NCHSR

RESEARCH PROCEEDINGS SERIES

Health Survey Research Methods

Third Biennial Conference

Reston, Virginia
May 16–18, 1979

This conference
was jointly sponsored
by the National Center
for Health Services Research
and the National Center
for Health Statistics,
through NCHSR grant no.
HS 03271.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Office of Health Research, Statistics, and Technology
National Center for Health Services Research

DHHS Publication No. (PHS) 81-3268

Editor: Seymour Sudman
Department of Business Administration and Survey Research Laboratory
University of Illinois at Urbana-Champaign

Library of Congress Card No. 80-600136

Contributors

Lu Ann Aday
Ronald Andersen
Marilyn Bergner
Marc Berk
Lisa F. Berkman
Charles F. Cannell
Diana Cook
Alfred Dean
Charlene E. Depner
Barbara S. Dohrenwend
Bruce P. Dohrenwend
Jack Elinson
Walter M. Ensel
Lynn A. Evans
Esther Fleishman
Floyd J. Fowler, Jr.
Matilda Frankel
Howard E. Freeman
Robert R. Fuchsberg
Helen C. Gift
Louis J. Goodman
Bernard G. Greenberg
Robert M. Groves
Ruth S. Hanft
Elizabeth B. Harkins
Mimi Holt
Daniel G. Horvitz
Morton Israel
Robert A. Israel
Lynn E. Jensen
Charles D. Jones
Lawrence A. Jordan
William D. Kalsbeek
Judith Kasper
Nan Lin
John D. Loft
Alfred C. Marcus
Kent H. Marquis
Linda S. McCleary
Peter V. Miller
Lois Oksenberg
Donald L. Patrick
Clyde Pope
Sharon Reeder
Gerald Rosenthal
George S. Rothbart
Naomi D. Rothwell
Sam Shapiro
Patrick E. Shrout
Eleanor Singer
Monroe G. Sirken
Seymour Sudman
D. Garth Taylor
Lois M. Verbrugge
Thomas T.H. Wan
Richard B. Warnecke
Gail R. Wilensky
Sherman R. Williams
Robert Wright
Richard Yaffe

Planning Committee

Charles F. Cannell, University of Michigan
Joseph L. de la Puente, NCHSR
Jack Elinson, Columbia University
Bernard G. Greenberg, University of North Carolina at Chapel Hill
Daniel G. Horvitz, Research Triangle Institute
William H. Kitching, NCHSR
Linda S. McCleary, Executive Secretary, NCHSR
Leo G. Reeder, University of California at Los Angeles
Seymour Sudman, Chair, University of Illinois at Urbana-Champaign
Elijah L. White, NCHS

Dedication

The Third Biennial Conference and these proceedings are dedicated to the memories of Leo G. Reeder and Elijah L. White, founding members of the Planning Committee for these conferences, who both died in tragic accidents in the year prior to this conference. We remember them with affection and shall miss them sorely. A discussion of their lives and impact will be found in these proceedings.

The Planning Committee

Foreword

The Third Biennial Conference on Health Survey Research Methods was held in Reston, Virginia in May 1979. The National Center for Health Services Research (NCHSR) and the National Center for Health Statistics (NCHS) have continued their sponsorship of these conferences because of the importance of survey research in gathering data for use in policy formulation for the Nation. These agencies, mindful of the need for increasingly valid and reliable information by means of surveys on measures of health levels, and on the use of health services and their availability, have sponsored research in survey methods designed to improve these techniques.

Although there have been considerable advances in survey research methods, systematic studies in this area are of relatively recent origin. Professional methodologists are well aware of the limitations of survey methods and understand the many areas in which much improvement can be achieved by means of careful studies; such is not the case, however, among many research scientists. Moreover, it is important that the information developed by methodological studies be shared and disseminated rapidly and in a form that is usable. These Biennial Conferences serve the function of disclosing current findings and developing strategies for future work that will advance the state-of-the-art of survey research. The published reports of these deliberations permit the dissemination of current thinking to the health services research community in general.

NCHSR and NCHS, in response to the requirement for better survey methods and the dissemination of these findings, supported this conference through the Survey Research Laboratory of the University of Illinois at Urbana-Champaign. This publication contains information that will aid survey researchers in carrying out better studies and will be a significant contribution to the goals of dissemination and use of research findings. The actual use of these findings will be the ultimate test of these conferences. It is our hope that the outcome of this conference will be significant in the design and quality of research and in the analysis of studies based on survey methods.

Gerald Rosenthal, Ph.D.
Director
National Center for Health Services Research

Dorothy P. Rice
Director
National Center for Health Statistics

May 1981

Acknowledgments

We are pleased to acknowledge both the financial and intellectual contributions of Gerald Rosenthal, Director, National Center for Health Services Research, and Dorothy P. Rice, Director, National Center for Health Statistics. This is the third conference jointly sponsored by these two agencies, and their continuing encouragement and support are most gratifying.

Mr. William Kitching of the National Center for Health Services Research continues to be the key person in virtually all phases of conference and post-conference planning. His efficiency and good judgment are vital to the conference's success. He has been ably assisted by Linda McCleary, who, in her role as Secretary of the Planning Committee, has ensured that lines of communication remain open.

Mr. James Smith and his capable staff in the Conference Management Branch, National Center for Health Statistics, were enormously helpful in planning conference logistics and solving problems while the conference was in session.

Mati Frankel of the Survey Research Laboratory, University of Illinois, played a vital role in pre-conference communications with participants and in the compiling and preliminary editing of the proceedings. The final editing and preparation of the index were done with great skill by Mary A. Spaeth, SRL's Coordinator of Survey Research Information Services, with the assistance of Antje H. Kolodziej.

The Planning Committee

Contents

Introduction

Seymour Sudman, University of Illinois at Urbana-Champaign

This is the third in a series of conferences to synthesize the current state of the art of survey procedures relevant to health surveys. These conference proceedings aim to disseminate the conference findings to a broad audience of health researchers, teachers, and other users and collectors of health survey research data.

The format for the conferences has changed over time. At the first conference, held in 1975, topics were specified but there were no papers prepared in advance. Invited researchers came and reported their experiences in each of the selected topic areas. The second conference, in 1977, was somewhat more structured. For each topic discussed, a speaker was invited to present a position paper summarizing major findings and indicating the major unsolved problems and research needs. Each presentation was followed by a lengthy floor discussion.

In this conference, even more structure has been introduced. From among a large number of papers contributed on the designated topics, three papers were selected for presentation. A discussant was invited to comment and elaborate on each paper. Open discussion on each paper and on the general session topic followed. All of the formal papers and discussant comments and summaries of the open discussions are included in these proceedings.

As at previous conferences, the key roles in preparing the proceedings have been those of the session chairpersons and recorders, who have summarized and synthesized the open discussions. From this synthesis, specific recommendations are made for new research or for policy action in health data gathering activities. While the prepared papers are, in our judgment, of high quality and represent the frontiers of research in their respective areas, a fuller understanding of the meanings and limitations of the results is obtained by reading the open discussions.

The Planning Committee for this conference, as for the earlier ones, assumed the responsibility of selecting topics of the highest methodological and substantive interest on which significant research was being conducted. Thus, some of the topics continued discussions started earlier, while some topics were entirely new. Among the topics that had not been formally discussed in earlier conferences were "Provider or Physician Surveys," "Use of Informants and Multiplicity Estimates," and "Methodological Issues in Developing Standardized Measurement of Long-Term Behavior."

Topics such as "Health Interviews by Telephone" and "The Reinforcement and Feedback to Respondents by Interviewers," as well as "The Use of Diaries and Panels in Health Services Research," have been discussed at earlier conferences, but the research and uses of these procedures continue to grow rapidly, so that additional discussion is valuable. The conference ended with a presentation of some preliminary papers and discussion on the methodological implications of the National Medical Care Expenditure Survey. This large study conducted for the National Center for Health Statistics and the National Center for Health Services Research is a rich source of methodological findings of which only a small sample was presented.

The conference proceedings also include the special discussions that occurred during the luncheon and dinner sessions. From a policy standpoint, "Implications of Survey Research for Health Policy and Programs" by Ruth Hanft and "The National Health Interview Survey—Recommendations by a Technical Consultant Panel" presented by Bernard Greenberg both precipitated vigorous discussions as well as some recommendations by the conferees. The discussion by Charles Jones of the U.S. Bureau of the Census on plans for data collection and processing should be of value to users of the 1980

census in indicating not only what the census does but why.

Finally, two sessions were devoted to a discussion of the contributions to health research of Leo G. Reeder and Elijah L. White. Both were among those who founded these conferences and both died in tragic accidents while this conference was being planned. This conference is dedicated to their memory.

It has always been necessary to limit the participation at these conferences to ensure that the group is small enough to permit and encourage active discussion and exchange of research findings and ideas. Every effort has been made, however, to make these proceedings as complete and accurate as possible so that readers who did not attend may share in the latest developments in this important and growing field of research.

Recommendations

The following recommendations emerged from the presentations and discussions at the individual sessions of the conference.

Surveys of physicians and health care providers

1. Reliability of information obtained from provider log-diaries and questionnaires may be improved by furnishing detailed instructions and definitions to respondents. Apparently even frequently utilized concepts such as "inpatient" versus "outpatient," "practice arrangement," and "specialty" are often not uniformly interpreted. However, long and explicit definitions may discourage respondents from thoroughly reading the instructions. Response rates may also be reduced by elaborate directions. These possible tradeoffs need to be systematically evaluated.
2. A purported advantage of the log-diary approach is that experiences are recorded soon after they happen. This minimizes recall problems and inaccurate reporting. If, however, the recording is done later, these potential benefits may be lost. The effects of provider characteristics and of the techniques used to elicit log-diary information on the time when the information is recorded and the completeness of that information require further investigation. In addition, a better understanding is needed of the effect that a delay in log completion has on the reliability and validity of the information provided.
3. It is plausible that physician characteristics such as speciality, type of practice, supporting facilities and personnel, and the kinds of patients seen will influence responses to requests for survey information. Additional efforts are needed to document such relationships. This work can lead to a better understanding of the differential approaches needed to obtain valid and reliable information from different kinds of practitioners.
4. Provider participation in social surveys may be influenced by (a) efforts to convince them of the importance of the study; (b) gains that they will derive from participation, such as compensation, feedback that might be useful for practice management, association and interaction with sponsoring professional, academic, and other esteemed organizations, and public service; and (c) encouragement of providers' staffs, who often assume major responsibility for completing the required forms. Additional studies should investigate the relative importance of these and other incentives to increase participation.
5. A major concern about provider surveys, as well as other social surveys, is the effort needed to maintain acceptable response rates and factors that might be associated with the increasing difficulties of obtaining provider cooperation. More should be done to establish what are the trends in cooperation on provider surveys. It is important in such studies to take into account the effort necessary to obtain a given response rate.
6. It is not clear what factors are critical in determining level of provider response. A number of factors are suggested as important: (a) auspices under which the study is conducted; (b) respondent burden in a particular survey; and (c) number of surveys to which a provider is asked to respond in a given time period. Efforts need to be undertaken to develop a general framework for considering various factors in order to establish their relative importance in determining the current pattern of provider cooperation.
7. It was suggested that a clearinghouse for physician surveys be organized. Such a clearinghouse might contain (a) a listing of organizations that have in the past conducted, or are currently conducting, physician surveys, with descriptions of those surveys; (b) published and unpublished reports of the findings of these studies; and (c) data from

these studies that might be used for secondary analyses. Such a clearinghouse would have the potential advantage of reducing physician burdens in filling out forms by offering an opportunity for coordinating the demands made on physicians by organizations and researchers. It would also reduce duplicate efforts and might thereby result in more cost-effective research. Finally, it has the potential to improve the quality of research by making it much easier to learn the current state of the art in the field.

Although appealing, this proposal would need to overcome a number of administrative and political barriers: (a) developing auspices under which the clearinghouse would be set up; (b) establishing a continuing financial and administrative base to support it; (c) convincing data-gathering organizations to participate and contribute to a data bank; and (d) ensuring the privacy of provider respondents. Despite obvious problems, the idea of a clearinghouse does seem worth pursuing so that more definitive recommendations could be made.

Telephone interviewing and respondent behavior

1. With respect to computer-assisted telephone interviewing, a number of technical issues need to be worked out. In particular, the most cost-effective mix between tasks needs to be developed. This mix should be assessed from the point of view of start-up and processing time, from the point of view of the effect of computer operations on the interviewer-respondent interaction, and from the point of view of the impact on interviewers' morale and satisfaction with their job. At this time, the potential of computer assistance cannot be adequately assessed on any of these criteria. Clearly, the whole area needs quite a bit more experience before it will be possible to make sound judgments about the best kinds of computer-assisted telephone interviewing systems, let alone make more detailed statements about the kinds of projects and applications that are or are not appropriate for such interviewing.
2. While some studies have found almost no differences between the aggregate figures based on telephone and personal interviewing procedures, others showed some tendency for sensitive data to be less well reported on the telephone. Although these differences are not very large, to understand why there are study-to-study variations further research is needed to produce better generalizations about when personal and telephone procedures are interchangeable and when they are not.
3. In particular, there is a need for better understanding of the interactive patterns on the telephone compared with those in personal interviews. It seems almost certain that the interaction is affected by the mode of communication. Understanding better how the interaction differs will open the way to designing more effective telephone procedures and training procedures for interviewers. Controlled laboratory experiments and tape recordings of actual interviews should be used as a tool for comparing face-to-face and telephone interviewing procedures.
4. Research is needed to identify what differences, if any, exist between "good" telephone interviewers and "good" personal interviewers. Some interviewers have a definite preference for one mode of interviewing over the other and have styles that seem particularly effective in one kind of interviewing or another. There has as yet been virtually no research on these matters; such research would be helpful.
5. Response rates are of particular concern in telephone interviewing. In particular, there is a difficulty with the random-digit-dialing strategies, for which no advance letter is possible. Telephone response rates are not always lower than those for personal interviews. There are certain groups, such as those with good security systems or urban populations in general, who respond more often to telephone interviews than to personal interviews. Nonetheless, research on effective ways of presenting telephone interviews is needed in order to improve further the value of telephone interviewing strategies. High response rates are particularly important in telephone samples because 7 percent of the population is excluded since they have no telephones.
6. In the preceding two conferences, there was discussion of the need to develop practical procedures for structuring interviewer behavior, making that behavior a positive force rather than a random or even negative force in the production of good respondent behavior. Results presented at this conference provide further evidence of the potential value of such efforts. Yet we still are very far from having a set of accepted practical guidelines for incorporating the results of experimental studies into standard interviewing procedures.

Use of informants and multiplicity estimates and the use of diaries

1. Multiplicity estimation and network sampling as techniques need further investigation regarding their applicability to health survey research.
2. Multiplicity estimation appears to be particularly useful for locating rare respondents, and its applicability in this regard should be further evaluated.
3. Information about the accuracy of reporting is another particularly critical issue, and further research into this technique must be carried out.
4. Another issue related to the accuracy of information obtained by these techniques is the counting rule used to define the network for sampling purposes. The tradeoffs between widely extending the network to reduce sampling error versus tightening the network to reduce response effects need further study.
5. Confidentiality and the implications of network sampling and multiplicity estimation for invasion of privacy are critical issues that will have to be addressed if this approach is widely used.
6. Diaries have been shown to be an effective procedure for obtaining health information about less obvious health events. Because health is such an important topic to respondents, cooperation rates with health diaries have been high, with little loss of respondents over time. We recommend that diary procedures be utilized in continuing studies of health behavior and expenditures such as the Health Interview Survey (HIS) and the National Medical Care Utilization and Expenditure Survey (NMCUES).
7. Further research on diary format and whether diaries should be used as a primary data collection procedure or as a memory aid is necessary.
8. With the increased use of diaries, more needs to be known about the conditioning effects of diary reporting. As a corollary, whether the diary is a viable means of public health education in the field of health behavior should also be explored and experiments using it in this way should be developed.
9. Research is needed on the cost of improving diary cooperation versus the impact of noncooperation on total survey error.
10. The issue of the quality of health data obtained from telephone versus face-to-face interviews remains unresolved. More research is needed on the relative costs, benefits, and alternative mixtures of these two procedures.
11. A careful investigation is needed to determine whether studies that combine telephone and personal interviewing procedures should use separate cadres of telephone and personal interviewers and under what conditions a single instrument is adaptable to both techniques.
12. We recommend further study of the interaction of effects due to procedure (telephone, face-to-face, diary), population characteristics (age, socioeconomic status, rural/urban residence), and design (cross section versus longitudinal). It seems likely that there will be optimal combinations of these three critical characteristics that can be recommended for future work.

Health data needs

1. Data for surveillance of environmental hazards are becoming an urgent need.
2. Longitudinal data on individuals are clearly essential. The induction period for many cancers may be 20–30 years. Other chronic, degenerative diseases may develop as a consequence of even low levels of exposure to a hazard over an extended time period. The capacity to identify individuals exposed to certain occupational or environmental factors during a certain time period and then to monitor health events in their lives would constitute a powerful tool in the search for health effects. It could serve also to measure the effects of health programs and other social change in the health status of the individual, another capacity we now lack. Such data would influence national health policies, as other data have done.
3. Small area data on infant mortality, measles, other disease concentrations, resources in the area, and use of services are needed to focus scarce resources on specific issues.
4. Better ways are needed to inform researchers and policy analysts about existing data and to assist them in using these data.

Mental health scales and health status measurement

1. Additional studies of reliability of diagnoses are needed, particularly in the area of mental health.

2. Mental health researchers should follow American Psychological Association guidelines for tests and presentations and provide more validation of methods.
3. In cross-cultural health research, more attention should be paid to concepts and less to literal translation.

The National Medical Care Expenditure Survey

1. Additional research is needed on the value in cost-benefit terms of providing respondents in health panel studies a Summary of previous information reported so that cost and other information may be completed or modified.
2. The process of reviewing and updating the Summary needs research specific to understanding the stimuli for updates that modify the originally reported data and the implications on the quality of the data.
3. The specific demographic groups for which the use of a Summary is more appropriate need to be researched and identified.
4. The cost of obtaining health provider data for validating household reports should be related to the value of the data in reducing total survey error.
5. Alternative measures of medical charges and insurance coverage require additional comparison and evaluation of both theoretical foundations and reliability of response.
6. More post-survey researching of interviewer perceptions is needed to add to our understanding of respondent limitations, instrument limitations, and limitations resulting from the interviewer-respondent interactive process.
7. It is very important to recognize inadequately researched methodological techniques that are being implemented for the first time in health surveys and to imbed controlled evaluations in the survey design.
8. Serious consideration should be given to budgeting for methodological research of large-scale health surveys proportional to the expected survey operational budgets.
9. Methodological research for large-scale health surveys should be concerned with identifying the significant sources of error and choosing those designs that are *optimum* in terms of total survey error.

The Health Interview Survey

The following recommendations of the Technical Consultant Panel on the national Health Interview Survey (HIS) were reported at the conference and received wide approval by conference participants. They are repeated here for the reader's convenience.

