtain--e.g. requiring a large number of call-backs--and the overall rate of missingness in the end is higher than it would have been if the followup decisions were made by field staff, the end result will be that the missing data for nonrespondents that are not included in the followup sample will be of the type "ignorable."

As a scientist, I would be willing to trade a few percentage points of missingness for a guarantee that (at least most of) the missing data are ignorable. I suspect others would as well. The tradeoff between the cost of missing data versus the benefit of knowing the missingness mechanism is a subtle but important issue to which statisticians ought to pay more attention in the future.

2. Comments on "Model-based reweighting for nonresponse adjustment"

This paper by Binder, Michaud, and Poirier (BMP) discusses in detail the methods used by StatCanada to model response propensities in two of its ongoing surveys. By reading this paper, and examining the tables of regression coefficients, one develops an excellent sense of what factors may be related to nonresponse in demographic and establishment surveys. It is interesting to note that in regard to these two surveys, BMP and StatCanada seem to be following the recommendations of the authors of the previous paper: clearly documenting the rates of nonresponse, the factors related to nonresponse, and the methods used for nonresponse adjustment.

Throughout this paper, the value of adopting a model-based approach to nonresponse adjustment clearly shines through. By constructing an intelligent model for the nonresponse mechanism, one is able to carry out a nonresponse adjustment using many more explanatory variables than would otherwise be possible using a more traditional approach. In a more traditional approach, one would form adjustment cells by crossing the classes of every explanatory variable. This would be equivalent to building a response-propensity model that includes all possible interactions among the explanatory variables, whether or not those interactions have much predictive power (and they often don't). The modeling approach adopted by BMP allows them to exclude unimportant high-order interactions, and instead include main effects for a larger number of explanatory variables.

One issue that may deserve a little more attention is what should be done with the estimated response propensities once they are calculated. On this point, statisticians north of the U.S.-Canada border tend to use the reciprocals of these probabilities as factors in the nonresponse weighting adjustment. Statisticians south of the border tend to form classes--e.g. by dividing the estimated propensities into quintiles--and reweight the observations within these classes. Little and Rubin (1987) comment that the latter may sacrifice a little bias for the sake of reduced variance and robustness against model failure. I wonder if anyone has done a comparison of the two methods in a realistic setting to see which one tends to perform better.

Another important issue, which is perhaps beyond the scope of the BMP paper, relates to the underlying philosophy of response-propensity weighting. Response-propensity weighting attempts to control and reduce nonresponse bias. The theory of propensity scores (Rosenbaum and Rubin,

1983) says that if reweighting could be performed on the basis of the actual (as opposed to estimated) propensity scores, then the reweighted distribution of the respondents would not be systematically any different from that of the respondents--in other words, nonresponse bias would be eliminated. But of course, bias is only one component of error, the the other being variance. As pointed out by Little (1986), response-propensity weighting may do very little to control variance. One might also want to consider forming weighting classes on the basis of variables that are highly correlated with the survey variable of interest (if any are available), so that variance might also be reduced. Perhaps forming weighting classes on the basis of two variables--a linear predictor of the response propensity, and a linear predictor of the survey variable of interest--may be a reasonable approach.

The methods described in the BMP paper also raise a number of theoretical issues that deserve a closer look by statisticians in the future. One question involves the use of complex, automated variable-selection procedures to choose a model for nonresponse adjustment. Whenever the form of the model is chosen through examination of the sample data, the procedure used to select the model should be considered a part of the overall method of nonresponse adjustment. Any calculations of bias and variance--whether carried out analytically, or by simulation, jackknifing, etc.--should recognize that the model itself is sample-dependent and therefore random, and the model selection procedure must therefore be included in the calculation.