1. Core items on the HIS provide for continuous, long-term series of data; no item should enter that category unless time trends are absolutely essential. Any net increase in core means shortening or eliminating supplementary items because the current interview is considered full and any time increase for one section requires a compensating reduction elsewhere. The supplementary questions should be responsive to the needs of federal agencies and other entities, and routine solicitation of newly suggested items should be obtained every two or three years.
2. The Panel did not feel that the detailed data on chronic conditions justified the considerable costs in collecting and tabulating them. Instead, it recommended that the list of chronic conditions for which prevalence data are collected be shortened to eliminate those conditions for which useful prevalence data cannot be estimated. Furthermore, the members of the Panel felt that the coding of all chronic conditions to the four-digit code of the ICDA (8th revision) was not only a delaying factor but also an unnecessary expense, since such detail implies a level of accuracy and reliability that is not warranted.
3. Much of the criticism leveled at the way in which data on chronic conditions are collected and tabulated by NCHS was felt to be equally applicable to acute conditions. In particular, the Panel challenged the coding of all diseases to the four-digit code. The Panel also suggested that data on problems, symptoms, and complaints presented by a patient when consulting a physician might be tabulated.
4. There was doubt expressed about whether data on height and weight need be core questions or might be obtained periodically every few years. The original reason for including these questions as core was to identify groups of persons with problems of obesity and to relate them to hypertension. Using these data on height and weight for detailed epidemiological studies might be challenged because of their unreliability as well as the accuracy of the health condition itself. The HIS was urged to assess the utility of these items on height and weight in order to justify their retention as core.
5. To replace the deleted items, the Panel recommended the inclusion of some questions on mental health, occupational and environmental health, and health insurance

and expenditures. NCHS was urged to consult with other federal agencies for specific suggestions on how their data needs might easily be met in these areas. In reviewing the requests for supplementary items, the Panel urged the use of an external group of advisors to review the requests. There were other related functions that such an advisory committee or panel might fulfill.

6. Regarding the sociodemographic variables included in core, the Panel urged that attention be given to achieving some degree of uniformity so that all federal programs, or at least NCHS, would use the same categories to collect these data. The present system has too many differing definitions of terms, and even data on age are inconsistently collected within NCHS. Expanding the items to include religious preference and disaggregating the data for those 65 years of age and older were also recommended.
7. The Panel believes that the HIS has not yet made a systematic attempt to evaluate significant alternatives to the present design of the survey such as the use of bounded interviews and partial rotation of the sample. The Panel recommended that NCHS undertake a careful and comprehensive evaluation, based on Total Survey Design (TSD) principles, to determine the feasibility and desirability of several alternatives to existing HIS data collection and sample design features. Design changes resulting from this evaluation should be timed to coincide with the redesign of the sample following the 1980 Census of Population and Housing.
8. The Panel also recommended that TSD principles be followed in providing technical assistance to help state and local governments meet their health data needs. Moreover, low-cost survey methods such as computer-assisted telephone interviewing with random-digit dialing should be given full consideration as alternatives to replication of the procedures used in the national HIS.
9. Increasing demand for wider coverage of subject matter and greater detail for analysis means that the HIS must be strengthened by increasing its capacity through a more efficient sample design. This will facilitate more detailed and reliable data for different demographic, socioeconomic, and geographical sectors of the population on a more timely basis.
10. Considering all the elements involved, including interviewer training and bias, the Panel felt that continuation of a continuous survey was preferable to periodic surveys.
11. The HIS should start immediately to make an intensive review of its own design features (e.g., longitudinal versus cross-sectional approach, respondent rules, length of reference period, use of diaries, interview methodology, and other features) to take advantage of the opportunities provided by the post-1980 Census Bureau design. Moreover, in order not to defer the study, this review should be undertaken by staff who carry no responsibility for ongoing HIS operations. If such staff are not available, the Panel recommended use of an appropriate contractual arrangement.
12. Although a tremendous wealth of survey knowledge has come from the HIS, the Panel believes the opportunities for whole new approaches in data collection and analyses are not being fully exploited. There are new data requirements, new techniques of interviewing and data collection, new processing technologies, and whole new concepts of error structure that require research by HIS staff. The Panel therefore recommended that not less than 10 percent of the HIS budget be allocated to methodological and developmental research.
13. The HIS is only one of a large number of data sets compiled by the federal government. Analysis of HIS data in isolation from all of the other data sets deprives the nation of valuable linkages. Examples of typical linkages are provided in the full report, and these illustrations reflect possible usefulness of cross-analyses for policy issues as well as etiological issues in epidemiology. NCHS was charged with establishing a procedure for cross-analysis of HIS with other data sets.
14. More publicity should be given to the fact that the HIS can provide raw data and tabulations on data tapes to people who need such information and can reimburse NCHS.
15. Quick estimates of some items could be obtained with two weeks of interviewing. The current lag seems to be due to the lead time needed to develop and incorporate questions. More attention should be devoted to ways of incorporating special purpose items in the HIS without too much delay.
16. The ongoing analyses are less than desirable because of lack of staff availability. In this respect, highest priority should be given to increasing staff positions to strengthen the analytic capability. The HIS should explore ways of offering special incentives to staff

members for producing and disseminating analyses of HIS data. Also, expanded use of the IPA mechanism should be sought to recruit talented persons to devote themselves to analysis of HIS data. Finally, when responding to requests for special data collections from other federal agencies, part of the joint commitment should include staffing for the analysis of the collected data.

17. The HIS should continue to study users of their data and to evaluate the usefulness of given data sets.
18. The demand for subnational data has increased exponentially during the past decade because of federal promotion of health planning and regulation at these levels. These changes, plus the inflationary effects of increases in all health expenditures, have made it essential to study the distribution of health services as well as their impact. Thus, there are state and county needs as well as the needs of Health Systems Agencies and Professional Standards Review Organizations.

    Other federal legislation, such as support for Health Maintenance Organizations and the placement of professional manpower in the National Health Service Corps, has also added to the need for local data. The underserved and shortage areas have tremendous stakes in having such data available on a timely basis.

    Health interview data are unique in that they do not arise through routine administrative procedures such as licensing of professionals or reimbursement of health expenses. Such interview data are useful because they provide information on persons whether or not they use the health system.
19. Some modifications of current procedures could be adopted, or should be explored, to strengthen the subnational data program after 1980. One simple method would be to base subnational data on 104 weeks of interviewing; there would thus be biennial publications for most of the global statistics. Another possibility would be to divide sample resources into two parts. Part A would be the continuation of the present system and would provide global statistics for the nation and for larger states and metropolitan areas. Part B would distribute resources differently, reduce the number of items queried, and experiment with lower-cost collection techniques such as mail and telephone.

    Other solutions involve greater use of synthetic and composite estimates for the subnational areas. The HIS has used synthetic estimates, and the Panel recommended greater use of them as well as the Composite Estimator in which a synthetic estimate is one component.

    The growing demand for subnational data will probably not be fulfilled by any of the foregoing procedures, including the strengthening of the design of the survey, or by increasing the sample size within reasonable bounds.
20. A more suitable and practical method of meeting subnational data needs is for NCHS to provide a complete range of technical assistance and demonstration activities on health interview survey methods and the uses of such data. This technical assistance should be provided at varying times and at minimal cost to those agencies and organizations in need of such assistance. This program of technical assistance should take advantage of relatively low-cost data collection techniques when they can be expected to produce data with adequate reliability for local data needs. Moreover, the HIS should develop guidelines for the local staffing, funding, and uses of the data in these local areas. Contractual support and IPA personnel may be used for this technical assistance, but the primary source of this service should be the professional staff of NCHS, whose numbers and resources will probably need to be increased to accommodate this demand.

SPECIAL SESSIONS

In Honor of the Memory of Elijah L. White

Chair: Seymour Sudman, Department of Business Administration and Survey Research Laboratory, University of Illinois at Urbana-Champaign

---

Memorial Session in Honor of Leo G. Reeder

Chair: Bernard G. Greenberg, School of Public Health, University of North Carolina at Chapel Hill

Recorder: Linda S. McCleary, National Center for Health Services Research

---

Implications of Survey Research for Health Policy and Programs

Chair: Gerald Rosenthal, National Center for Health Services Research

Recorder: Matilda Frankel, Survey Research Laboratory, University of Illinois at Urbana-Champaign

---

Selected Methodological Features of the 1980 Census

Chair: Charles F. Cannell, Survey Research Center, University of Michigan

Recorder: Sherman R. Williams, National Center for Health Services Research

---

In honor of the memory of Elijah L. White

Robert A. Israel, National Center for Health Statistics

Charles F. Cannell, Survey Research Center, University of Michigan

Robert R. Fuchsberg, National Center for Health Statistics

ROBERT ISRAEL:

On behalf of Dorothy Rice and the entire staff of the National Center for Health Statistics, I take this opportunity to tell you how much pleasure we have in cosponsoring, with the National Center for Health Services Research, this Third Biennial Conference on Health Survey Research Methods.

I also want to thank you for your participation, which is so essential if this Third Biennial Conference is to be a success.

Mrs. Rice sends her regrets that she cannot be present with us at this time, but she is in Geneva as a member of the United States delegation to the World Health Assembly.

Conference mechanisms such as this bring us in close contact with innovations in survey research methodology, so that we can apply the best possible techniques to our data collection and analytical activities, as well as stimulate our own research activities. Tremendous strides have been made over the last ten years in the development of improved measurement instruments and new approaches to data gathering. Although health interview data have been widely used for many years, the increase in validity and reliability of measures of health and disability have broadened their utility as input to health policy and decision making.

Elijah White, whose recent untimely passing is keenly felt by all of us, has been a significant contributor to and supporter of methodological research as well as practical applications of the fruits of such research. Elijah and his wife, Mary, as well as our colleague Feliks Sawicki, were killed in Poland in April while engaged in a survey research project. Elijah came to the Public Health Service as Assistant Chief of the Health Examination Survey in 1957 from Community Studies, Inc. in Kansas City, Missouri, where he had served as field director and survey statistician. In 1960, he became the Deputy Director of the Division of Health Interview Statistics; in 1966, the Director of that Division; and in 1973, the Associate Director for Data Systems, a position that encompassed all of the data collection mechanisms of the National Center for Health Statistics. In addition to his professional positions, he has contributed significantly to our profession through his interest in young statisticians and their training and development. Also, he served as a professional recruiter for the National Center for Health Statistics, and he has left his mark on at least one member of the NCHS staff, today being my 13th anniversary with NCHS, having been recruited to the Center by Elijah White.

CHARLES CANNELL:

It is my assignment and my privilege to comment on Elijah's role in survey methodology. As I considered the methodological work of the National Center for Health Statistics, it became clear that to isolate a particular activity and say "This is the work of Person A or Person B" is not possible. It is not possible because the activity is the product of more than one person within the Center working cooperatively with persons from other organizations.

Students of organizational theory tell me that this pattern identifies a "person-centered" organizational form. This is a particularly effective style that generates cooperation and utilizes the abilities and skills of its members, to the benefit of both the individual and the organization. This characterizes Elijah's operational style and also indicates why discussing his methodological contributions cannot and should not be isolated from those of others on the NCHS staff.

By way of apology for slighting or omitting the work of others, these comments relate principally to the methodological activities of Elijah and his staff and of me and my colleagues at the Survey Research Center. These are, then, per-

sonal reflections, not an exhaustive cataloging of activities.

Because of his position in NCHS, Elijah had a unique opportunity to initiate, guide, support, and critically evaluate health survey methods. It is as a research administrator that he made particularly significant contributions to studies of survey methods—especially those relating to response effects. He and his colleagues promoted good survey methodology and influenced research to improve methods through several activities. First, he was involved in promoting, guiding, and financing research studies designed to improve the quality of Health Interview Survey (HIS) data, such as the following:

- Validity of reporting of hospitalizations and visits to physicians (which we did at SRC)
- Madow's work on response validity in reporting chronic conditions
- Observation and coding of interview behavior, a project for which Jack Fowler and Kent Marquis had major responsibility
- Quality control and measurement of nonsampling error in the Health Interview Survey (David Koons at the Bureau of the Census)
- Effects of an extensive questionnaire and diary procedure (Laurent and Marquis)
- Statistical methods for analysis of HIS data (Temple University)

Some of these studies developed when NCHS recognized a topic that needed to be explored and then located someone to do the research. It was quite likely, however, that an outside researcher identified a topic that needed attention and found a receptive hearing with Elijah and his staff. The working relationship between the contracting parties characteristically developed as a collegial relationship among fellow researchers.

Elijah's second influence on methodology was in his concern over the quality of surveys being conducted in regional, state, and local health organizations. He devoted considerable time to consulting with health agencies, helping them toward sound methods. I recall that the St. Louis Public Health Conference on Records and Statistics included a session that Elijah organized on survey methods. Following the session, members from several state delegations kept Elijah busy well into the dinner time and took up much of his time for the next day, seeking his advice and assistance for their particular health survey needs.

Third is Elijah's activity in fostering and supporting groups dedicated to promoting sound methods. I note here only two. The first is an organization that many of you may not have heard of, the Field Directors' Organization. This group of field directors of government and academic survey organizations meets once a year to talk over operational and methodological issues of field procedures. The first of these meetings was held in Ann Arbor some 15 years ago. Elijah asked if he could participate even though technically he was not eligible. We brashly assigned him the job of organizing a discussion on "methods of improving the coverage, response rate, and quality of response in central cities." I'd like to be able to report a real breakthrough on this topic. Unfortunately, this was one task in which Elijah—along with the rest of survey research—failed.

Elijah has also been a member of the planning committee for these Biennial Conferences on Health Survey Research Methods. From the beginning, he has been an active participant in the sessions.

It was through working together on research contracts that Elijah and I first became acquainted. In earlier days, arranging contracts was considerably simpler than at present. Agreeing on a topic to be investigated, hypotheses to be tested, and research designs to be used were all arranged in a couple of meetings between Elijah and his colleagues and those of us from SRC. These sessions were a give-and-take among researchers, sometimes with sharp differences of opinion, but with mutual respect for the opinions and expertise of the others. From these interactions Elijah and I formed a close friendship.

Elijah's major contribution was as a research administrator who was technically trained, highly qualified, and experienced in survey research. His dedication to promoting sound methodology and his efforts to extend knowledge of survey techniques through methodological research have worked toward the goal of improving the quality of health studies and of survey research in general.

It is both as a friend and as a research colleague that I cherish his memory.

ROBERT FUCHSBERG:

I was asked to speak about the contributions of Elijah L. White because I worked very closely with him for over 20 years. I first came to know Elijah when he joined the staff of the National Health Survey in 1957. His first job in this program, which was the forerunner of the National Center for Health Statistics, was often described as a challenge. I would describe it as an impossible task that was carried out by Elijah and his

colleagues despite overwhelming obstacles. They designed, developed, and tested procedures for the Health Examination Survey. He had the principal administrative and statistical responsibility for the development and implementation of a nationwide survey to conduct health examinations on a sample of the population. In carrying out this major assignment, he coordinated efforts required to develop a national sampling plan. He initiated a research program for developing and evaluating the examination content and techniques for the standardized clinical examination. He negotiated with state and local health authorities to facilitate cooperation in the administration of the program.

He was responsible for directing several large-scale pilot studies to test methods of operation and quality-control procedures. Elijah worked with contractors to develop methods and techniques designed to ensure a high level of cooperation from the sample persons selected for the survey. But the task that sapped his energy and unique diplomatic skills called for him to recruit and direct a staff to meet the logistical needs of supplying examining units and an examination team that were deployed over the United States at each sample location. He not only coordinated these data-gathering activities but also developed a program of analysis and interpretation of the data derived from the survey and related research.

In 1960, he shifted his activities within the National Center for Health Statistics to the Health Interview Survey (HIS). In his positions as Assistant Chief and Director of the Health Interview Survey, he continued to strive for excellence in each aspect of survey research. His ability to motivate colleagues to overcome impediments provided the leadership that was required to make the HIS one of the most respected data-gathering systems in the country. He was proud that the data from this survey were used both by persons who advocated and by those who opposed key health care programs such as Medicare and Medicaid when Congress was debating the legislation to establish these programs.

Elijah was always on guard to ward off any threat that the National Center for Health Statistics might be politicized and lose its reputation as an independent fact-finding agency. In addition, he constantly strove to improve HIS procedures by encouraging methodological research on every aspect of survey design and analytical techniques. Elijah demanded that we broaden our horizons and encouraged the development of new initiatives such as the National Medical Care Expenditure Survey, the National Medical Care Utilization and Expenditure Survey, the Telephone Health Interview Survey, and the Survey Intelligence Service.

With the reorganization of the National Center for Health Statistics, Elijah became Associate Director for Data Systems. In this position, he was responsible for encouraging and coordinating all data-gathering activities of the Center. In this capacity, he applied his firm principles to all of the Center's data-gathering activities so that his colleagues could be proud of their organization and the data users could be assured of valid and reliable health statistics.

Memorial session in honor of Leo G. Reeder

Bernard G. Greenberg, School of Public Health, University of North Carolina at Chapel Hill

Howard E. Freeman, Institute for Social Science Research, University of California at Los Angeles

Alfred C. Marcus, School of Public Health, University of California at Los Angeles

Sharon Reeder, School of Nursing, University of California at Los Angeles

BERNARD GREENBERG:

I am pleased to welcome and thank you for participating in this session to honor the memory of Dr. Leo G. Reeder.

It might be well to tell you how this session will be structured and why it has been associated with the evening meal. When the Planning Committee for this Third Biennial Conference met in San Diego, California, on August 15, 1978, in conjunction with the annual meeting of the American Statistical Association, the decision was made that the initial evening session would be devoted to a summary by me of the final report of the Technical Consultant Panel on the national Health Interview Survey. Leo Reeder and I were members of that Technical Consultant Panel, and what a tragedy it was to learn only about one month later that he was a passenger on the commercial plane in the air disaster over the San Diego airport.

The members of the Technical Consultant Panel decided soon thereafter that its report, which was in the final stages of preparation, should be dedicated to the memory of Dr. Reeder because of his key and stimulating role on that Panel. Inasmuch as the Planning Committee for this conference also wished to honor Dr. Reeder, we voted to have this memorial session in his honor and to combine it with the presentation of the Technical Consultant Panel's report that had already been dedicated to him.

Before presenting a summary of that report, two of Dr. Reeder's colleagues from the University of California at Los Angeles have been invited to comment on Leo as an individual and on the relative importance of his contributions. The first speaker is Dr. Howard E. Freeman, a distinguished sociologist who was recruited from Brandeis University and the Social Science Research Council several years ago by Leo Reeder to replace him as Director of the Survey Research Center at UCLA.

HOWARD FREEMAN:

The suddenness of Leo Reeder's death brought home to me our dependence on ritual. Although the evidence was clear that Leo was on the fated PSA plane, failure to identify his body—first hours and then days after the crash—left his family, friends, colleagues, and students in a state of persistent despair. A ceremony was desperately needed to initiate the process of adjusting to his death. With wisdom and courage, his wife, Sharon, arranged a memorial service on the Saturday—some five days—after Leo's death. The event was sad, and the grief was evident among the several hundred people who attended.