Another, perhaps more basic, issue pertains to the criteria used for selecting the model. BMP emphasized the principle of parsimony, eliminating variables whose effects were not significantly different from zero. They also included variables tended to reduce the number of extreme residuals. In the end, they were left with models that had very few parameters relative to the size of the dataset. I was left with the feeling that they could have included more variables, provided that the model-fitting could be accomplished in a reasonable amount of time. The usual criteria given by textbooks on regression modeling--high R^2, low prediction error, all coefficients statistically significant, and so on--are usually appropriate when the goal is to acquire some scientific understanding of how the response variable is related to the pool of potential predictors. When the goal is not necessarily scientific understanding, but adjusting for nonresponse, however, it is not yet clear what model-selection criteria statisticians ought to be using.

Finally, another issue that deserves further investigation is the proper role of sample-design information in the construction of response-propensity models. I suspect that many statisticians would attempt to include design information by fitting logistic regressions with standard software, including the case weights (inverses of the sample-selection probabilities) in the fitting procedure. The correctness of such a procedure is not at all clear. If the goal were to estimate regression coefficients for predicting nonresponse for the entire population, then including the case weights would be appropriate. The goal in response-propensity modeling, however, is to estimate the probability of response for the units in the current sample. To the extent that this response propensity is related to covariates describing the sample design (e.g. stratum or cluster indicators), those covariates ought to be included in the model somehow. But merely weighting the cases by their inverse probability of selection is probably not sufficient to guarantee that the special features of a dataset that arise from complex sampling, such as clustering effects, are appropriately described.

Additional references

Little, R.J.A. (1986), "Survey nonresponse adjustments for estimates of means," International Statistical Review, 54, 139-157.

Rosenbaum, P.R. and Rubin, D.B. (1983), "The central role of the propensity score in observational studies for causal effects," Biometrika, 70, 41-55.

TOWARD AN AGENDA FOR THE FUTURE

**OLD DIRECTIONS, NEW DIRECTIONS:
CONCLUDING REMARKS**

Norman M. Bradburn
NORC

I have been asked to take on the unenviable task of giving an overview of the last two days and, on the basis of that overview, to give some thoughts on future directions. In order to have some empirical idea of what the directions--new or old--have been, I looked at the topics covered in this symposium and compared them with the topics covered in the first symposium four years ago.

If we view the programs of the conferences as unobtrusive measures of what are the central research directions in statistical methodology, at least in the minds of the organizers of these conferences, we can get an idea of how things have changed in the last four years. For comparison, I coded the session topics into four themes: 1) conceptualization of statistical measures--that is, what we are measuring; 2) design and analysis of statistical surveys; 3) operational issues in data collection; and 4) dissemination of data to users. The result of the comparison is seen in Table 1.

Table 1

Themes of Papers Given at 1990 and 1994 Seminars

	1990	1994
Conceptualization	0	1
Design/Analysis	5	4
Operations	6	5
Dissemination/Use	1	2
Total	12	12

Six (or seven) of the 12 sessions in each year are on the same topics. They are: 1) Administrative Records, 2) Longitudinal Surveys, 3) Data Editing, 4) Disclosure Limitation, 5) Computer Assisted Data Collection, 6) Cognitive Testing of Questions. The seventh, Incomplete Surveys, could be considered at least partially the same topic, i.e., Coverage Problems in 1990 and Non-Response in 1994.

Without looking at the new topics, one might view this as little change; but when we look at the particulars of those new topics, it seems clear that there are major shifts in direction. New session topics were: 1) Customer Surveys, 2) Respondent

Incentives, 3) Small Area Estimation, 4) Time Series revisions, and 5) Economic Classification Revision.

What is it about the new topics that leads me to think we are moving along some significant new directions? I see the primary theme of these new topics as the need to rethink some old problems in order to cope with three types of change: 1) changes in social conditions, 2) growth in demand for data and 3) technological change. Let me comment on these in turn.

1. **Changes in social conditions.** Society is becoming more diverse along many lines. It is not just the increasing cultural diversity that is so much remarked upon but also the increasing diversity in household structures, in a changing occupational structure, and in the globalization of the economy. Our traditional ways of collecting data and the statistical categories we have used are creaking and in need of revision. Surveys are harder to carry out, so we become concerned about questionnaire development and incentives for respondents. It is even necessary to rethink how we conduct the decennial census. Our classification schemes for economic and social categories are no longer adequate, so we rethink the SIC and SOC and the problems of revising time series when you change the measuring instruments.