I must confess, however, that I sat through the service with a smile on my face, and I had difficulty refraining from laughing. Why? One of Leo's retired faculty colleagues was serving as an usher. When I walked through the door, he said, "Leo wants you to sit in the front row." I smiled because I literally could see a big grin on Leo's face, feel a poke in the ribs, and hear Leo say, "How does he know?"

Well, I know that Leo would feel extremely honored that we are here tonight. He would feel it a personal testimony to our friendship that I stopped after two martinis before dinner so that I would not mumble too badly. He would be deeply touched that so many colleagues from near and far are here tonight. This group meant a lot to Leo, in part because he helped found this series of meetings and served in a leadership role in them, and in part because of the many close friends who attended them, but mostly because these meetings represent what Leo most looked forward to, opportunities to share ideas and findings with persons from different disciplines and outlooks.

Since this is a forum of Leo's friends, he would want me to be brief. But there is a lot to say.

First, he wants it known that he really did

burn the midnight oil. He always thought he deserved the recognition that he received within his university and the honors that he accrued on the outside.

Second, he still insists that stress is related to heart disease, although he is not sure about its relation to cancer.

Third, he strongly advocates the continued promotion of the consistent use of the same indicators and scales in our studies. He argues, of course, for a careful consideration of the ones that he had a part in developing, and he reminds us of the book of measures he coedited several years ago.

Fourth, he is not angry with editors and referees who rejected his papers. Over the years, he had almost all of them published in one of our major journals. Even his early papers, now over 20 years old, are found on current reference lists.

Fifth, he is very grateful for the various federal groups and foundations that supported his research. Nevertheless, he does want it known that he could have done his work faster with a little more money, and sometimes with fewer site visits.

Sixth, he wishes that his wife, his student Al Marcus, and the rest of his research team would talk less, write more, and get on with finishing his preventive health behavior study. He continues to be optimistic about its substantive and methodological importance.

Seventh, Leo hopes his graduate students will double their research efforts and finish their dissertations more quickly. With respect to the ones whom I inherited, they should understand that most of us do not have Leo's high standards, and they can get by with much less now.

Eighth, he is violently mad over the recent episode of poor administration at UCLA. He thinks not having an up-to-date faculty list is inexcusable, but he is proud of the way Sharon handled it. What happened was that the committee to raise money for Leo's memorial fund at UCLA, of which I am a member, sent a solicitation letter to our university colleagues. Back to us came a short, calm letter from Sharon, commending us for our efforts but pointing out that it is rather unusual to ask the wife to contribute. She also wrote that it was even more unusual that she received a second letter addressed to Leo.

Ninth, Leo knows there is a lot more I could say. However, he realizes that I have written a more formal and dignified obituary for the American Sociological Association's *Footnotes* and that I have been asked to comment about him at other meetings and have to save some material for them.

Finally, Leo wants to be loved and missed but not mourned. He really was an unpretentious person who would have wanted to smile and be pleasant at his own memorial service and would have wished this meeting to be a happy, as well as productive, one.

BERNARD GREENBERG:

Thank you, Howie. Obviously, the Planning Committee must have had some ESP itself in selecting you as the lead-off person, since you appear to be spiritually in communication with Leo.

One of the most notable achievements of Leo Reeder was the close working relationships and empathy he had for graduate students. He not only was a superb teacher himself but a friend and adviser to any student who indicated a willingness to work. Therefore, the most fitting person to comment on Dr. Reeder's achievements in this realm is one of his own doctoral students. Working with Leo on one of his projects was Alfred C. Marcus, whom I should like to call on to provide us with the perspective of how a student perceived Dr. Reeder.

AL MARCUS:

As most of you know through your own experiences with Leo, he was a warm, friendly, and gregarious man who had boundless energy and a real passion for life. Some of my fondest memories of him include the times we played volleyball and tennis and attended the Los Angeles Dodgers baseball games together. He especially enjoyed hiking with his son Andy. Leo also loved to party, and he reveled in the camaraderie of professional meetings such as this.

I recall, for example, attending a technical session of the meetings of the American Statistical Association in San Diego to which Dr. Greenberg referred. The session was entitled "Improving Response Validity," and Leo had arranged the program and was serving as chairperson. I have a vivid memory of all the distinguished participants sitting on that raised platform in front of the audience. During the reading of one of the papers, Leo took a piece of paper and wrote something down. As he did so, a smile crossed his face and he passed the note to his close friend sitting next to him. It was Charles Cannell; after accepting the note, he read it quickly and gave a big smile to Leo in return. Then, he too scribbled a note on the paper and returned it to Leo. I don't know the contents of that message, but whatever it was, it made Leo smile noticeably but not enough to disturb the speaker. Sitting there in the audi-

ence and watching these two grown men, full professors in their universities, passing notes back and forth, I began to appreciate how much Leo really enjoyed himself at scientific meetings. He was fun to be with no matter how serious the occasion.

I remember another incident from that same conference. Leo, Larry Jordan, and I were sitting beside the pool in our swimming trunks on one sunny afternoon. The meetings were still in session, so it was obvious that we were "playing hooky" from the afternoon schedule. Soon, we were joined by another truant, Dan Horvitz, who was also wearing swimming trunks. As Larry and I listened intently, the two of them passed the hour exchanging stories and anecdotes and playing a friendly game of one-upmanship. Leo loved dialogues like that, especially when the other person was good at it.

I feel certain that everybody here who knew Leo well could tell his or her own favorite story about him, probably in connection with a professional meeting or conference attended together. However, there was another side of Leo that you may not know, and that involves his role as a professor and teacher at UCLA. To prepare myself for this memorial, I met with several of my colleagues to reminisce about Leo. Everybody agreed that he was a very accomplished lecturer and had an excellent reputation for his class preparations. Several students glowed in recalling some of his more inspiring lectures, especially those concerning the relationship between stress and coronary heart disease.

In quite a different vein, one student told of the time that she took the qualifying examinations for the doctoral program. Being terribly anxious and nervous, she simply could not complete some of the questions and thus failed. Leo was on sabbatical leave at the time, but he was the only faculty member in the Division who called to comfort and reassure her. At the time, she was thinking of leaving the University. Leo convinced her not to do so in reaction to the test and helped her to reschedule the examination; she later passed with high marks. The amazing part of this tale is that Leo was not even her faculty advisor and was not that familiar with her background. You can readily imagine how she respects Leo as a person and professor.

When this story was told in the presence of other students, the remarkable effect it had was that another student told a similar story. He too was persuaded by Leo not to leave school after failing an examination and was grateful for this little extra that Leo always provided. These incidents are part of Leo's legacy at UCLA that are probably not known by many of his colleagues outside the University.

I think that my favorite story about Leo was the time when an impostor in Los Angeles began calling people on the telephone and posing as Dr. Leo G. Reeder, Director of the Survey Research Center at UCLA. After making proper introductory remarks, this person would then ask questions about the most intimate details of the interviewee's sex life—all in the presumed name of science. You can imagine how embarrassing this was for Leo and the Survey Research Center when Leo started to receive complaints from irate housewives and their husbands. However, after a while the impostor was apprehended by the police. In later years, Leo loved to tell about this incident, and with a twinkle in his eye, he would say that he was going to publish the transcripts of those interviews—under the name of Dr. Howard E. Freeman.

I also have a clear recollection of the time Leo returned from one of his usual trips to Washington. He had attended a working conference as a member of the Technical Consultant Panel on the Health Interview Survey. Upon arriving at work the next day, he called me into his office and, with great enthusiasm in his voice, repeated the details of the meeting. Leo felt that he was being of service to the nation and was proud of what the Panel was trying to accomplish for health surveys. I think it is most fitting that this particular memorial session has been dedicated to Leo's memory and that the first public hearing of the Panel's report be given tonight. He took enormous pride in feeling that he had a role in improving the national Health Interview Survey.

It is too bad that we cannot have enough time to allow everybody here tonight to share his favorite anecdote about Dr. Reeder. I feel privileged to have been invited to share the perspective of a student and even more so to have been one of Leo's students. Dr. Freeman struck a responsive chord in me when he described this gathering as a forum of Leo's friends. That is exactly what it is, and I want to thank you for the opportunity of this privilege.

BERNARD GREENBERG:

Thank you very much, Al, and I know that Leo would echo my feelings in wishing you well and an early success in completing your dissertation. Before calling on Sharon Reeder tonight, I would feel left out if I did not tell one personal anecdote, such as Al Marcus suggested, that demonstrated Leo's warmth and humanism.

During the meetings of this Technical Consultant Panel, my wife and I introduced Leo to the Kennedy Center and the fine theatre that helped to neutralize the stress of the day's work. We always tried to include Leo whenever we could get tickets for a performance. To show his appreciation, Leo informed me around four or five months before the statistical meetings that there was a fine Shakespearian festival in the evening at one of the parks in San Diego. With the newspaper clipping that he had enclosed, he indicated which performances he, Sharon, and their son Andy would attend and suggested that if I obtained tickets for any of them, I could be assured of transportation in his car to and from the park. This was most fortunate because later, when the American Statistical Association started to notify its members about the performance, tickets were hard to obtain. That evening was the last one that I spent with Leo and his family. It serves to remind us what a kind, thoughtful person he was to think of me four months before the meeting.

As indicated, we have asked Mrs. Reeder to share this evening with us. Unfortunately, their son Andrew could not join us because he is in school back home. Sharon is on the faculty at the UCLA School of Nursing, and she was a colleague of Leo's as well as his wife. They published several papers jointly, and each was the severest critic of the other's research.

Sharon is not unknown to many persons here tonight. She and Andy accompanied Leo to the Second Biennial Conference at Williamsburg, Virginia, in 1977. Leo was a family person who always tried to have his family with him. In part, Leo's love of his family was why he was on the ill-fated plane that day. He was scheduled to participate in a morning conference on cancer in San Diego, and he took the early commuter plane in order to spend an extra evening at home with his family. Others had gone to San Diego the evening before.

Sharon, would you like to say a few words at this time?

SHARON REEDER:

Howard has done very well tonight, martinis notwithstanding. I believe that brevity is the soul of wit, and so I intend to be brief.

There are special memories for me that relate to this conference. Some have to do with the people involved, others have to do with the substance of the conference and what it has come to stand for. With respect to the "first mention," several of you have seen the Reeders through some difficult times. Throughout those years, and finally, when Leo was taken so abruptly, you have given time and expressed your concern. I want each of you to know that your efforts have been truly appreciated.

For the "second mention," namely, this conference, I'd like to tell you what Leo and I both feel. You see, I too have a special connection like Howard.

In one of Leo's obituaries, concern was expressed that life being what it is, Leo and his work might be swept out of our memories too soon, the ocean of time obliterating the accomplishments of Leo the man. I don't believe this will happen because Leo knew what was important. While he was less concerned with self-aggrandizement than with getting on with matters that have substance and importance, he was also smart enough to structure things so that they would not fizzle out the minute his back was turned. To be fair, he also did not want to have himself associated with what he called the "dummies." He wanted to win, and he did love a good pat on the back for a job well done. So it was with this conference. The Biennial Conference was one of his fondest projects—and he had many. As the idea began to germinate, he was excited. He began to assemble the cast in his head. He had a vision of a true interdisciplinary group who could come together periodically to share their work, ideas, and progress. Leo knew the value of different outlooks to provide a healthy leavening for any discipline. He knew how to cooperate with others to get a good thing going, and he knew how to pass the ball to let competent people carry it along. He did not necessarily want to be in the driver's seat, he just wanted to be sure that what he believed important would continue. He did want me to let Seymour Sudman know, however, that he's disappointed that we are not enjoying the sun out in New Mexico. He really did like the Bishop's Lodge!

Ralph Turner, a good friend and colleague of ours, has stated Leo's philosophy and style very aptly. Leo, he said, was like a fireman going to wherever the fire was. He was persistent, dogged, and concerned. In the case of the Biennial Conference, it was not a first priority that he be the fire chief so long as the fire got attended to properly.

I think that all of us here know that things will be taken care of and that these conferences will continue. The rest of the people on the Biennial letterhead and you, his friends and colleagues, will see to that.

In closing, as Howard has said, let us not mourn him or be sad. The best tribute of all to celebrate his memory would be a happy productive conference this week and many more of the same in the future.

BERNARD GREENBERG:

Sharon, thank you ever so much. Please convey back to Leo that we got his message, and the Planning Committee does not want him to think of us as "dummies." In fact, the Planning Committee wanted to record its deep love for and appreciation of Leo and has prepared a special testimonial certificate. If you will join me up here at the lectern, I would like to read and present it to you.

In beautiful style reminiscent of the medieval illuminated manuscripts, the citation reads as follows:

Biennial Conferences on Health Survey
Research Methods
Certificate of Appreciation
is awarded to
Leo Glenn Reeder

as a recognition of his dedicated service and outstanding contributions in planning, organizing, administering, and editing of publications for these biennial national conferences on health survey research methods sponsored by the National Center for Health Statistics, National Center for Health Services Research, Department of Health, Education, and Welfare, and by other participating agencies and universities.

It is dated May 16, 1979, and has been signed by all members of the Planning Committee.

Sharon, we want you and Andy to know that our love and prayers go with this testimonial, and we hope that God's blessings will shine upon you both in the future.

The national Health Interview Survey—recommendations by a Technical Consultant Panel*

Bernard G. Greenberg, University of North Carolina at Chapel Hill (Chairperson)

Members of the Panel:

Carol W. Buck, University of Western Ontario

Rodney Coe, St. Louis University School of Medicine

Floyd J. Fowler, Jr., University of Massachusetts

Robert R. Fuchsberg, National Center for Health Statistics

Thomas Jabine, Social Security Administration

Roger Kropf, Alpha Center for Health Planning

Leo G. Reeder, University of California at Los Angeles (deceased)

Anne A. Scitovsky, Palo Alto Medical Research Foundation

Walt Simmons, Alexandria, Virginia

John F. Wennberg, Dartmouth Medical School

Introduction

The Health Interview Survey (HIS) was established in July 1957 and is the world's oldest, continuous survey that collects data on health status, illness, injuries, disability measures, and related information involving the social, economic, and demographic attributes of the population in the United States. It was created as a result of the National Health Survey Act in July 1956 and is the responsibility of the National Center for Health Statistics. The objectives of the survey have been modified periodically as a result of subsequent legislation, some as recently as 1974 and 1978.

The HIS is a household interview survey based on a probability sample of approximately 800 households per week, with data collected on about 40,000 households and 116,000 persons annually. The recall period for most items is the two calendar weeks prior to the interview, thereby providing a system for a continuous overlap of one calendar week. Originally, the survey studied only the most urgent needs of health planners and provided data on illness, disability, hospitalization, and doctor visits. As these initial data needs were met, the HIS added more topics and special-purpose items in order to probe health issues in more detail. The sampling design has had three modifications (1959, 1963, and 1973) after the original sample in 1957. Each modification was based on data from the previous decennial census. The 1963 change increased efficiency by permitting better estimates of components of variance. There was a major evaluation of the HIS in 1963–67; the survey instrument was redesigned in 1967–68 to improve validity of the estimates, and new procedures were instituted to improve quality-control techniques. (For details of these changes, see U.S. National Center for Health Statistics, 1975.)

The current format of the questionnaire consists of a set of "core" items that are repeated every year and not modified for at least ten years. These core questions require about 30 minutes of the interview and gather data on the incidence of acute illnesses, prevalence of chronic conditions, days of disability, physician and dental visits, hospitalization, and a set of sociodemographic variables including age, race, sex, marital status, education, employment, and income. The remainder of the interview, about

* This paper represents an overview of the Panel's report to the National Committee on Vital and Health Statistics. The complete report is dedicated to the memory of Leo G. Reeder, a member of the Panel who died in an airplane crash in September 1978. Prior to his death, Dr. Reeder had played a key and stimulating role in discussions and recommendations of the Panel.

15–20 minutes, is devoted to supplementary items of current interest, such as vaccination status, smoking habits, use of hearing aids and dentures, and health insurance coverage. These supplementary items are used for only one to three years, and the duration may even be as short as one-quarter of a year.

Data from the HIS are published in *Advanced Data Reports* for rapid access to limited amounts of data and in the Series 10 reports of NCHS's so-called "rainbow series" of publications, *Vital and Health Statistics*. Sometimes there are special reports outside of Series 10, such as those on methodological issues and important topical subjects. Data are also made available on tape and through special tabulations.

Creation of Technical Consultant Panel

The National Committee on Vital and Health Statistics established this ad hoc Panel as a subcommittee in 1977 in order to review the status of the HIS and to ascertain how well it was meeting the legislative mandate. The members of the Panel are listed above and are considered coauthors of this summary report.

The charge to the Panel was as follows:

1. To recommend, in order of priority, the topics for which the type of general purpose data produced by the Survey are most needed on a national, regional, or state basis.
2. To review relevance of the core questions that are asked on an ongoing basis in terms of the changing emphasis or direction of national public health policy.
3. To recommend on an annual basis whether the core component of the questionnaire should be expanded or contracted in relation to the amount of interviewing time to be left for supplemental topics.
4. To consider and make recommendations regarding any biases resulting from the sampling methods, the survey collection techniques, or the data processing and analytical procedures that may produce inaccurate or invalid estimates for any numerically significant group.
5. To review and make recommendations regarding the types of sociodemographic data collected on each person with regard to both the variables to be included and the type of information to be sought on each variable.
6. To make recommendations regarding the most appropriate form for disseminating the data—with special emphasis on the question of the tradeoff between the amount of time devoted to analysis and the timeliness of the release of the data.
7. To make recommendations regarding the optimum sample size and the frequency with which this survey should be conducted.

Without going into the details of the legislative mandate by Congress in 1974 (PL 93–353), the National Center for Health Statistics was charged with the responsibility of collecting data on "comprehensive health statistics." This included the extent and nature of illness and disability, impact of such impairment on the economy, environmental and social health hazards, health resources, utilization of health care facilities, health care costs and financing, and patterns of family growth. The law also authorized NCHS to undertake and support methodological research on new or improved methods of obtaining current health data.

In 1978, the mandate was broadened (PL 95–623) to add epidemiological activities for the purpose of improving the effectiveness, efficiency, and quality of health services and health statistics. Singled out for special emphasis were the special health problems of low-income and minority groups and the elderly. Most significantly, NCHS and NCHSR were also authorized to undertake and support training activities.

Another new element in the law was that NCHS may collect, tabulate, and analyze data on health and health care on the request of public health and nonprofit entities, with the requester paying the cost of the service provided. NCHS was also authorized to provide technical assistance to such entities in the effective use of health statistics compiled by the Center. These two sections are pertinent to one of the Panel's charges concerning the issue of providing subnational data. These data encompass information not only for individual states but also for cities and counties and clusters of the latter coinciding with Health Systems Agencies and other health areas. Discussion by the Panel on how to fulfill this charge consumed a major portion of the time involved.

The 1978 law also detailed the kinds of data to be collected in order to highlight environmental and occupational effects. This was not completely novel because in 1964 and 1974 the HIS had estimated numbers of persons exposed to varying levels of radiation from medical and dental x-rays. What was new in the legislation was the extent of the specifications regarding diseases related to community air quality and water quality and the data required for identification of health problems in occupational groups.

Recommendations

The recommendations of the Panel are grouped

into four convenient categories for easy reference.

1. **Content of questionnaire.** The relative role and contributions of core and supplementary questions are defined in the full report, and guidelines are provided there on how to decide what should be considered a core question. Core items provide for continuous, long-term series of data; no item should enter that category unless time trends are absolutely essential. Any net increase in core means shortening or eliminating supplementary items because the current interview is considered full and any time increase for one section requires a compensating reduction elsewhere. The supplementary questions should be responsive to the needs of federal agencies and other entities, and routine solicitation of newly suggested items should be obtained every two or three years.

The items in core relating to chronic conditions were the ones about which the Panel had serious reservations. It did not feel that the detailed data on chronic conditions justified the considerable costs in collecting and tabulating them. Instead, it recommended that the list of chronic conditions for which prevalence data are collected be shortened to eliminate those conditions for which useful prevalence data cannot be estimated. Furthermore, the members of the Panel felt that the coding of all chronic conditions to the four-digit code of the ICDA (8th revision) was not only a delaying factor but also an unnecessary expense, since such detail implies a level of accuracy and reliability that is not warranted.

Much of the criticism leveled at the way in which data on chronic conditions are collected and tabulated by NCHS was felt to be equally applicable to acute conditions. In particular, the Panel challenged the coding of all diseases to the four-digit code. The Panel also suggested that data on problems, symptoms, and complaints presented by a patient when consulting a physician might be tabulated.

There was doubt expressed about whether data on height and weight need be core questions or might be obtained periodically every few years. The original reason for including these questions as core was to identify groups of persons with problems of obesity and to relate them to hypertension. Using these data on height and weight for detailed epidemiological studies might be challenged because of their unreliability as well as the accuracy of the health condition itself. The HIS was urged to assess the utility of these items on height and weight in order to justify their retention as core.

To replace the deleted items, the Panel recommended the inclusion of some questions on mental health, occupational and environmental health, and health insurance and expenditures. NCHS was urged to consult with other federal agencies for specific suggestions on how their data needs might easily be met in these areas. In reviewing the requests for supplementary items, the Panel urged the use of an external group of advisors to review the requests. There were other related functions that such an advisory committee or panel might fulfill.

Regarding the sociodemographic variables included in core, the Panel urged that attention be given to achieving some degree of uniformity so that all federal programs, or at least NCHS, would use the same categories to collect these data. The present system has too many differing definitions of terms, and even data on age are inconsistently collected within NCHS. Expanding the items to include religious preference and disaggregating the data for those 65 years of age and older were also recommended.

2. **Methodology**

a. *Total Survey Design.* In the historical development of sampling survey methodology, the initial emphasis was on minimizing sampling error for a given cost in order to estimate the size or frequency of one or two items. More recently, new research has focused on designing surveys to minimize total error that incorporates response bias and variance with sampling error. This has led to Total Survey Design (TSD), an approach whose effective application requires knowledge of data objectives as well as of the error structure. The Panel believes that the HIS has not yet made a systematic attempt to evaluate significant alternatives to the present design of the survey such as the use of bounded interviews and partial rotation of the sample.

The Panel recommended that NCHS undertake a careful and comprehensive evaluation, based on TSD principles, to determine the feasibility and desirability of several alternatives to existing HIS data collection and sample design features. Design changes resulting from this evaluation should be timed to coincide with the redesign of the sample following the 1980 Census of Population and Housing.

The Panel also recommended that TSD principles be followed in providing technical assistance to help state and local governments meet their health data needs (see Section 4.c below). Moreover, low-cost survey methods such as computer-assisted telephone interviewing with random-digit dialing should be given full consideration as alternatives to replication of the procedures used in the national HIS.

b. *Strengthening the sample.* Increasing demand for wider coverage of subject matter and greater detail for analyses means that the HIS must be strengthened by increasing its capacity through a more efficient sample design. This will facilitate more detailed and reliable data for different demographic, socioeconomic, and geographical sectors of the population on a more timely basis. The specific change in sample design was not prescribed except to point out that the new one should be cost-effective.

c. *Continuing versus periodic surveys.* The question was raised whether the HIS should be a continuous survey or a periodic one every n years. Considering all of the elements involved, including interviewer training and bias, the Panel felt that continuation of a continuous survey was preferable.

d. *Redesign of HIS in the 1980s.* The sample for HIS is closely linked to the Current Population Survey and other sample surveys of the Bureau of the Census. In fact, HIS data are collected in a subset of primary sampling units (PSUs) of the Census Bureau that is based on 1970 census data. When the Bureau of the Census updates its PSU design after the 1980 census, HIS will need to make its needs known or adopt an alternative strategy. Inasmuch as the change in the Bureau of the Census design is likely to be a major one, the opportunity for a strengthened sample and TSD to be incorporated will be unique for HIS. The HIS should start immediately to make an intensive review of its own design features (e.g., longitudinal versus cross-sectional approach, respondent rules, length of reference period, use of diaries, interview methodology, and other features) to take advantage of the opportunities provided by the post-1980 Census Bureau design. Moreover, in order not to defer the study, this review should be undertaken by staff who carry no responsibility for ongoing HIS operations. If such staff are not available, the Panel recommended use of an appropriate contractual arrangement.

e. *Methodological research.* Although a tremendous wealth of survey knowledge has come from the HIS, the Panel believes the opportunities for whole new approaches in data collection and analyses are not being fully exploited. There are new data requirements, new techniques of interviewing and data collection, new processing technologies, and whole new concepts of error structure that require research by HIS staff. The Panel therefore recommended that not less that 10 percent of the HIS budget be allocated to methodological and developmental research.

f. *Cross-analysis with other data sets.* The HIS in only one of a large number of data sets compiled by the federal government. Analysis of HIS data in isolation from all of the other data sets deprives the nation of valuable linkages. Examples of typical linkages are provided in the full report, and these illustrations reflect possible usefulness of cross-analyses for policy issues as well as etiological issues in epidemiology. NCHS was charged with establishing a procedure for cross-analysis of HIS with other data sets.

3. **Analysis, use, and timeliness of dissemination**

a. The value of the HIS is ultimately dependent on the extent to which it is used. There are three strategies for serving users that should be strengthened.

(1) More publicity should be given to the fact that the HIS can provide raw data and tabulations on data tapes to people who need such information and can reimburse NCHS.

(2) Quick estimates of some items could be obtained with two weeks of interviewing. The current lag seems to be due to the lead time needed to develop and incorporate questions. More attention should be devoted to ways of incorporating special purpose items in the HIS without too much delay.

(3) The ongoing analyses are less than desirable because of lack of staff availability. In this respect, highest priority should be given to increasing staff positions to strengthen the analytic capability. The HIS should explore ways of offering special incentives to staff members for producing and disseminating analyses of HIS data. Also, expanded use of the IPA mechanism should be sought to recruit talented persons to devote themselves to analysis of HIS data. Finally, when responding to requests for special data col-

lections from other federal agencies, part of the joint commitment should include staffing for the analysis of the collected data.

b. The HIS should continue to study users of their data and to evaluate the usefulness of given data sets.

4. **Role of HIS in meeting subnational needs**

a. Subnational data needs were defined in an earlier section when discussing new legislation in PL 95-623. This demand for subnational data has increased exponentially during the past decade because of federal promotion of health planning and regulation at these levels. These changes, plus the inflationary effects of increases in all health expenditures, have made it essential to study the distribution of health services as well as their impact. Thus, there are state and county needs as well as the needs of Health Systems Agencies and Professional Standards Review Organizations.

Other federal legislation, such as support for Health Maintenance Organizations and the placement of professional manpower in the National Health Service Corps, has also added to the need for local data. The underserved and shortage areas have tremendous stakes in having such data available on a timely basis.

Health interview data are unique in that they do not arise through routine administrative procedures such as licensing of professionals or reimbursement of health expenses. Such interview data are useful because they provide information on persons whether or not they use the health system.

The full report documents several illustrations whereby the use of health interview data can provide useful information at the state and local level on resources, utilization of facilities, expenditures, health manpower shortage areas, etc.

b. *Capabilities of the present national sample.* Each week's interviews constitute a probability sample of the nation, but the primary focus is on estimates for the entire country based on 52 weeks. Some estimates, however, are based on 13 sample weeks, others on 104, and occasionally a single week for a high-incidence phenomenon. Past practice has restricted publication of direct estimates for subnational data, and the present design is not even well adapted for estimates of individual states or smaller areas. With redesign of the survey after the 1980 census, the present general structure may make it possible to produce direct estimates for the 8 largest metropolitan areas and 8 states. Even if the sample size were quadrupled, there would be 22 states for which annual direct estimates would not be sufficiently reliable for publication, and subnational data below the state level would still be unavailable.

Therefore, some modifications of current procedures could be adopted, or should be explored, to strengthen the subnational data program after 1980. One simple method would be to base subnationa' data on 104 weeks of interviewing; there would thus be biennial publications for most of the global statistics. Another possibility would be to divide sample resources into two parts. Part A would be the continuation of the present system and would provide global statistics for the nation and for larger states and metropolitan areas. Part B would distribute resources differently, reduce the number of items queried, and experiment with lower-cost collection techniques such as mail and telephone.

Other solutions involve greater use of synthetic and composite estimates for the subnational areas. The HIS has used synthetic estimates, and the Panel recommended greater use of them as well as the Composite Estimator in which a synthetic estimate is one component.

The growing demand for subnational data will probably not be fulfilled by any of the foregoing procedures, including the strengthening of the design of the survey, or by increasing the sample size within reasonable bounds.

c. *Technical assistance.* A more suitable and practical method of meeting subnational data needs is for NCHS to provide a complete range of technical assistance and demonstration activities on health interview survey methods and the uses of such data. This technical assistance should be provided at varying times and at minimal cost to those agencies and organizations in need of such assistance. This program of technical assistance should take advantage of relatively low-cost data collection techniques when they can be expected to produce data with adequate reliability for local data needs. Moreover, the HIS should develop guidelines for the local staffing, funding, and uses of the data in these local areas. Contractual support and IPA personnel may be used for this technical assistance, but the primary source of this service should be the professional staff of NCHS,

whose numbers and resources will probably need to be increased to accommodate this demand.

Summary

The HIS is of vital importance to the health data system of the United States. It has served this country well during the past 22 years, but new developments in TSD, improved data collection techniques, new methods of processing and analyzing data, and other methodological improvements must be studied and applied promptly if NCHS is to accomplish its mandate to fulfill the growing demands at all levels of government for comprehensive health statistics.

Open discussion: The national Health Interview Survey

In reply to a query on what process the Technical Consultant Panel had undergone in order to arrive at its recommendations, Greenberg indicated that this was a good question since the summary report had not allowed time to go into such details.

The Panel met every three of four months over a period of two years with a planned agenda to cover two days of deliberations at each meeting. The agenda was prepared in order to inform Panel members about the status of the HIS and its history, sample design, current operations, and problems. The Panel heard reports on how core and supplementary questions were selected, tested, and tabulated; and it met with members of one or two other federal agencies to learn of their data needs and the difficulties encountered in having items included as supplementary questions. A representative from the Bureau of the Census explained plans for sampling redesign after the 1980 census. The Panel heard reports from members of NCHS both inside and outside the Division of Health Interview Statistics (DHIS). One day was spent with Dan Horvitz learning about the National Medical Care Expenditure Survey and the design and survey techniques used in that survey. The Panel requested some research calculations by the DHIS staff on the effects of changing sample size and how well such changes would permit direct, reliable estimates of subnational data. It also studied an analysis of the users and uses of HIS data.

The purpose of all this inventory taking was to help the Panel arrive at a diagnosis of where the HIS was relative to where the Panel felt that it belonged. The members of the Panel then considered possible strategies so that the HIS could advance in the direction from its current position toward where the legislative mandate specified it should be. In considering options for change, the members of the Panel, who were all researchers or administrators with research backgrounds, recognized that it would be futile to recommend certain changes unless adequate resources were available to implement them. Therefore, whenever appropriate, they indicated a need for more funds, more staff positions, a greater use of IPA and contract mechanisms, and other necessary measures. Of course, not all of the suggested modifications require additional funding, and such changes should be planned as soon as possible.

Another question was asked on how the report was to be disseminated and its recommendations implemented. Greenberg replied that the Panel has no authority to prescribe changes. It was a subcommittee of the National Committee on Vital and Health Statistics (NCVHS), and the formal report will be submitted to that group. Unofficially, of course, a copy will go directly to Dorothy Rice as Director of NCHS. If NCVHS approves the report, it will be forwarded to Ruth Hanft, Director of the Office of Health Research, Statistics, and Technology, and to Julius Richmond, Assistant Secretary for Health, DHEW. It is hoped that by supplying all these other individuals with a copy of the report, means will be found to fund those changes requiring additional resources.

The report itself was prepared to be considered as an internal document and is not necessarily scheduled for publication by the Government Printing Office. On the other hand, sufficient copies will be distributed so that anyone who wants to, or should, see the report will be able to do so. The report has been written in a style appropriate for publication of all or part of it by NCHS in the "rainbow series." These conference proceedings are an additional form of publication. The publicity generated by these various means of dissemination should facilitate early implementation of those major recommendations considered sound and feasible.

Implications of survey research for health policy and programs

Ruth S. Hanft, Office of the Assistant Secretary for Health, DHEW

It is a pleasure for me to address this audience, particularly at a conference jointly sponsored by the National Center for Health Services Research and the National Center for Health Statistics. These two Centers are beginning the analysis of a major research survey, the National Medical Care Expenditure Survey. Not only will this survey provide much-needed current data on utilization, expenditures, charges for health care and in-depth information on health insurance status, but it also served as the pilot survey for the National Medical Care Utilization and Expenditure Survey. The cooperation of the two Centers in this survey, along with the use of the data by the Department and others, illustrates the importance of survey data in health services research, policy analysis, and policy planning.

Health surveys are relatively new in this country, but over the past 20 or 25 years they have helped to shape health policy and programs. For example, when the final push for health insurance for the aged began in the 1960s, survey data on health and use of services by the aged were used to document need and to plan the Medicare program that emerged. Similarly, data on differentials in health status among income groups were important in designing the programs of the War on Poverty and other efforts to improve delivery of health services in the United States.

Health care surveys are used for a wide variety of purposes including the identification of health status problems, health care use, and financing, for management of specific programs, and more and more for program planning purposes.

When the President's proposals for national health insurance go to Capitol Hill, they undoubtedly will contain supportive data from health surveys. As the committees of the House and the Senate debate the President's proposal and others that have been introduced and determine their positions, they too will be drawing data gathered in health surveys. People in and out of Congress will undoubtedly differ on what the data really mean. Further, decision making is tempered by limitations of resources and relative priorities among many societal goals. Nevertheless, the fact that extensive information on health status and health services is available, in spite of all its limitations, means far better decision making than we would have had a few years ago.

Policy decisions have not always been so well based. Data for such use, although there are still many gaps, are taken for granted today. It is a development within the lifetime of many at this conference. The techniques and tools did not exist earlier for measuring morbidity and other aspects of health in sample surveys. Thus, it was not until 1935–36 that there was a nationwide health survey of the U.S. population, although there were earlier community studies.

Not until 1956 did Congress authorize, in the National Health Survey Act, continuing studies of illness and disability in the U.S. population. By this time, the basic methods had been devised for producing the data. Research conducted during the 1940s and 1950s within universities, as well as methodological work sponsored by the federal government, had produced refinements of questionnaire design and survey techniques that substantially reduced errors in survey data. The theory and application of techniques of sampling human populations was developed to a point where relatively small samples of the population could provide information about the population as a whole at reasonable cost and within stated margins of error.

Since 1956, the National Health Survey has become the umbrella title of several survey systems, all of which have benefited by continuing improvements in survey methodology. This family of surveys now includes the Health Interview Survey, the Health and Nutrition Ex-

amination Survey (HANES), the Family Growth Survey, the National Ambulatory Medical Care Survey, the Hospital Discharge Survey, the Survey of Nursing Homes, and now the National Medical Care Utilization and Expenditure Survey. These surveys provide great flexibility to address current issues as well as to provide trend data on a variety of issues. For example, a special emphasis in the next HANES cycle will be the nutritional status of the Hispanic population. The Family Growth Survey will address teenage pregnancy.

We have not only a wider range of data to bring to bear on current issues but also more valid and reliable data. I would like to share with you a few illustrations of the use of survey data in the health programs of HEW. Deliberations in Congress and the White House may be the highest level of use in our government, but survey data are relied on at every level.

Many of the issues that we face are continuations of the past. Costs are a paramount concern. There remain problems of access to care, particularly for inner-city and rural residents. Health resource utilization and expenditures data illuminate specific problems. We have begun to question the philosophy of the sixties that more is better—more health manpower, more utilization of services, more technology. We have recognized that advances in health status will be dictated to a considerable degree by the personal actions that people can take to protect their health, and we have begun to stress health promotion. Given the vast number of options that might go into a health promotion program and the scarcity of resources, it is imperative that the component initiatives be systematically analyzed to minimize redundancy and to attack as broad a range of preventable illness as possible.

Data from health surveys, including knowledge of the extent and nature of the burden caused by various health conditions, aid us in determining the modifiable personal habits and environmental factors that occur with sufficient frequency and consequences to warrant intervention. Sample surveys provide information that has been extremely valuable in targeting health promotional efforts. In planning a program, it is necessary to identify the population groups with the greatest need for preventive and treatment activities of various types. For instance, obesity has been shown to increase the risk of early death. The Health and Nutrition Examination Survey points to women in low-income households and certain racial and ethnic groups as being particularly prone to obesity. In other words, the data suggest target groups for prevention programs. Additional information regarding eating habits, physical activity, attempts to lose weight, and other considerations relevant to prevention are available to us from sample surveys of NCHS. Similarly, data are provided that are essential in planning, monitoring, and evaluating prevention activities aimed at smoking, control of high blood pressure, accidents, and a variety of other personal and environmental actions.

The current drive to eliminate measles is partly the result of the sample surveys conducted in recent years by the Center for Disease Control and the findings that immunization levels in children were declining while reported cases were rising. We are beginning to reverse that trend.

With respect to national health insurance, we will soon begin to feed into the planning process data gathered in the National Medical Care Expenditure Survey, conducted jointly by the National Center for Health Services Research and the National Center for Health Statistics. Although we have a fair amount of knowledge of how much we as a nation spend for health, we need to know a great deal more about who uses the services, who ultimately pays the bill, and trends over time.

The group of surveys that is labeled National Medical Care Expenditure Survey will answer many of our questions and provide baseline data for the pre-national health insurance era. In this study, data were obtained from a panel of households followed throughout the year 1977 and from the providers of care, insuring organizations, and employers. It will make available to us more detailed data than we have ever had before on utilization and charges for medical care services, including hospitals, physicians, and dentists, and on such expenses as drugs. From the data, NCHSR will construct econometric models to study the various options for the financing of health care. The President, the Congress, the Department, and the people themselves will be far better prepared to make informed decisions about national health insurance than we would have been without this survey.