2. **Growth in demand.** It is a truism these days that we live in an information society. The demand for data grows apace with the proposals for health care reform and the health data network and the increasing demand for evaluation of social services and social programs. Concern for international economic competitiveness leads us to rethink the categories we use to collect economic statistics, such as how we measure the trade balance and what industrial and occupational classification scheme we use. It is important to have an economic statistical system that is consonant with that of other countries.

There are two implications for statistical methodology of this increase in demand for data. The first is that it requires a strong customer orientation to be responsive to the changes in demand. We need to know what the consumers of statistical information require in order to carry out their work efficiently. Second, because of the high cost of data collection in a world of tight budgets, we need to develop alternative ways of doing things--not just tinkering around the edges, but some fundamental changes ("reengineering," to use the current buzz word).

3. **Technological change.** CATI, CAPI, CASI, Audio-CASI. computer-assisted everything. The ability to handle large data files cheaply has changed the economics of data analysis, if not of data collection. Data are accessible on Internet; you can call in for SIPP data. All of these advances whet the appetite for more data--more quantitatively, in more detail, and more easily accessible.

At the same time, increased availability of large data sets increases the risk of disclosure of data that compromises the privacy and confidentiality promised to respondents and upon which our ability to get accurate data depends. Confidence in the statistical system must be maintained, and that means preventing disclosure on individually identifiable data for non-statistical purposes. Hence our renewed concern for data disclosure limitation principles and methods.

Where does all this change lead us? First, I think it leads us to a fundamental rethinking of some of our major data systems. We have already seen the revision of the CPS. We have heard here about proposed revisions in the Standard Industrial Classification system and the Standard Occupational Classification. These are truly major changes in our statistical thinking.

Work has begun on a revision of the CPI. We are rethinking the way we conduct the census. Dare we think of an administrative records enumeration in 2010 with the long form data supplied by continuous measurement? Can the national accounts be recast to include supplemental accounts for the environment (so-called Green Accounting) and recast for other non-market sectors as well? Will data bases become public utilities that can be used for many purposes? For example, the basic health care file, if it were to come to pass, could be a prime source of data for an administrative records census or a data base for assessing the outcomes of health care services. A Master Address File that was available to everyone, the Census Bureau, the Post Office, state and local governments, the private sector, would be a true common good.

A final theme that cuts across all of our sessions here is that we must take a more international perspective. We can learn from other countries that have been confronting some of the same problems. There is also increased pressure to have consistent statistical measures across the developed countries at least, so that international comparisons can be made.

The spur of competition can be good. The State governors adopted a set of Educational Goals for the year 2000. One of these is for the United States to be first in science and math education by the year 2000. In the Economist magazine's annual rankings of governmental statistical systems, the United States has ranked about 5th for the last few years. Perhaps we should take as our goal to replace Statistics Canada as first in the world among statistical systems by the year 2000.

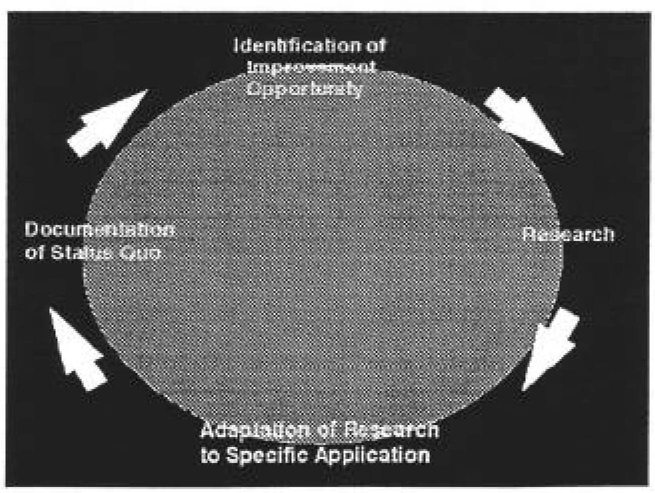
Future directions? From what we've heard here, we may not be able to know precisely how to proceed, but we will have to move forward with an international perspective, a willingness to rethink our methods, our tools, and our purposes.