One final example relates to improved delivery of health care services. Sample surveys are one of the means by which we have watched the increasing volume of health care, much of it probably not emergency in nature, provided in emergency rooms and outpatient departments of hospitals, and the greater use of these facilities by low-income groups. Routine primary care in these settings is expensive and often inappropriate. In the next fiscal year,

funds permitting, we will start a new program to convert traditional outpatient departments of hospitals into family primary care centers.

Our efforts to encourage growth of Health Maintenance Organizations similarly are bolstered by survey data that show less hospital use by members of such groups than by people with other forms of insurance.

These are a few examples of the ways in which survey data highlight needs for resources and services and indicate where public and private investment should be used to improve health status.

Probably too often overlooked, except by you in the field, is the contribution of agencies such as NCHS to methodological advances. The National Health Survey Act also provided that NCHS "study methods and survey techniques for securing ... statistical information, with a view toward their continuing improvements." Under this provision, NCHS has supported a great deal of research in survey design and methodology, in addition to its own in-house program. Speakers at your tribute to Elijah White gave ample evidence of this aspect of the Center's work in connection with the Health Interview Survey, and there are many examples from the Center's other data systems. The goal has been not only to advance the science but also to evaluate the data collected and to improve their precision and reliability. In addition, advances that I can sum up in the words "the computer" have enabled faster processing and greater timeliness in the release of data and in new ways of displaying findings.

Since this is a working conference, I thought it might be appropriate to indicate major areas where traditional survey data have, so far, been less helpful and to stress the need for greater attention to the analysis of existing data. Data for surveillance of environmental hazards are becoming an urgent need. And last year, in the Health Services Research, Health Statistics, and Health Care Technology Act, Congress required that the National Center for Health Statistics collect statistics on "environmental, social, and other health hazards."

Some data are being gathered in the Health and Nutrition Examination Survey, but in limited amounts. Many other agencies have bits and pieces of data. But as we began to look at the Congressional mandate to the Center, we found little information that would enable us to define the problem—what is it that we need to measure—and even less to suggest how the data might be gathered in the ongoing data systems. National samples large enough to differentiate significant levels of exposure in the various geographic regions may be prohibitively expensive. So, too, may be the collection of residential and occupational histories in the Health Interview Survey.

In addition, longitudinal data on individuals are clearly essential. The induction period for many cancers may be 20–30 years. Other chronic, degenerative diseases may develop as a consequence of even low levels of exposure to a hazard over an extended time period. The capacity to identify individuals exposed to certain occupational or environmental factors during a certain time period and then to monitor health events in their lives would constitute a powerful tool in the search for health effects. It could serve also to measure the effects of health programs and other social change in the health status of the individual, another capacity we now lack. Such data would influence national health policies, as other data have done.

Another area of concern is small area data. Although we now have much-needed national data on health status, health resources, use, and expenditures, with resources becoming constrained we will need to target on problem hot spots. Small area data on infant mortality, measles, other disease concentrations, resources in the area, and use of services are needed to focus scarce resources on specific issues.

More effort is also needed on utilizing the data that we have in hand. Health services researchers and policy analysts have not made optimal use of data that already exist. Many of these analysts are not aware that these data are available for analysis. We need to find ways to inform researchers and policy analysts about existing data and to assist them in using these data.

We must also be on our guard against misuse, abuse, and misinterpretations of data. I need not provide this audience with examples of the problems of misuse of data for advocacy and political purposes. It is essential that statisticians, researchers, and policy analysts be outspoken against inappropriate use of data and that we clearly and carefully spell out our assumptions and caveats in our use of data and surveys.

Informed decision making—in HEW, in Congress, among our people generally—is one end of a long chain of investigation, experimentation, refinement, and information gathering. The other end is in conferences such as this, where people from university centers for survey research, from private associations, and from agencies such as NCHS and NCHSR are exchanging ideas and information.

Open discussion: Implications of survey research for health policy and programs

Morton Israel began the discussion after Hanft's speech by noting that in a recent New York study, deaths from respiratory disease were found to be higher in certain geographic areas. He wondered if there is any interest among federal agencies in funding small area environmental studies. Hanft responded that there are several agencies gathering data in this area, and a council will begin to coordinate efforts in the area of environmental affairs.

When asked whether RFPs or reports would be the product of this coordinating council, Hanft explained that the council would have three charges:

1. To determine the feasibility of a study of exposure to environmental hazards;
2. To define a study to be carried out by the Institute of Medicine; and
3. To look at all the environmental data currently being collected, define the gaps, and make recommendations for new data collection efforts.

It is not likely that RFPs will be issued, owing to lack of funds. It has not been decided yet just what will be done in the environmental area.

One of the conference participants wondered about the usefulness of research data to policy formation. Hanft responded that research data are examined by staff and incorporated in policy speeches. She has been asked to report shortly on the quality of the data being supplied to the Department (on costs, incidence of disease, etc.).

It was noted that the speaker had not mentioned psychological health statistics. Hanft indicated that there certainly is an interest in the mental health area. However, the data on incidence of mental health problems are quite variable. The National Institute of Mental Health is interested in improving the data base. But this improvement will be a long, slow process, probably taking several years. ADAMHA, NCHSR, and NCHS are trying to arrange for an exchange of data between agencies.

A question was raised about the existence, mentioned by the speaker, of a great deal of data that need analyzing: Do special areas of interest exist of which we should be made aware? Hanft suggested that the "gold mines" of what is already available should be explored before anyone sets out to collect his or her own new data.

The suggestion was made from the floor that it would be useful if the agencies would issue some RFPs letting research organizations know what data sets are available for analysis and what issues are of interest, but once again agency shortage of funds appears to make this impossible.

It was pointed out that special methodological problems are arising in some EPA-sponsored community surveys on hazard exposure, for which it is necessary to draw blood samples, gather hair specimens, etc., and that perhaps NCHS might work to further public cooperation in these studies. Hanft responded that this is being worked on now.

In answer to a question on how we will get more manpower for health and survey research, Hanft indicated that she spends at least 10 percent of her time on this issue. It is a constant uphill battle just to maintain current manpower levels, but increased funds have been promised for training personnel in the public health area.

In an attempt to cheer up the American participants, one of the Canadians reported that after work on a health study had been in progress in his country for a year, funds necessary for its completion were cut off—so things appear to be worse in Canada than in the U.S.

Sirken reminded the group that there are administrative and legislative procedures available to move people temporarily from universities and other organizations to government, but the agencies are not taking advantage of this

machinery. Hanft added that there is also another technique available for filling research and analysis positions—the "Service Fellow." However, the number of such temporary positions is strictly limited.

Selected methodological features of the 1980 census

Charles D. Jones, Bureau of the Census

Basic census procedures

In April 1980 the Census Bureau will set out on its every-ten-year effort to enumerate the population of the United States. This time around, the population will number about 220 million persons. Over 90 percent of the population will be enumerated by what we call the mailout/mailback/follow-up procedure. The remaining 10 percent, which is primarily located in the sparsely settled western Mountain States, will be enumerated by the more traditional door-to-door canvass.

The mailout/mailback/follow-up procedure has three major steps:

1. Prepare a listing of the addresses for all living quarters, which will serve as the mailout and control record.
2. Mail a questionnaire to each address, requesting the householder to complete the questionnaire and mail it back to us.
3. Use the control listing to identify households that did not return a questionnaire by mail. An enumerator is then sent to each such address to conduct the interview.

The door-to-door canvass proceeds as follows:

1. In late March, the post office delivers an unaddressed questionnaire to every residence. The householder is asked to complete the questionnaire and hold it until an enumerator calls.
2. Beginning on April 1, enumerators systematically canvass the area, recording each living quarter's address.
3. Simultaneously, the enumerators call at each housing unit to pick up the completed questionnaire. If the questionnaire has not been completed by the householder, the enumerator fills out a questionnaire at that time.

These are the basic steps, but a number of other procedures are important to the operation. For example, there are procedures to check completeness of the questionnaires, to improve population and housing coverage, and to enumerate special populations who do not live in households, such as transients, migrant workers, and persons in group living arrangements.

Although the basic procedures for the 1980 census are similar to those used in the 1970 census, there was considerable testing prior to the upcoming census. Beginning as early as 1975, we conducted a series of pretests focusing on methods to improve coverage, field procedures, and response accuracy.

The procedures to be used for 1980 also reflect the results of research conducted in earlier censuses. As you can see, the procedures tend to maximize self-enumeration. Studies conducted in prior censuses show that enumerators can have an effect on the data that they collect. In a door-to-door canvass without a prior mailout, an enumerator can influence the responses for all units in his or her assignment. The effect on the data is to decrease accuracy, in somewhat the same way as the clustering effect in sampling works to increase sampling variances. To reduce this effect, which can be large for local area data, we have been attempting over the past three censuses to have householders enumerate themselves, that is, to have them act as an "enumerator" to complete the work for only one household.

In the area of coverage improvement, the focus has changed from trying to make marginal improvements for *existing* procedures to investigating *new* procedures, many of them redundant, to improve the counts. In earlier censuses, for example, we attempted to improve coverage by improving the basic listing procedure, instituting quality control, increasing the supervisor-to-enumerator ratio, and developing "better training." We probably have achieved the biggest portion of positive effects from such

changes, so that further gains from improvements in these areas would be marginal. The approach for the current census is to use a number of independent sources, allowing each to make its individual contribution to improving coverage. The development of the control list of addresses in urban areas will illustrate this point.

We start by purchasing a list of mailing addresses. For some areas, we buy the list from two different firms and merge them. In other areas, we merge the purchased addresses with those from the prior census. The merged list of addresses is then submitted to the post office in mid-1979 for a specially arranged update and correction process. During this operation, the postal carriers report addresses that are not on the list. These addresses are then added to the list. In early March 1980, questionnaires are addressed from this updated list and sent to the post office for another address check. At about the same time, enumerators make a physical canvass of each block and building to identify units missing from the list. Units from both of these checks are added to the control listing and questionnaires are addressed. Next, in late March, as postal carriers deliver the questionnaires, they report back on any units for which there were no questionnaires.

In addition to these major checks, there are a number of others. For example, each respondent is asked to report the number of living quarters at his or her address, and if more units are reported than are in the control listing, the situation is investigated. Any listing deleted from the control register is checked again by a second enumerator. Next, primary counts are produced for each city block and census tract, and these are subjected to two reviews as part of what is called a "local review program." The local census office itself compares these counts with counts from past censuses and from projections to identify significant deviations. The counts are also submitted to local officials so that they can identify areas where the counts are suspect. Any significant deviations resulting from either the office review or the local officials' review are investigated.

As you can see, we have developed a control listing of housing units by having input from commercial vendors, the prior census, postal carriers on three occasions, enumerator canvass, the respondent, and local officials.

Of course, in developing a program with redundancy (that is, different procedures aimed at picking up the same missed units), one has to evaluate whether the unique contribution from each source is worth the additional cost and effort. Research conducted in our pretests was designed to evaluate the efficiency of each of a variety of sources. For example, to evaluate the contributions of the enumerator canvass, in pretesting we performed the canvass but did *not* include the missed units from the canvass in subsequent operations. Later, we matched the results of the enumerator canvass to the control list. By this technique we were able to estimate the missed units identified solely by the enumerator canvass, as well as those that were picked up by other sources.

Coverage improvement procedures for ensuring completeness of the listing of household members within an enumerated unit have followed the same principle of redundancy, although this task is more difficult because independent data sources are lacking. To a large extent we are depending on the household respondent to list everyone. In an attempt to encourage respondents to report all household members, we have developed extensive public information and community service programs. We hope both to convince people about the need to be enumerated and to ease their concerns about confidentiality. In addition, we have introduced some direct procedures to try to improve within-household coverage. We have included probes on the questionnaire to identify households where someone may have been left out—for example, to find out about persons whom the respondent was unsure whether to list. We recheck the roster of listed persons with the household when we recontact the household for other purposes. For example, some questionnaires are mailed back not completely filled out. These households are telephoned to collect the missing data. During this contact, we verify whether all household members are listed. Finally, we are matching independent lists of names to the questionnaires in selected areas and adding persons identified from such a match as missed in the census. For example, persons in drivers license files and in Immigration and Naturalization Service files are checked against the census, with unmatched persons followed up for enumeration.

Mail response rate

The mail response rate is a key figure in conducting the census by the mailout/mailback procedure. The rate describes the proportion of households that filled out the questionnaire and mailed it back to the Census Bureau. The complement of this rate, indicating the households that remain to be visited by enumerators to conduct the enumeration, combined with other data such as vacancy rates, estimates the volume

of work remaining, the funds that will be needed to accomplish this work, and the staffing levels needed to accomplish this work on schedule.

In 1970, about 87 percent of all households in mailback areas returned their questionnaires in the mail. For 1980, we have estimated this rate at 80 percent and have budgeted, planned, and organized the field work on the basis of this figure. The lower estimate for 1980 is based on our pretest censuses conducted over the past few years, in comparison with similar pretests carried out prior to the 1970 census. We do not have a definitive answer to why we are getting slightly lower response rates in this decade, since we have not made major changes in questionnaire content or burdens on the respondent. I noted in the last health survey research conference that there was a discussion on the similar problem of reduced cooperation rates in surveys.

In addition to working on our public information efforts to alleviate the problem, we tested whether mailing a reminder card to nonrespondents would boost the mail response rate. Much of our publicity is aimed at an April 1 mail return day. The reminder card was to be mailed about April 4–5 and emphasize that "it is not too late to send in your questionnaire."

In our pretesting, in a split-sample experiment, we observed a small increase in the mail response rate through use of the reminder card, but the increase was not worth the effort and expense of sending the reminder cards. Thus, we do not plan to undertake this activity in 1980. Instead, we plan to emphasize the "it's not too late" message in our post-census-day publicity efforts.

In classroom experiments, we have found indications that the format and appearance of the census questionnaire, if changed, might induce more people to complete the questionnaire. In these experiments, the regular census questionnaire and experimental questionnaires with alternative formats were studied. There were some indications that the regular form appeared to be more complicated than some of the experimental forms that we were testing. However, that evidence was not based on a sample of people, and the analogy between the classroom and the real world is limited since classroom experiments cannot tell us about the mailback rates that we can expect under actual census conditions. Consequently, we are undertaking research with alternative forms in the 1980 census by including national samples of them in the original mailout. One of two variants that we are testing in the 1980 census is a reformatting of the present questionnaire by laying out the answer categories so that the respondent can work across the page rather than from top to bottom. The other variant is a non-machine-readable questionnaire with a layout that is not constrained by the necessity for a machine to read the answers without prior clerical coding.

Evaluation and Research Program

I would now like to discuss the Evaluation and Research Program for the 1980 census. Starting with the 1950 census, the Census Bureau has had a major program of evaluation and research carried out in conjunction with each decennial census. The purpose is to develop estimates of errors so that users can better understand the strengths and limitations of census data and so that methodological improvements can be planned for future censuses.

For the past year or so, staff throughout the Census Bureau has been involved in planning the 1980 Evaluation and Research Program. The program is being developed in four major areas.

First, experimental procedures are being planned to be imbedded in the 1980 census field and office work to test procedures that may become the procedures for future censuses. The census, with its condition that cannot be fully simulated in a pretest, is the best place for this testing. Thus, one may view the upcoming census as a test bed for developing and testing future census procedures. A variety of suggestions are currently being considered: (1) testing an alternative procedure to the mailout of questionnaires in which an enumerator would leave a questionnaire for mailback as he or she updated the address listings (this is in place of having the post office deliver the questionnaires); (2) investigating the feasibility of telephoning some non-mail-return households rather than sending an enumerator to each one; (3) testing the use of alternative training techniques to train temporary personnel for census work; (4) developing curricula with universities by which students would, on an experimental basis, become enumerators to fulfill some research course requirements; (5) developing computer systems to replace clerical operations in the editing of questionnaires; and (6) as I mentioned earlier, testing the use of alternative questionnaire formats to increase mail response rates.

Second, evaluation studies are being planned to estimate response biases and variability for the various data items collected in the census. The techniques to be used here include a reinterview study, a series of record checks, and an

enumerator variance study. In the reinterview study, a sample of households is visited by specially trained interviewers to ask a more detailed set of questions than was feasible in the census. These detailed questions and procedures attempt to develop a more accurate response by investigating those areas where questions and instructions may have been unclear or ambiguous and by interviewing the most knowledgeable respondent in the household. In some cases, the reinterview is also used to repeat the identical question in order to obtain an estimate of simple response variance. Record checks are conducted when the data in the record source may be considered of higher quality than in the census. For example, the responses in the Current Population Survey in March 1980 will be matched to the census responses, and an estimate of response differences will be made. The assumption is that the Current Population Survey has more accurate responses owing to its permanent staff and their continued training on the concepts and methods of obtaining accurate responses.

For a few questions in the 1980 census, there is a serious problem of how to measure response quality. For example, the question on ethnicity depends on the respondent's self-perception about the ethnic group to which he or she belongs. There are no objective criteria that we can specify to decide in which category each individual belongs. The enumerator variance study would estimate the effect that the enumerator had on the data collected. With self-enumeration, the enumerator's role is reduced, but he or she can still affect the quality of data through enumeration of nonresponse households and other activities.

Third, a program of coverage evaluation is being designed to describe the characteristics and distribution of the undercounted population. For the U.S. estimates, a technique of demographic analysis will be used. In this technique, U.S. data on births and deaths, data from prior censuses, and data on immigration and emigration are combined to estimate the size and characteristics of the true total population. Census counts are then compared with these estimates, and differences are considered to be principally errors in the census counts. Unfortunately, we do not yet have an estimate of illegal migration. Since it is generally believed that there is substantial net illegal immigration, the demographic estimates will be flawed unless we can develop an estimate for this component of the population.

We have produced demographic estimates of census error for the United States as a whole in conjunction with the past three censuses, but as yet, we have not found a satisfactory way to disaggregate these counts for subnational areas within the U.S.—for example, for each state, city, and county—owing to the lack of reliable data on internal migration independent of the census.

There has been an increasing demand for subnational estimates of the undercount, and indeed, there is mounting pressure to correct census figures for fund allocation purposes. The pressures to adjust census counts are becoming enormous. Producing subnational undercount estimates is one of the significant problems that the Census Bureau will have to solve in conjunction with the 1980 census.

In response to these needs, we have undertaken the development of a large-scale post-enumeration survey and administrative records match to produce subnational estimates of the undercount. The survey will be designed to produce reliable estimates of undercounts for each state and for a few large cities and SMSAs.

In the past, post-enumeration surveys have been undertaken by selecting a sample of areas, having specially trained enumerators canvass and enumerate these areas independently of the census, and matching the names obtained to the census records to identify missed persons. Unfortunately, these efforts have yielded estimates with serious biases, especially for adult males. For example, in 1960 the post-enumeration survey estimated undercounts for black males aged 20–34 at 2.5 percent. Correspondingly, demographic analysis, which we feel is a more credible technique, estimated the rate at 18 percent. We believe these biases are due to correlation between the census and the post-enumeration survey. That is, persons missed in the census are missed in the post-enumeration survey for some of the same reasons. For example, persons deliberately concealed from census workers and persons with tenuous household attachments tend not to be reported either to the census or in the post-enumeration survey. On the other hand, persons already reported to the census are fairly easy for the second interviewer to locate. In other words, the census and survey are not independent. Census enumerations and census omissions do not have the same probabilities of being listed in the post-enumeration survey.