Toward an Agenda for the Future

Robert M. Groves
University of Michigan and Joint Program in Survey Methodology

It is obvious that the rate of change of the society and economy being measured by statistical agencies is increasing. We have only to observe the ubiquity of merger, acquisitions, and turnover of firms, partly a reflection of the ongoing creation of novel combinations of services and goods. In the demographic domain, we are faced new living arrangements, nontraditional household definitions, and homelessness at levels of occurrence that challenge traditional methods of measurement. It now seems clear that the rate of change of social and economic measurement systems needs to approximate that in the society. We have heard over the last two days much evidence of this.

Figure 1 is a version of charts that are popular among those in the quality movement. It is used to describe the process of continuous improvement key to that philosophy. As "reinvention," "restructuring," "re-engineering," "customer orientation," "total quality management," and related concepts make their way into statistical agencies, it might be useful to reflect how the technical research and development presented over the last two days relate to the process described in such a chart.



The process begins, even prior to the contents of this chart, with the identification of customer needs and desires. Given that as a basis, first the organization needs to document what it does and how the processes of the organization are now conducted. This documentation leads to identification of potential areas of improvement, "identification of improvement opportunity." Whether the process can or cannot be improved is a matter to be resolved in a "research" step. Then the results of the research are applied in "adaptation of research to specific application." At this point, the process begins again, this time with the production process operating at a higher level of overall efficiency, but still subject to improvement.

One aspect that differentiates research in government statistical agencies from research in academia is the last step, application of research findings in practical settings. This fact is accompanied by organizational roles that perform these activities in conjunction with the researcher. The researcher is joined by a manager, who is charged with the responsibility of directing the processes undergoing change.

It's useful to review each of the four steps from the vantage point of a program manager of a data series and a researcher or scientist, both located in a government statistical agency.

Documentation

Documentation is often viewed as the bane of those who develop and conduct data collection, processing, and estimation programs. It is most often not done (or not done completely) for many statistical series. When done, it is often done long after the introduction of a new process, and then completed by someone who was not part of the creation of the process. The lack of good documentation on design, processing, and estimation features of statistical series can be a nontrivial impediment to their improvement. For example, when there is clear evidence that a process is not performing as desired (e.g., nonresponse rates too high, edit processes yielding contradictory data), improvement is often delayed by lack of clear documentation of exactly what is being done. When lack of documentation is joined by high turnover among staff and computer based systems, it is often the case that workers implementing a computer-based process truly don't know what they are doing.

From the management perspective, documentation allows knowledge of what processes are the staff are performing. It permits the manager to look for opportunities to streamline processes, combining steps and improving flow. For the researcher or scientist, documentation is the equivalent of gathering observations about possible causes of outcome variables. It is the inductive part of the scientific method -- assembling information that suggest hypotheses for sources of weaknesses (and thereby opportunities for improvement).

Why is documentation an important step in the process of continuous improvement? Improvements on undocumented systems are limited to those performing the *status quo* activities because only they know what is being done. This limits the set of persons who can be called upon for suggestions for improvement. This limitation is exacerbated in organizations where staff implementing programs do not have rudimentary testing and experimentation skills. It seems clear that without documentation of best practice, agencies can't get to "the cutting edge" that the benchmarking aspect of the quality movement espouses.

Opportunities for Improvement

The use of Pareto curves prompts managers to focus on the problems that are causing the greatest loss of quality or efficiency in production. This is a vehicle to reallocate budgets to save money on expensive processes by investing money in design and retooling phases. Given most government budget systems, an investment in a research and development project to improve one component of a statistical series generally means that some other components will incurred reduced funding. The manager needs the courage of his/her convictions that such an investment can lead to cost efficiencies at a later point.

For the scientist the "identification of improvement opportunities" step often also means a search for experiments on design features common to many surveys. Often important weaknesses of one statistical series are shared by other statistical series. If the researcher can identify such cross-series problems, then single research projects can offer improvements to several data series at once.

In one sense, this seminar can fulfill a similar function. If one agency discovers improvements that are applicable to the work of another, significant research savings are possible.

Research

The research step from the management perspective is a search for independent evidence that change involving risk was well-founded. The research step tests some model of the real change in the production system. To the extent that the research step perfectly mimics the real production setting, the inference from the experiment will apply to the production process. In this sense, research is a risk reduction tool for the manager. Change without research runs higher risks of no improvement or even loss of quality in statistical series.

For the scientist, only experimentation offers the proper grounding for application of some discovery. The scientific method underscores the need for explicit contrast under similar conditions of the new process with the old, before recommendations about change can be well-founded. Much research in social and economic statistics is "applied," with its findings relevant to ongoing statistical series. Applied research is the type most often practice in statistical agencies (and reported in conferences like this one). The cheapest research (in the long run) is theory-driven, often motivated by basic research. This is the research that offers solutions for large sets of surveys and census operations. Choosing the right blend of support for basic and applied research in government agencies takes great wisdom. One without the other is vacuous.

Adaptation of Research to Specific Application

This step is typically the focal point for program managers. From the standard managerial perspective, the real work (and associated risks) of change begin here.

The issues of concern to the manager always concern whether the costs of the change are smaller than the benefits. Sometimes real courage on managers' parts is required in organizational cultures where failed change is punished by withdrawal of the support of superiors.

From the scientific perspective, this step typically requires speculation about the inferential limitations of the research into alternative solutions. Scientists are generally well-trained in hypothesis generation, but not particularly well-schooled in decision-making with imperfect information. In the world of statistical agency programs, there is rarely or never perfect information. Indeed, in any application of a scientific discovery there is some inferential leap that goes beyond the research. Many scientists are uncomfortable with such leap, but scientists in statistical agencies, to be responsive and useful to their agencies, must develop these skills. Clearly, this is most easily accomplished when they work in partnership with their colleagues on the program side.

All of the steps in Figure 1 are difficult, but this step seems most fraught with difficulties and failure to communicate between the managerial and the scientific cultures within statistical agencies.

Implications for Future Symposia, the Working Paper Series, the System

What do these observations have to do with a symposium like the one we have all just experienced? Let's examine the content of this symposium. There was clearly more attention on documentation and adaptation of research to applications, than on methods used to identify opportunities for improvement in data series, or in the findings of more basic research. This is an appropriate mix, I believe, for the purposes of the conference -- an attempt to disseminate new developments of applied utility.

Given that focus, future symposium might reconsider the format of a scientific conference, where a researcher presents his/her work and a discussant critiques the scientific merits of the work. Perhaps for the purposes of application of new discoveries, workshops and didactic seminars might be more useful.

One method of doing that is to use the Federal Committee on Statistical Methodology to identify agencies using "best" practices in various components of statistical work, and to encourage them to mount such workshops, perhaps with some collaboration with scientists and practitioners outside the Federal statistical system.

It is clear that the Federal statistical system can profit from the successes of integrating science and management, of adapting research discoveries to practice that exist in some agencies, but need to be spread more widely across all agencies. What the system does not need is more papers assessing ideas alone. It does need more papers describing how new ideas were assessed, molded to a specific problem, and systems changed to incorporate them. Problems of statistical system are not solved by ideas alone, but by ideas implemented to produce innovation in data series. Presentations in future symposia might be collaborations of program managers and scientific staff, addressing the four steps of Figure 1 for a specific survey or set of statistical activities, from their two different perspectives.

I feel more strongly that the symposium should continue annually, than that it should take on a particular format or offer particular content. Regardless of its format, it is a gathering of the clan within the statistical system, allowing them to meet and converse with one another, to compare techniques, to try out new ideas removed from their organizational cultures, to renew their commitments to improved quality in statistical activities. This will always have value.