Our hypothesis on one way to overcome this problem is to use an independent data source in which an individual's probability of inclusion would be independent of the probability of being reported to the census worker. This independent source needs to have current informa-

tion on addresses but does not need to be complete—it only needs to be representative of the population group being studied. To serve this purpose, we are investigating the use of IRS files and Medicare files to estimate the working and retired males through matching with census records. For females and children, who are missed at relatively lower rates, the post-enumeration survey seems to produce reasonably good estimates. The results from the record matching and from the post-enumeration survey would be combined to prepare the estimates for each state and area.

Fourth, we are planning a series of studies to evaluate the effectiveness of several new procedures that have been developed to improve coverage. We want to evaluate these in terms of the accuracy of application, costs, and effectiveness in order to provide information for use in planning future censuses.

Summary

In summary, the basic census-taking procedures for 1980 are similar to 1970, but with increased emphasis on improving the coverage. Coverage improvement approaches have changed from just refining existing procedures to taking advantage of independent, although in many cases redundant, sources of information. We can expect to achieve substantial improvements in housing unit coverage. Less progress has been made on improving within-household coverage by redundancy, owing to the lack of independent data sources. The mail response rates are expected to be lower in 1980 than in 1970.

The Evaluation and Research Program for the 1980 census is currently being planned and developed. Estimates of response error would be derived from reinterview studies, from a series of record-check studies, and from an enumerator variance study. Estimates of coverage error will be made by use of demographic analytical technique and, for subnational areas, by use of a post-enumeration survey and matching with administrative records. Experiments are being planned to be imbedded in the census to test possible future census procedures. New procedures in the 1980 census will be evaluated to determine their effectiveness in improving quality.

Open discussion: Selected methodological features of the 1980 census

A number of points were clarified in the floor discussion following Jones's speech.

In response to a question on whether there are any incentives to postal workers to provide complete and accurate information on addresses, Jones indicated that through cooperation with the Postal Service, procedures have been developed to produce complete and accurate information on addresses. Furthermore, an extensive training program has been planned and implemented for the postal workers who will perform the work. Finally, quality-control procedures have been built into the process to ensure a quality product.

When asked if there are any local assistance offices to help households complete the census form, Jones said that there will be about 700 offices, nationally, set up to provide assistance. The telephone number for each area will be listed on the census questionnaire.

Regarding the extent of double counting in the census, Jones thought that with all the checks and controls in the census process, their experience indicates this is a minimal problem.

The speaker was asked about provision in the census for follow-up surveys. Jones replied that certain questions tested appear to produce weak data when asked in the context of census methodology. For instance, the series of health questions relating to disability and impaired mobility seem to be underreported based on experiences in the pretests. A follow-up survey with about 100,000 respondents is planned to collect better and more detailed data on this subject.

In reply to a question about the usefulness of the Internal Revenue Service and Medicare files as verification sources, Jones explained that some testing carried out suggests that these files may contain representative proportions of missed persons. The Census Bureau has found that some individuals who earn minimal income file tax forms to obtain their refunds. Medicare also seems to provide a good source of names representing the elderly population.

When asked if multiplicity estimation has been tried, Jones noted that the Census Bureau has tried multiplicity estimation in conjunction with two pretests, but the results were not encouraging.

In making projections about the availability of the 1980 census data, Jones said that the Census Bureau expects to do generally better than for 1970. For example, plans are to publish block data in the period January–July 1981, compared with July 1971–January 1972 for the previous census.

When asked about how the Census Bureau is handling the Mid-Decade Census, Jones indicated that active planning has started. One full-time person is assigned to head the activity and coordinate planning. The first order of business is to determine what sorts of data are needed and at what levels of geographic detail. Once that is determined, they can address the type of design and data collection methodology required to meet the objectives.

As responses to a series of questions, Jones stated that a small number of individuals have been prosecuted for not complying with the census. He said that any rewards or punishments for responsible census staff to reduce undercounting come in the context that their personal performance is evaluated in terms of how well the agency performs its mission.

Jones said that there are no plans to collect religious affiliation in the 1980 census. Data for Spanish heritage will be comparable for 1970 and 1980. In both censuses, the individual identifies himself or herself in the appropriate category based on self-identification.

When asked if he thought that the machine-reading procedures used in the census are applicable to health surveys, Jones said that they were. However, for any specific application one needs to evaluate the costs, feasibility, and error rates of alternative data-entry systems.

In reply to a question on whether the Census Bureau has considered carrying out the census jointly with the Internal Revenue Service and tax returns, Jones stated that this has not been given serious consideration because of privacy and confidentiality concerns.

When asked about the nature of the specific health questions on disability, Jones explained that they determine whether a person has a physical, mental, or other health condition that has lasted for six or more months and that (1) *limits* the kind or amount of work this person can do at a job, (2) *prevents* this person from working at a job, or (3) *limits or prevents* this person from using public transportation.

Jones was asked whether some form of monetary incentive might improve completeness and accuracy of reporting. He replied that the idea cannot be entirely discounted, because of lack of data. However, Census Bureau research has shown that undercoverage or inaccurate reporting may be largely due to deliberate misreporting or lack of knowledge about what is wanted. These problems probably could not be overcome by the usual types of incentives considered in survey work.

In response to another question, Jones said that the Bureau of the Census has never tried sending a second census form to nonrespondents rather than just a reminder card. The forms are expensive to print and mail, so there would be a question of whether the benefits outweigh the costs.

When asked the total cost of the 1980 census, Jones gave the figure of about $960,000,000, which includes all phases of the census, that is, planning and pretesting, implementation, data collection, analysis, and reporting the data.

Replying to a final question on why April 1 was chosen as the date to begin each census, Jones said that in some earlier censuses, the enumeration began on dates other than April 1. Starting with the 1920 census, Census Day has been chosen as April 1. Considering possible alternative dates, (1) April 1 is a date when most of the population is at home (few persons are on vacation), which minimizes enumeration and coverage problems; (2) the weather is generally good enough throughout most of the U.S. to allow outdoor canvassing work to be done; and (3) it provides sufficient time to collect and tabulate the data for reporting the results to the President by the end of the year ending in the digit "0."

SESSION 1: Provider or physician surveys

Chair: Ronald Andersen, Center for Health Administration Studies, University of Chicago

Recorder: Helen C. Gift, Bureau of Economic and Behavioral Research, American Dental Association

An evaluation of the reliability of data gathered from three primary care medical specialties using a self-administered log-diary*

Elizabeth B. Harkins, Health and Population Study Center, Battelle Human Affairs Research Centers

Introduction

As our research questions become more sophisticated, our needs for finer and finer bits of information frequently grow. These needs are often expressed in surveys that place considerable burden in time and effort on the responding subject. As was discussed at the Second Biennial Conference on Health Survey Research Methods in 1977, respondent burden can have several effects, including a lowered response rate, inaccurate reporting, and a high turnover among interviewers (Rothwell and Bridge, 1978). The study described here addressed the second of these potential problems: the reliability of data collected from physicians using a self-administered log-diary to describe their activities over a week.

The University of Southern California (USC) is engaged in gathering detailed data that describe the practices of a national sample of physicians representing approximately 24 medical specialties and subspecialties.[1] The data for their studies are gathered using a log-diary that requests detailed information on how a physician spends his/her time over a prespecified three-day period. Information on the patients who are seen or talked with over the telephone, as well as details concerning what was done during each patient encounter, is included. Summary data on the number of patient encounters and professional hours worked are also gathered for a full seven-day week.

This paper is based on a study of certain aspects of the quality (primarily reliability) of the data generated in the USC manpower studies of three medical specialties surveyed between August 1977 and May 1978.[2] Our analysis focused on describing the physician's experiences in using the USC log-diary, on estimating the reliability of physician, practice, and patient encounter characteristics, and on evaluating potential sources of unreliability, including both log-diary and respondent characteristics.

Methods

This study of the quality of the data gathered in the USC surveys utilized a comparison of the data collected by USC with those gathered in a follow-up survey through a telephone interview with the physician and a self-administered practice audit completed by a member of the physician's office staff. The study focused primarily on an assessment of the degree to which the same question elicited the same response on repeated occasions and addressed to a lesser degree the question of whether the instrument accurately reflected what it was intended to measure.

Sample selection. The universe to be sampled was defined as those physicians in the second of two USC samples of general practitioners, the first of two samples of family practitioners, and the first of two samples of pediatricians who responded to the USC survey within three months of the assigned data collection week. The two USC samples within each specialty were random replicates from the total sample of each specialty, so the follow-up study sampling frame was a random subsample of the total number of physicians responding to USC for each specialty. The population of respondents was stratified by (1) specialty, (2) promptness of the physician's response to USC, and (3) practice arrangement.

The three specialties were selected because they were the ones scheduled for survey by USC at the time that the reliability study was planned. The need to have the reliability study conducted promptly was felt to outweigh the desire for a sample of specialties that represented greater variation across levels of care and

*The author is indebted to Margaret Marini, Edward Perrin, and Malcolm Peterson for their contributions to this research and to earlier reports on which this paper is based.

other key characteristics. Inclusion, for example, of a surgical subspecialty was desirable, but not feasible, given USC's schedule of surveys.

The population of respondents was also classified into two groups on the basis of rapidity of response to the USC survey. Those returning the log-diary within four weeks after the first completed booklet was received were classified as "early responders." This group was distinguished from "late responders" who returned the booklet 5 to 13 weeks after the first completed booklet was received. It was expected that the booklets received substantially after the study week had probably been completed, at least in part, from records, notes, or memory. The particular distinction drawn between early and late responders was based on an examination of the response curve experienced by USC in prior studies of other specialities.

Finally, the population of respondents was stratified by practice arrangement into four categories: solo, partnership, group, and institutional and other arrangements. Differences in the conditions of practice across arrangements and in the characteristics of the physicians who chose to work in each arrangement might affect the quality of the responses provided. Since it was likely that institution-based practitioners would be unable to complete the practice-audit booklet, separation of this portion of the sample into a distinct stratum for the purposes of data collection was also desirable. Physicians in this stratum were interviewed only; no practice-audit booklet was requested for them.

A total sample of approximately 300 physician respondents, divided equally among the 24 cells of the $3 \times 2 \times 4$ design, was felt to represent the minimum size necessary to carry out an analysis of factors affecting the quality of the data and an approximate upper limit of the sample size that could be obtained given (1) an anticipated USC response rate of 60 percent or less and (2) an expected response to the follow-up survey of as low as 40 percent. The size of the sample to be drawn was inflated to compensate for the expected nonresponse. Sampling of respondents was carried out biweekly as the responses were received.

Data collection instruments. Two different and complementary data collection protocols were developed and employed for the reliability study. First, a telephone interview schedule was designed to gather information from all sampled physicians on (1) their experiences in completing the log-diary, including such information as when they filled out each part, who on their staff assisted with the completion of the booklet, how participation in the study affected their usual professional activities, and how they defined certain terms used frequently in the log-diary; and (2) certain characteristics of their practices, including their principal practice arrangement, primary specialty, and a profile of their office staff. The interview took about 15 minutes to complete and elicited only information that was readily available to the physician without consulting office or patient records.

Second, a practice-audit booklet was developed, to be completed by staff of the office-based practitioners in the sample, for reporting the number of patients seen during the USC study week and the characteristics of face-to-face patient encounters on one day during that week. This booklet was modeled after the log-diary with only minor variations designed to gather data in a slightly more disaggregated fashion and to omit requests for information not likely to be available in office records. The instructions and definitions provided were essentially identical to those in the log-diary. Only *one* day of detailed recording was required rather than the three days requested by USC. Some parts of the log-diary, including a 72-hour diary and the descriptions of telephone encounters, were excluded, since resources needed to provide those data were not likely to be available. The practice-audit booklet was to be completed from records by a member of the physician's staff who was familiar with the practice.

Physicians in solo, partnership, and group practice arrangements were offered up to $50.00 as partial reimbursement for staff time and other costs involved in completing the booklet. Approximately 40 percent of the physicians who completed the booklet requested reimbursement, almost always in the full amount.

Response rates. An interview or booklet was considered complete if it contained any usable data. A few bits of data were missing from many responses; occasionally only a few bits of data were present. The response rate for the booklet part of the reliability survey (48 percent) was lower than that for the telephone interview (84 percent). Many physicians were willing to spend 15–30 minutes of their own time on the telephone but were unwilling to allow a member of the office staff to obtain the requested information from office and patient records. Overall, 219 usable practice-audit booklets were returned and 374 interviews were completed.

Selected characteristics of the physicians sampled for the reliability study were explored for systematic differences between those who responded to the follow-up survey and those who

did not. Any such differences would be suggestive of possible biases in the findings of the reliability study and might also be indicative of respondent-nonrespondent differences in the USC data. No major differences were found between physicians responding and not responding to the practice-audit booklet or interview on hospital appointment level, practice location, faculty appointment, practice arrangement, and A.M.A.-designed specialty.

A potentially significant finding is that physicians responding to the reliability study reported fewer outpatient encounters during the assigned study week and on the one detailed recording day than did those not returning a booklet. The magnitude of this difference is 33 encounters per week and about 4 encounters on the detailed recording day. These differences could readily arise from the respondent burden in the reliability study being positively related to the number of patient encounters during the study week, particularly the number on the detailed recording day. Thus, a physician would find it easier to participate if he or she saw relatively few patients. A bias of this sort could result in an overall estimate of patient volume that is lower than the true patient load.

On average, physicians completing the practice-audit booklet were 48.9 years of age; those not completing the booklet averaged 51.0 years of age. These two groups do not differ statistically on age. The mean ages of physicians who were interviewed (48.3 years) and those who were not (51.7 years) differ significantly, however, since the difference is somewhat greater and the sample larger.

Finally, respondents to the booklet are significantly more likely than nonrespondents to be board certified in at least one specialty. The observed difference is rather small, however: 67 percent of respondents, compared with 58 percent of nonrespondents, are board certified. Although a similar trend is apparent between those who were interviewed and those who were not, the difference is not statistically significant.

Matching patient encounters between the data sets. In order to examine the reliability of variables measured for each patient encounter described in the log-diary, it was necessary to match patient encounters reported in the initial survey with those reported in the follow-up study. Encounters in the two surveys could not be paired directly because, in order to protect privacy and confidentiality, no meaningful identifying number or name was associated with the encounter descriptions in either of the studies. It was, therefore, necessary to match the encounters reported for a given physician in the two data sets on the basis of selected variables measured for each encounter.

Six variables were used to pair patient encounters: (1) age, (2) sex, (3) location where the patient was seen, (4) primary problem type, (5) focus of the primary problem, and (6) etiology of the primary problem. Matching was accomplished in a stepwise fashion, moving from greater to lesser precision in the amount of agreement required between two encounters to constitute a match.

The same nine-step computerized algorithm was used in matching encounters recorded by both general and family practitioners. The sequence of steps alternately allowed for greater tolerance in the amount of agreement on age and on primary problem type, focus, and etiology.[3] Although the algorithm used to match the encounters recorded by pediatricians employed the same variables as the one used for data from general and family practitioners, it was redefined into eight steps to allow for the more constrained age range of patients seen by pediatricians.

A manual matching stage based on the same six variables was carried out following the computerized sequence using the same procedures for all three specialties. The focus and etiology of secondary and tertiary problems, where they were recorded, and the order in which encounters were listed in the two data sets were also considered during this phase of the matching. This additional information sometimes provided a basis for choosing one of several multiple match possibilities or for allowing slightly greater disagreement on the six principal matching variables. In all cases a match was identified only if a unique pair of matching entries could be found. A total of 71 percent of the patient encounters reported in the follow-up survey could be matched with encounters reported in the initial survey.[4]

Analysis of the unmatched encounters indicated that although some nonmatches were due to missing data and some were due to the availability of multiple match possibilities, the overwhelming percentage of nonmatches in each data set was due to the inability to find any match for a given encounter on the basis of the selected variables. Since the criteria for matching were fairly conservative in order to avoid mismatching encounters and biasing estimates of reliability downward, it is not surprising that most of the nonmatches were encounters for which no close match could be identified.

It should be emphasized that the matching of patient encounters was a data reduction step in

the reliability study. The percentage of encounters matched should not be considered an estimate of the reliability or validity of the log-diary data. Although the reliability of the encounter characteristics did constrain the ability to match encounters, other factors unrelated to the quality of the data also affected the percentage of encounters that could be matched.

Analytic approach. The evaluation of the quality of the data from the log-diary was approached from both a descriptive and an analytic viewpoint. The physicians' responses to questions about their experiences in the log-diary were described in order to provide a more qualitative framework within which to review the quantitative estimates of reliability.

Where similar data were collected in both the initial and the follow-up surveys, a quantitative estimate of reliability, i.e., agreement between the two reports, was made. Where the characteristic of interest was assessed in terms of discrete, nominal categories, Cohen's (1960) kappa and/or Meltzer and Hochstim's (1970) index of reliability were used to measure agreement between the USC and Battelle data. Kappa is defined as the proportion of agreement observed after chance agreement is removed from consideration.[5] The index of reliability is defined as the observed agreement as a proportion of the *maximum observable agreement* after chance is removed from consideration.[6] Kappa was used as a measure of reliability for all nominal variables; the index of reliability was also computed where the data were supplied in the two surveys by different types of individuals or where records were used as the information source in one study and concurrent experience in the other.

In order to understand and describe the degree of agreement between any two ratings for a continuous variable as fully as possible, both Spearman's rho and the bivariate regression coefficient between the two ratings were estimated. Since perfect agreement between the data sets was expected on theoretical grounds, the regression analyses were performed using an intercept constrained to zero. Where perfect agreement exists, not only should the correlation coefficient be one but also the intercept of the best-fitting regression line should be zero and the slope should be one.

Krippendorff (1970) has demonstrated that kappa and rho belong to the same family of coefficients. Chance agreement is estimated in the same way in both measures; rho differs from kappa in the manner in which deviations from agreement are weighted. Kappa weights all deviations equally; rho weights deviations more heavily as they become larger (specifically, weights increase with the square of the deviation).

In evaluating reliability, one usually demands a correlation coefficient of at least .6 to .7, with reliabilities of .8 and above being clearly preferred (McKennell, 1970). The measures of agreement for both unordered and ordered variables are closely related and, in fact, take on identical values under a number of conditions. The following labels have been suggested to describe the strength of association indicated by the corresponding levels of kappa: less than .00, poor; .00–.20, slight; .21–.40, fair; .41–.60, moderate; .61–.80, substantial; and .81–1.00, almost perfect (Landis and Koch, 1977). When one is concerned with reliability, it is important to have substantial to almost perfect agreement.

Results

The results of the evaluation of the quality of the data collected using the log-diary are presented in four parts. First, a description of the physician's experience in using the log-diary is given. Second, the reliabilities of physician and practice characteristics measured in the log-diary are explored. Next, the characteristics of the patient encounters described by the physicians in the log-diary are evaluated. Finally, the effects of various potential sources of error on the observed reliabilities are investigated.

Physicians' experience with the log-diary. The manner in which the log-diary was filled out, specifically whether the physician received help in filling out various sections and when each section was completed, could have a substantial effect on the reliability of the data reported. If the data were recorded in the log-diary long after encounters took place and if other staff in the practice completed large parts of the diary, the reliance on records for information is likely to have been greater than in instances where the physician provided the data as the patients were encountered. Overall, physicians received the greatest amount of help with the sections of the log-diary devoted to description and enumeration of patient encounters; about 34 percent of physicians who responded to the reliability study received help with these sections. A somewhat smaller percentage, 21 percent, received help with the 72-hour diary; and 19 percent received help with the questionnaire about practice characteristics. Both of these sections were devoted to describing the physician's activities and characteristics.

The reliability study also provided information about when each section of the booklet was

completed. Slightly more than half of the physicians indicated that they filled out each section as things happened each day, and about half indicated that they usually filled out each section at the end of the day. Only about 5–8 percent of the physicians indicated that they completed any part of the log-diary after the day in which the activity took place.

We also examined the consistency with which key terms used in the description of patient encounters were interpreted by physicians responding to the log-diary. The consistency in use of the terms "inpatient" versus "outpatient" was examined for patients seen in emergency rooms and nursing homes, since these locations appear to be less clearly defined on the in/outpatient dimension than other settings. Of those responding to the reliability study, 81–91 percent said that they classified patients seen in the emergency room as outpatients. In contrast, there was considerable disagreement about the classification of nursing home patients. Almost half of the general and family practitioners classified them as inpatients, and a little over half classified them as outpatients. These differences in the classification of nursing home patients would likely affect the reliability of estimates of the relative division of a physician's practice between inpatient and outpatient settings to the extent that nursing home patients represented a significant proportion of encounters.

Physicians were also asked to note whether they provided the majority of care for each patient whom they encountered during the detailed recording days of the study week. Of six definitions presented to each physician responding to the telephone interview, there was a high degree of agreement on one definition of "majority of care." Almost all physicians (93 percent) indicated that providing the majority of a patient's care meant providing "all physical and emotional care—both therapeutic and preventive." The remaining 7 percent of physicians indicated that it meant providing "all physical care—both preventive and therapeutic."

The physician's definition of a regular patient was also examined. Five attributes that might define a regular patient were presented to the physicians in the telephone interview; four were identified by the physicians as important in classifying a patient as "regular." In order of importance, these were (1) whether the physician expected to provide all or most of the patient's care, (2) the number of times that the patient had been seen, (3) the length of time since the last visit, and (4) the length of time that the physician had known the patient. One of three combinations of these definitional attributes was offered by 61 percent of physicians. The expectation that he/she would provide all or most of the care for the patient was common to each of the three most-preferred definitions; the number of prior visits was included in two of the definitions; and the duration since the patient's last visit was noted in one of these three definitions.

Although there was substantial agreement on the definition of a regular patient (61 percent of physicians constructed three rather similar definitions), there was also considerable disagreement (39 percent of the physicians together constructed over 10 different definitions). These discrepancies would give rise to interrater unreliability, which would make analytic use of this patient-encounter characteristic problematic.

It should also be noted that the definitions of a regular patient and majority of care are in many respects quite similar. The physicians indicated that providing the majority of a patient's care is providing *all* their care and that a regular patient is one for whom he/she expects to provide all or most of the care. The Spearman correlation in the data provided in the log-diary between regular patient and majority of care is .82 over 2,316 encounters.

Reliability of physician characteristics. The reliability of two major types of variables was examined: (1) variables for which the physician was the unit of analysis and (2) variables for which the encounter was the unit of analysis. Key physician-level variables were primary specialty and practice arrangement, the number of patients seen, the number of professional hours spent each day of the week, and the composition of the office staff.

The same primary specialty was reported by 82 percent of the physicians in both the log-diary and the follow-up interview. Not surprisingly, the level of agreement, adjusted for chance, observed in the data is substantial (73 percent). There appears, however, to be one focus of unreliability in the data. Of the 69 physicians who reported initially that they were general practitioners, nearly one-third reported in the follow-up interview that they were family practitioners.

Table 1 contains a cross-tabulation of primary practice arrangement as reported in the log-diary and follow-up interview. A substantial 65 percent agreement is observed after chance agreement is removed. A third of the total deviation from perfect agreement could be identified as arising from shifts between the partnership and group categories. Obscuring

Table 1
Cross-tabulation of principal practice arrangement as reported in log-diary and follow-up interview

Practice arrangement and follow-up interview	Log-diary (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	Total
Solo (1)	95	3	1	0	1	1	0	0	0	101
Partnership (2)	4	63	6	2	0	0	1	0	2	78
Single specialty group (3)	1	13	36	0	0	1	0	0	1	52
Multispecialty group (4)	0	10	3	34	0	0	3	0	1	51
Academic department (5)	0	0	0	0	11	0	0	0	2	13
Hospital (6)	0	0	0	1	6	11	1	0	1	20
Clinic (7)	1	2	0	3	2	1	4	4	2	19
Government (8)	0	0	1	0	0	1	0	1	0	3
Other (9)	7	2	5	2	2	0	2	2	5	27
Total	108	93	52	42	22	15	11	7	14	364

Observed proportion of agreement 714
Agreement expected by chance 183
Proportion of agreement after chance is removed: Kappa ... 650

the distinctions among these three categories, the level of agreement beyond chance is 71 percent.

Table 2 contains the results of an analysis of the number of patient encounters per day and the number of professional hours spent by the physician each day during the study week.[7] The results indicate that reports of the number of outpatients encounters are highly reliable. With the exception of the data for Sunday, the reliability is in the vicinity of .90. Whether the degree of agreement is measured by r, r_s, or the slope of the best-fitting zero-intercept regression line, the number of outpatient encounters reported in the follow-up survey agrees closely with the number of encounters reported in the log-diary. The reports for Sunday appear to be in fair agreement. Greater inaccuracies in office records of Sunday encounters may contribute to the lower agreement observed for that day.

Reports of the number of professional hours spent on each day of the week were less reliable than reports of the number of encounters. As estimated by r_s, the agreement for Tuesday

Table 2
Reliability of number of outpatients and professional hours

Day of week	Intercept[a]	Slope	Standard error of slope	Standard error of estimate	r	r_s	Booklet Mean	Booklet S.D.	Log-diary Mean	Log-diary S.D.	N
I. Number of Outpatients Seen											
Sunday	0	.552	.035	1.45	.769	.636	0.74	2.39	1.16	3.52	123
Monday	0	.934	.023	6.82	.899	.916	21.64	15.47	22.76	15.48	119
Tuesday	0	.986	.024	6.66	.914	.929	19.20	16.37	19.37	15.32	123
Wednesday	0	1.003	.020	5.08	.952	.952	16.83	16.57	16.71	15.81	117
Thursday	0	.950	.037	8.32	.838	.883	15.15	15.13	15.20	14.16	115
Friday	0	.980	.022	5.76	.932	.932	19.32	15.42	18.78	15.79	116
Saturday	0	.912	.022	2.79	.955	.943	5.96	9.37	6.36	9.93	112
II. Number of Professional Hours Spent											
Sunday	0	.413	.051	1.56	.473	.586	1.15	1.67	1.77	2.61	96
Monday	0	.890	.027	2.78	.672	.517	8.53	3.57	9.23	3.60	107
Tuesday	0	.881	.026	2.52	.744	.663	8.00	3.53	8.68	3.90	104
Wednesday	0	.919	.032	2.71	.780	.782	7.08	4.29	7.51	4.01	101
Thursday	0	.866	.031	2.57	.748	.755	6.66	3.72	7.31	3.93	103
Friday	0	.873	.029	2.65	.678	.595	7.60	3.38	8.28	3.60	103
Saturday	0	.791	.064	3.02	.602	.724	3.23	3.68	3.49	3.40	95

[a]Intercept constrained to zero.

through Thursday and Saturday appears substantial. Reliabilities on the other three days of the week are only moderate. Some of the error in professional hours arises from a systematic difference between the two surveys of a little over a half-hour a day, with the physicians reporting the larger number in the log-diary. This difference may arise from the need in the reliability study to rely on records for data. A physician's office is not likely to keep accurate information on the total number of professional hours spent by the physician each day.

Three sources of information were available with which to estimate the reliability of the number of each of several types of personnel employed in the physician's office. Both the log-diary and the practice-audit booklet contained a grid in which the respondent noted the number of persons employed in each category of staff personnel at each of five levels of work time. In addition, the physician reported in the telephone interview the total number of persons in each category of employee.

The grid format in which the data on staff profile were collected presented major problems in differentiating implicit zeros from missing data. No specific request was made to record zeros where appropriate. Initially, the data were analyzed using an exclusion rule for missing data that required that a case be excluded only if the entire grid was blank. The total number of staff members in each category of personnel, collapsing distinctions between full- and part-time employees, was compared among the three sources of data.

The left half of Table 3 presents the results of this analysis. As is readily apparent, agreement is quite variable, ranging from slight to substantial. The agreement on number of physicians in the practice, for example, is only moderate. Only 6 of the 36 correlations for the 12 categories of personnel reach the .60 level of agreement usually expected of reliable indicators. Further examination of the data suggested that no sizable amount of the error could be explained by specialty, practice arrangement, provider of the information (i.e., physician or office staff), or delay in recording of the data. Rather, it appeared that the confounding of zeros with missing data was the basis of the problem.

The data were reanalyzed using a more restrictive exclusion rule to eliminate missing data. Any row within the office profile grid that was blank was considered missing in the second analysis, and, for consistency, known zeros were excluded from the responses to the physician

Table 3
Reliability of office staff profile by type of personnel and treatment of blank data (Spearman correlations)

Type of personnel	If all variables blank, case is omitted			Any blank variable is omitted		
	Booklet/ log-dairy	Interview/ log-diary	Interview/ booklet	Booklet/ log-diary	Interview/ log-diary	Interview/ booklet
Physician	.496 (199)[a]	.579 (316)	.542 (192)	.651 (114)	.714 (203)	.618 (169)
MEDEX	.205 (199)	.437 (310)	.247 (188)	— (2)	— (4)	— (1)
Nurse practitioner	.467 (199)	.553 (309)	.450 (187)	— (8)	.817 (19)	— (12)
Physician's assistant	.165 (199)	.217 (309)	.452 (187)	— (5)	.673 (14)	.786 (10)
Assistant trained in practice	.182 (199)	.300 (307)	.339 (186)	.485 (29)	.619 (42)	.588 (22)
Registered nurse	.577 (199)	.668 (305)	.819 (183)	.581 (80)	.738 (131)	.754 (97)
Lab or x-ray technician	.667 (199)	.674 (306)	.702 (185)	.792 (52)	.844 (96)	.678 (65)
Surgical technician	.181 (199)	.075 (306)	— (185)	— (3)	— (1)	— (0)
LPN or LVN	.584 (199)	.560 (305)	.750 (184)	.666 (47)	.604 (83)	.416 (64)
Community health aide	.032 (199)	.053 (306)	.026 (186)	— (3)	— (8)	— (1)
Secretary/receptionist	.519 (199)	.478 (299)	.550 (181)	.683 (132)	.609 (228)	.653 (157)
Other	.143 (199)	.177 (309)	.127 (187)	.311 (13)	−.072 (34)	−.297 (12)

[a]Numbers in parentheses are base numbers for correlations.

interview. The results from this analysis are given in the right half of Table 3.

Although the base n's for each correlation drop substantially, the correlations appear much higher. The magnitude of the missing data problem is immediately obvious from the size of the n's. For example, only 114 of the 219 physicians who completed both the log-diary and a practice-audit booklet reported any physicians in their practice; only 127 of the same respondents reported having either a registered or a licensed nurse in their office. Using the more restrictive exclusion rule, 17 of the 24 calculable coefficients are, however, .60 or greater.

If one is willing to collapse the data across all types of nonphysician personnel and use only an enumeration of total office staff, the correlations among the three sources of data range from .70 to .71, using the less restrictive missing data exclusion in which blanks are inferred to be zeros unless the entire grid is blank. This level of reliability is within the usually accepted range, and the missing data rule is unlikely to exclude many true zeros, thereby producing little upward bias in the estimates. A more inclusive collapsing, including physicians, yielded a reliability of similar level (.70–.74). Less inclusive collapsing was generally ineffective in raising the reliabilities of the estimates.

Reliability of encounter characteristics. In this section the results of analyses of the face-to-face patient encounters are presented. The units of analysis are matched patient encounters, i.e., the subset of encounters described in the log-diary and in the practice-audit booklet that have been determined to be descriptions of the same patient encounters.

In a preliminary analysis, three subsets of encounters—(1) those encounters matched in steps one through six of the computer matching algorithm, (2) those matched in the full sequence of the computer algorithm, and (3) the entire complement of computer and manually matched encounters—were analyzed separately to investigate the effects of the possible inclusion of mismatched encounters on the estimates of reliability. The criteria for a match were more restrictive in the early stages of computer matching than in the latter stages; therefore, the probability that some mismatches were included among the true matches increased as one moves successively through the algorithm and hand matching stages. The values of kappa across the three levels of matching were very stable; they decreased an average of only .04 from the most restrictive part of the computer algorithm to the total number of matches. For this reason, it appeared unlikely that any substantial number of mismatches were included.

In another preliminary analysis, the magnitude of the interrater effect on reliability from different individuals completing the log-diary and the practice-audit booklet was examined. In general, the physician completed the log-diary and a member of the office staff responded to the reliability study. There were, however, 21 office staff persons who asserted that they had completed 100 percent of the patient encounter section of the log-diary. The reliabilities of the encounter characteristics reported by this group of respondents were estimated separately and compared with the reliabilities based on all respondents. On average, the intrarater reliabilities as measured by kappa were approximately .09 above those based on all respondents.

This difference could arise from two sources. First, interrater differences can come from systematic differences in the way individuals with different professional backgrounds complete the instruments. Some variation occurred in who within each practice completed the log-diary; generally, there were differences in the background and training between respondents to the log-diary and respondents to the practice-audit booklet. If systematic differences occurred between physicians and office staff, there would be some unreliability in comparisons between the log-diary and practice-audit booklet arising from the reliability study design. Second, individuals are usually more consistent with themselves than with others of any background or training. A difference of about 9 percent agreement could easily occur from this source alone, even if all respondents in both studies were physicians. This source of unreliability is inherent in studies using self-administered data collection protocols, since a different person (even several different persons) completes the instrument in each physician's practice.

The characteristics of each patient encounter recorded in the log-diary included whether the patient had been seen before, whether the patient was a regular one, whether the physician provided the majority of that patient's care and the care to his/her family, the source of the patient referral, the physician's role vis-à-vis other providers, therapeutic and diagnostic procedures requested or performed, the disposition of the encounter, a description of the patient's problem, and a coding of that problem in three areas—type, focus, and etiology.

The majority of responses (85–98 percent) at both data collection times indicated that the patients had been seen before, were regular patients, were self-referred, and were referred to

no further provider after the encounter in question and that the physician provided the majority of their care and was providing care rather than assisting another provider. Net of chance, only 10–60 percent agreement was observed between the two observations for these encounter characteristics. These skewed distributions produce levels of chance agreement of between 76 and 95 percent.

The responses to whether the physician provided the majority of care to the patient's family were more evenly distributed across the various categories than the responses to the other encounter variables considered so far. Chance agreement was estimated at only 54 percent. Net of chance, however, only 46 percent agreement was observed.

Analysis of the item on the physician's next contact with the patient revealed an observed agreement after chance of 38 percent. On this item, 45 percent of the disagreement appeared to be among 3 of the 14 response alternatives: "no next contact," "call back as needed," and "return as needed." Approximately 54 percent agreement is observed after chance when the distinctions among "none," "phone," and "return as needed" are ignored.

Validating information on whether each patient had been seen before was gathered in the practice-audit booklet. For each encounter on the detailed recording day, the respondent was asked to note the dates of the three visits immediately prior to the one being described. Blanks were recorded as missing data and excluded from analysis; the encounters with no prior visit are those for which the respondent actually wrote in "none."

The agreement within the data provided in the follow-up survey is high—85 percent after chance is removed—but not as high as one might expect. In 65 of 2,297 encounters, the respondent noted at least one prior visit date and then recorded that the patient had not been seen before, or the respondent wrote in "none" for prior visit dates and then recorded that the patient had been seen before. The agreement between the recording of prior visit dates in the practice-audit booklet with whether the patient had been seen before according to the log-diary is substantially lower than the agreement within the practice-audit booklet. Only 60 percent agreement is observed, not of chance. It should be noted that the reliability estimate is essentially equal to the validity estimate for this encounter characteristic.

Therapeutic and diagnostic procedures. Analysis of the reliability of the therapeutic and diagnostic procedures was limited to those procedures that were recorded in the log-diary with relatively high frequency. The respondents were presented with a list of 35 to 40 therapeutic procedures and 35 to 40 diagnostic procedures and asked to note the code numbers of those procedures requested or performed during each encounter. Table 4 presents the cross-tabulations of the eight therapeutic procedures that were noted in at least 5 percent of the encounters: immunization, other injection, patient education, listening/reassurance, systemic drugs, topical drugs, exercise/diet, and treatment program counseling.

It is readily apparent that, with the exception of immunizations, there is little agreement in the reporting of these procedures. The probability that inadequate notations existed in the records used as the source of information for the reliability study with regard to patient education, listening/reassurance, exercise/diet, and treatment program counseling is sufficient to justify discounting somewhat the estimates for these treatments. This methodological problem is less useful, however, in explaining the lack of agreement in the report of drugs and injections. It would seem that these treatments are likely to be recorded in the office record as immunizations, which demonstrate an agreement level at least twice that of the drug and other injection procedures. An examination of the data suggests that there may be a substitution between the drug codes and the injection code. In 42 percent of the encounters where an injection is recorded, no drug code is noted.

Because of the infrequent recording of diagnostic procedures, those reported in 4 percent or more of the encounters were analyzed. Table 5 presents the cross-tabulations of the five procedures that met this criterion: routine lab, blood chemistry, culture, chest x-ray, and other radiology. The reliability levels of the routine lab, culture, and other radiological procedures are moderate, ranging from .52 to .58. Blood chemistry and chest x-ray, however, demonstrate agreement at only 26 to 32 percent, after chance. No combinations of codes of substitution effects were found that could help to explain the levels of reliability observed. Notations in office records of lab procedures and x-rays would be expected to be fairly complete. Therefore, differences due to the varying source of information for the two studies should not be great. Moreover, such differences, if they did exist in significant degree, would not explain why certain procedures (e.g., chest films) are recorded with only fair agreement and others (e.g., other radiology) with moderate agreement.