TOWARD AN AGENDA FOR THE FUTURE

Katherine K. Wallman
Office of Management and Budget

Over the past two days, we have heard many ideas and suggestions for how we can build upon and improve the efforts we are undertaking through the Federal Committee on Statistical Methodology (FCSM). The "new directions in statistical methodology" we might pursue range from updating and expanding our technical contributions to broadening and deepening the understanding and application of our work. Underlying all of the proposals are two fundamental and, I believe, shared perceptions: first, that the FCSM's activities to date have been of considerable value, and second, that there are opportunities we should pursue to further the FCSM's goal of improving the quality of data produced by the Federal statistical system for our community of users within and outside the government.

In his keynote address for this seminar, Graham Kalton reminded us of the four functions for the FCSM that Margaret Martin outlined at our 1990 Seminar on the Quality of Federal Data. These functions do indeed serve as a useful framework for considering the committee's future agenda, and I would like to return to them in highlighting in particular a few of the suggestions that have been made for ways we might broaden and deepen the understanding and application of the FCSM's work.

The first of these functions is the exchange of knowledge, techniques, and experience among committee members. As has been noted, the members of the Federal statistical community who serve on the FCSM and its subcommittees often carve out time for these activities from their already overly full calendars. The result may well be that they must focus rather narrowly on the task at hand and may forego the opportunity for more informal dialogue. We need to think of ways to give these professionals "permission" to engage in less structured conversation. As a first step, the leaders of the statistical agencies should be encouraged to view the work of the FCSM as a priority for their staff members. In part, we may be able to address this matter by familiarizing the agency leaders more fully with the committee's products.

The second function outlined for the FCSM is the provision of "state of the art" reports to encourage best practice among a broader group. With respect to the "state of the art" aspect of the committee's work, I think there is some considerable agreement that while an FCSM report may meet this challenge when it is issued, in some cases the art is changing so quickly that a working paper may become if not obsolete at least outdated more rapidly than we would hope. Perhaps we should consider new forms of publication -- "loose leaf" printed versions and/or electronic versions that could have sections updated, rather than waiting several years to revise the entire paper. The challenge of sharing our work with a broader group is key. On the one hand, as Graham Kalton elaborated, we need to reach out beyond our Federal walls in developing our working papers to capitalize on the expertise resident in the private sector, foreign governments, and other arenas. His remarks provide a number of useful suggestions the FCSM should pursue. On the other hand, it is absolutely essential that we continue to strengthen efforts that have been undertaken to disseminate the FCSM's products far more widely, both within and outside the Federal Government. This means not only distributing our product to additional audiences, but also extending the seminars, workshops, and other forums where the subject matter of the papers can be explored in greater detail and become more useful to those less familiar with the content.

In terms of the third function suggested by Margaret Martin, recommending areas for improvement and needed directions for research, once again the need for greater outreach has been highlighted, and some useful paths to pursue have been suggested. At the same time, we need to take some care in reaching out that we do not ask the FCSM to be "all things to all people." As has been noted, the members' plates are quite full already, and we should not overload them. Yet there should be new ways to meet reasonable demands, and we as a community should put some of our energy into brainstorming about alternative ways of operating to address these needs.

The consensus building role suggested as a fourth function for the FCSM is surely an area where the committee can make a contribution. In my view, however, to suggest that the FCSM should be responsible for obtaining consensus on issues such as definitions, concepts, and classifications moves us to an arena that is beyond its mandate. Many of these activities involve domains of expertise beyond the specialties of the FCSM membership, as well as exceptionally labor intensive tasks. Through the cooperation of the statistical agencies and other relevant components of the Federal Government, we are pursuing a number of these challenges. We would welcome the suggestions for further work in this area that might arise as a result of the FCSM's deliberations.

In closing, let me note our appreciation to the Council of Professional Associations on Federal Statistics for bringing us together once again, to the session organizers and presenters for their substantial contributions to the success of this seminar, and last but not least to the members of the FCSM and its devoted leader Maria Gonzalez. We look forward to planning an encore to be held in 1996.

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