Table 4
Cross-tabulations of selected therapeutic procedures from the log-diary and practice-audit booklet

	Log-diary											
	Immunizations			Other injection			Patient education			Listening/ reassurance		
Booklet	Not coded	Coded	Total	Not coded	Coded	Total	Not coded	Coded	Total	Not coded	Coded	Total
Not coded	2,103	57	2,160	1,860	56	1,916	2,112	158	2,270	2,082	150	2,232
Coded	61	117	178	320	102	422	49	19	68	58	48	106
Total	2,164	174	2,338	2,180	158	2,338	2,161	177	2,338	2,140	198	2,338
Observed proportion of agreement			.950			.839			.912			.911
Agreement expected by chance			.861			.776			.900			.878
Proportion of agreement after chance is removed:												
Kappa			.638			.281			.118			.273
Index of Reliability			.645			.568			.221			.401

	Systemic drugs			Topical drugs			Exercise/diet			Treatment program counseling		
	Not coded	Coded	Total	Not coded	Coded	Total	Not coded	Coded	Total	Not coded	Coded	Total
Not coded	1,054	523	1,577	2,175	89	2,264	2,117	138	2,255	2,089	200	2,289
Coded	241	520	761	39	35	74	69	14	83	44	5	49
Total	1,295	1,043	2,338	2,214	124	2,338	2,186	152	2,338	2,133	205	2,338
Observed proportion of agreement			.673			.945			.912			.896
Agreement expected by chance			.519			.919			.904			.895
Proportion of agreement after chance is removed:												
Kappa			.321			.327			.078			.006
Index of Reliability			.428			.444			.112			.016

Diagnoses. After receipt of the data, the problem descriptions recorded in the log-diary and practice-audit booklet were converted into standard ICDA codes. Seven relatively frequently used diagnoses were selected and examined for reliability using the three-digit ICDA code. The cross-tabulations for these diagnoses—obesity (ICDA 277), otitis media (381), essential hypertension (401), acute pharyngitis (462), acute upper respiratory infection (465), medical examination (Y00), and prenatal examination (Y06)—are presented in Table 6. The reliability of four of the five diagnosed conditions is between .55 and .65; the reliability of the report of upper respiratory infections is only fair (.39); and the agreement in the two "well" conditions is almost perfect (.84–.90).

An analysis of the uncoded information written in the log-diary and practice-audit booklet for a sample of the encounters in the inconsistent cells of the cross-tabulations was conducted to illuminate the sources of the apparent errors. A sample of the inconsistent cells across the seven cross-tabulations was randomly drawn and the booklets reviewed. Three sources of discrepancies were identified: (1) incorrect coding of the information provided, (2) mis-

Table 5
Cross-tabulations of selected diagnostic procedures from the log-diary and practice-audit booklet

Booklet	Log-diary: Routine lab			Blood chemistry			Culture			Chest x-ray			Other radiology		
	Not coded	Coded	Total	Not coded	Coded	Total	Not coded	Coded	Total	Not coded	Coded	Total	Not coded	Coded	Total
Not coded	1,707	166	1,873	2,190	83	2,273	2,117	85	2,202	2,205	58	2,263	2,222	52	2,274
Coded	163	302	465	40	25	65	39	97	136	47	28	75	22	42	64
Total	1,870	468	2,338	2,230	108	2,338	2,156	182	2,338	2,252	86	2,338	2,244	94	2,338
Observed proportion of agreement			.859			.947			.947			.955			.968
Agreement expected by chance			.681			.929			.873			.934			.935
Proportion of agreement after chance is removed:															
Kappa			.560			.264			.582			.325			.516
Index of Reliability			.562			.355			.690			.350			.641

Table 6
Cross-tabulations of selected diagnoses from the log-diary and practice-audit booklet

Booklet	Log-diary: Obesity (277)[a] Not coded	Coded	Total	Otitis media (381) Not coded	Coded	Total	Hypertension (401) Not coded	Coded	Total	Pharyngitis (462) Not coded	Coded	Total
Not coded	2,284	22	2,306	2,137	43	2,230	2,144	50	2,194	2,198	36	2,234
Coded	11	21	32	33	75	108	55	89	144	98	56	104
Total	2,295	43	2,338	2,220	118	2,338	2,199	139	2,338	2,246	92	2,388
Observed proportion of agreement			.986			.986			.955			.964
Agreement expected by chance			.968			.908			.886			.920
Proportion of agreement after chance is removed:												
Kappa			.553			.647			.805			.553
Index of Reliability			.650			.678			.617			.591

Booklet	Upper respiratory infection (465) Not coded	Coded	Total	Medical exam (Y00) Not coded	Coded	Total	Prenatal exam (Y06) Not coded	Coded	Total
Not coded	2,134	42	2,176	1,851	62	1,913	2,271	5	2,276
Coded	107	55	162	50	375	425	7	55	62
Total	2,241	97	2,338	1,901	437	2,338	2,278	60	2,338
Observed proportion of agreement			.936			.952			.995
Agreement expected by chance			.895			.699			.949
Proportion of agreement after chance is removed:									
Kappa			.393			.841			.899
Index of Reliability			.535			.855			.915

[a] Cross-tabulations were based on the 3-digit ICDA code.

matched encounter pairs arising from the encounter-matching procedures, and (3) differences in the information recorded in the log-diary and booklet. Of the encounter pairs examined, the discrepancy was due to a coding error 14 percent of the time and an apparent mismatch of encounters 12 percent of the time. The majority of the inconsistent encounter pairs (74 percent) were due to differences between what had been recorded in the log-diary and in the practice-audit booklet.

Further analysis of the differences in the recorded information revealed a couple of discrepancy patterns. One frequently occurring example of inconsistency arose from describing the purpose of the visit (e.g., well-child exam) rather than the findings from the visit (e.g., otitis media). Of the inconsistencies due to recording differences, 18 percent appeared caused by variation in coding either the purpose of or the findings from a visit. Another and more pervasive problem involved the distinctions among similar diagnoses, e.g., pharyngitis, upper respiratory infection, and, to a lesser ex-

tent, otitis media. Of the recording differences that were related to at least one of these three diagnoses, 47 percent appeared to involve a substitution of one of the diagnoses for another of the three. These two problems represented 43 percent of the total number of recording differences observed in this sample of inconsistencies.

Primary problem, type, focus, and etiology. In addition to the open-format problem description, each encounter recorded in the log-diary was coded using three closed-format variables to describe (1) the problem type, e.g., medical, surgical, obstetrical; (3) the focus of the problem in terms of body systems; and (3) the etiology of the problem in terms of basic pathological processes. These variables were used in the patient encounter-matching algorithm. The estimates of their reliability are, therefore, directly constrained by the computer algorithm used to pair encounters from the two data sets and can be interpreted only as upper limits on their "true" reliability. Because of the clustering of about 70 percent of the encounters in the "medical" category of problem type, the agreement expected by chance alone is about 52 percent. Net of this, 62 percent agreement is observed as a likely upper limit on the reliability of problem type. There is less skewness in the use of the focus and etiology categories, resulting in relatively low (7 –12 percent) levels of chance agreement. Only 48 percent agreement after chance is observed in etiology, however, and 63 percent agreement in focus.

Effects on reliability of specialty, response rapidity, patient load, and board certification. The final phase of the analysis focused on an examination of several potential sources of unreliability in the log-diary. Four sources were explored: (1) specialty, (2) delay in completion and/or return of the log-diary, (3) patient load, and (4) board certification.

Four physician characteristics and three encounter characteristics were selected to represent variables of higher and lower reliability and to represent characteristics that are more or less based on judgment. An evaluation was conducted of the effects that the potential sources of error had on these variables: number of outpatients on Sunday and on Monday, number of professional hours on Monday and Wednesday, whether the patient was a regular patient or had been seen before, and what next contact was planned. No systematic differences in agreement in the selected variables were observed among the three specialities, by when the log-diary was completed (i.e., during or after the assigned study week) or was returned to USC (i.e., during the first six weeks of responses or not), by patient load, or by board certification. There were no clearly discernible sources of unreliability in these characteristics of the physicians who responded to the studies or in the circumstances under which they completed the log-diary.

Adjustments to reliability estimates. Several analyses were carried out to estimate the effects on the observed reliabilities of the procedures designed for matching encounter descriptions. An average difference of only .04 in the kappa for several encounter characteristics was observed between those encounter pairs generated by the most restrictive part of the matching algorithm and the entire set of encounter pairs defined by the computer and manual matching steps. Moreover, when specific cases of inconsistency in diagnoses were investigated, only 12 percent of the disagreements were apparently due to mismatched encounter pairs. If one assumes (1) that mismatched encounters represent approximately 12 percent of the nonagreeing encounter pairs, (2) that correctly matched encounters would yield *perfect* agreement, and (3) that the marginal distributions would remain as observed, kappa rises an average of approximately .07 over several encounter characteristics studied. Since the assumption that correctly matched encounters would always demonstrate perfect agreement is unreasonable, it appears appropriate to estimate that the kappas based on matched encounter pairs may be depressed by about .04 as found in the first analysis described above.

A second feature of the reliability study design that might produce encounter characteristic reliability estimates that are biased downward is the difference in the training and experience of the individuals who completed the log-diary compared with those who completed the practice-audit booklet. The physician was the primary respondent to the practice-audit booklet. This difference involves variations both in individual respondent and in type of respondent across the two studies. The log-diary survey involves similar differences both within and across practices, but the comparison between the initial and follow-up surveys involves a greater number of those differences. Although we cannot directly estimate the effect of the reliability study design, we can estimate what the reliabilities might be if the same individual provided all the data. A comparison of those encounter pairs where the same person provided the data at both times with the total

number of encounter pairs yields an average difference in the kappa across several encounter characteristics of .09. This difference, by definition, must be greater than the effect of the study design on the observed reliabilities.

Taking these two estimates of biasing effects due to the study design into consideration, we propose that the reliabilities of the encounter characteristics as measured by kappa are possibly between .05 and .10 lower than the true reliabilities. This adjustment factor is somewhat arbitrary and crude, but it represents our best estimate of an easy-to-apply correction for methodological artifacts.

Discussion

Physicians' experience with the log-diary. We explored the physicians' definitions of certain encounter characteristics and the conditions under which they completed the log-diary. A grey area in the distinction between inpatients and outpatients is the nursing home patient. There was considerable disagreement on how to categorize these patients. Through detailed analysis of the components of the definition of a "regular patient," it becomes clear that a regular patient is, above all, one for whom the physician expects to provide all or most of the care. Although there is reasonable agreement on this definition, it presents a problem because among the primary care specialists studied, it is very similar to majority of care—a theoretically independent construct. Finally, analysis of the conditions under which the physicians completed the log-diary suggests that it was filled out in much the way that was intended. Overwhelmingly, respondents recorded the information as things happened or at least by the end of the day.

Reliability of physician and encounter characteristics. Not surprisingly, the observed reliabilities of the physician's practice arrangement and primary specialty were above the usually accepted lower limit of .60. In fact, the majority of disagreement on specialty arises from the lack of a clear distinction between general and family practice—a problem that is not likely to occur between most specialty designations. Practice arrangement, however, contains an inconsistency in the distinction between partnership and group arrangements that is likely to be fairly pervasive in medical practice.

The number of patient encounters appeared to be reliably reported. A source of bias in the estimates of number of patient encounters arises, however, from a differential response rate by patient load. The burden in responding to the study was highly associated with the number of patients encountered; the greater the number of patients, the more likely the physician was not to respond. Although the estimates of the number of professional hours are somewhat less reliable than the estimates of the number of outpatients, they are certainly acceptable.

The data on individual categories of office staff personnel are quite unreliable owing to a major confounding between implicit zeros and missing data. Moreover, because of the problem with inferring zeros from blanks in the log-diary, the data are systematically biased—downward if one interprets blanks as zeros, and upward if one excludes blanks as missing data. The problem in these data should remind us to be vigilant in designing self-administered instruments to ensure that true zeros can be discriminated from nonresponses.

Employing the estimated correction factor for reliability study design effects discussed above and the usually accepted lower limits of reliability (.60–.70), we can discuss the findings for the various encounter characteristics studied. Of the eight characteristics describing the physician's prior and planned future contact with the patient, we would estimate that the report of whether the patient had been seen before is reasonably reliable (.65–.70) and the report of whether the patient was a regular one is at the lower limits of usually acceptable reliability (.57–.62). The physician's next contact with the patient is an acceptably reliable variable (.59–.64), if one does not attempt to distinguish among the categories for no next contact, phone contact, and return as needed. The remaining four characteristics have reliabilities that fall below .60 even after allowing a correction for study design effects.

The problems with these characteristics seem to arise from the nature of the judgments being requested and the adequacy of medical record keeping. The physicians define regular patients much as they define those for whom they provide the majority of care. The source of the initial referral for the patient is likely to be poorly documented and is subject to all the usual problems of recall. The notion of assisting in the care process rather than providing is very poorly separated.

Of the therapeutic procedures studied, only immunizations reflect an acceptable to good reliability (.69–.74). The estimates for several of the procedures that are not likely to be clearly noted in the patient's record should be weighted carefully, since the reliability study depended on office records as the major source of infor-

mation on encounters. This consideration would not, however, appear adequate in explaining the less than acceptable reliabilities of reports of other injections or systemic and topical drugs, all of which are below .45.

Among the diagnostic procedures examined, several appear to have reliabilities at the lower range of that which is usually considered acceptable; routine lab work, culture, and other radiology demonstrate reliabilities in the .60 to .65 range. Somewhat surprisingly, given the reliabilities of these procedures, blood chemistry and chest x-ray reveal reliabilities below .45. Moreover, record-keeping differences are not a likely explanation for these findings.

Of the diagnoses examined, the well-patient codes, e.g., medical and prenatal examinations, demonstrate high reliabilities of greater than .80 without any correction for study design problems. A special correction factor was, however, derived from the detailed analysis of the sources of inconsistencies in the diagnosis codes. The correction factor assumes (1) that 18 percent of the observed error in diagnoses is due to mismatched encounters and coding errors from the follow-up study, (2) that perfect agreement would have been found in the absence of these design effects, and (3) that the marginals would remain unchanged. Applying this adjustment to the data on the selected diagnoses, we find that there is substantial to almost perfect agreement for all diagnoses, save upper respiratory infection, which demonstrates a moderate level of agreement, .50.

Finally, the upper limits of the reliabilities of the primary problem type, focus, and etiology were examined. Primary problem type and focus appear within the acceptable limits of reliability with kappas of just over .60. Problem etiology, however, demonstrates a reliability of less than .50.

Conclusion

Overall, we conclude from our studies of the data elicited by the self-administered log-diary that the characteristics of the respondents have much less of an effect on the quality of the data that they provide than do the characteristics of the data requested and the formatting of the questions presented to the respondents. The data are quite good where there is a need simply to enumerate something, such as hours or patients. The data are also quite reliable if the respondent is asked to record a familiar piece of information in a familiar manner, as in an open-format description of presenting problems. However, even simple enumerations become unreliable where the questions are formatted in such a way that zeros are not discriminated from blanks. The data are less reliable in cases that require recording even familiar information, such as presenting problems, in an unfamiliar manner, e.g., as separate codes for type, focus, and etiology. Even frequently utilized concepts such as inpatient versus outpatient, practice arrangement, specialty, and regular patient have foci of unreliable discriminations. Finally, it appears that those physicians who completed the log-diary did so in much the way that was intended; about half reported that they recorded the information requested as things happened. About half, however, recorded at the end of the day. Participation in the study was affected by the degree of burden inherent in responding.

Footnotes

[1] These studies are part of the Medical Activities and Manpower Projects supported by the Robert Wood Johnson Foundation and the Health Resources Administration, DHEW. See Mendenhall, Girard, and Abrahamson (1978) and Mendenhall, Lloyd, Repicky, Monson, Girard, and Abrahamson (1978) for more detail.

[2] The study was undertaken by the Health and Population Study Center of Battelle Memorial Institute, Seattle, with support under subcontract (M868565) from the University of Southern California and contract (PLD0595078) from the Health Resources Administration, DHEW. See Perrin, Harkins, and Marini (1978) for more detail.

[3] The specific criteria on which agreement was needed to constitute a match are described in the final report on the follow-up study (Perrin et al., 1978).

[4] A smaller fraction of encounters were matched in the data gathered by USC because physicians reported more encounters, primarily not-in-office encounters, during the concurrent recording of the initial survey than in the retrospective follow-up. The procedures used in the reliability study were expected to produce such differences, since they focused almost exclusively on abstracting data available in *office* records. Not-in-office encounters were generally not recorded in a way that was readily accessible to the physician's office staff and, therefore, were not reported in the practice-audit booklet.

[5] Kappa can range from -1.0 to $+1.0$. Zero indicates a chance level of agreement; $+1.0$ is perfect agreement; and values less than 0 indicate poorer than chance agreement. Since kappa should only reach $+1.0$ when there is perfect agreement, i.e., the off-diagonal cells are empty, inequality in the distribution of the marginals for the variables being compared produces a maximum kappa of less than $+1.0$. Skewness in the marginals alone, i.e., a nonrectangular distribution of responses among the various categories of the nominal variable, does not affect kappa.

[6] The index, like kappa, ranges from -1.0 to $+1.0$, with 0 indicating a chance level of agreement. It, however, takes its maximum value when the observed agreement equals the maximum possible agreement, given the marginal distributions of the item. Thus, inequality in the marginals does not affect the index of reliability.

[7] The analysis of the number of encounters is restricted to outpatient encounters, since information on inpatients was

not sought in the retrospective reliability study. It is also important to note that there were substantial amounts of missing data in this analysis even after excluding inpatient encounters. Of the missing data, 46 percent for week days and 34 percent for weekend days are due to a lack of information on not-in-office outpatient encounters in the follow-up study.

Discussion: An evaluation of the reliability of data gathered from three primary care medical specialties using a self-administered log-diary

Morton Israel, New York Department of Health

In her work, Harkins encountered three methodological problems: (1) instrumentation, (2) definition of terms, and (3) research design.

I have chosen to concentrate on research design problems and applications. First I would like to ask Dr. Harkins a few questions about the design of the study; then I plan to make some suggestions and comments about the conclusions of the study.

Israel: Your sampling plan called for a matrix of 24 cells, made up of specialty, type of practice, and early-late respondents. Since the institutionally-based physicians were administered only a telephone interview, did you finally end up with 18 cells for reliability analysis?

Harkins: Yes, for the log-diary variables.

Israel: Thus, you ended up with a target sample of approximately 700 for direct interviews. Did this allow for adequate cell size for analysis?

Harkins: Yes.

Israel: Would you please describe the flow or process of the implementation of the research plan?

Harkins: The University of Southern California sent out, in batch, the questionnaires, which included a log-diary. The respondents' names were sent to us every two weeks. A sample of returns was selected for a reliability check. This sample was sent a follow-up log-diary. This mailing was followed by a telephone call.

Israel: How many women physicians were in the study?

Harkins: I am not sure. I am sure there were some, but we did not take this factor into account in sampling.

Israel: Was any consideration given to the particular patient population characteristics served by your sampled doctors, that is, age, sex, or social class?

Harkins: No, because data on patients were not available from USC until much later.

Israel: Did you have any opportunity to do any pretesting of the instrument or items employed?

Harkins: Because sample and instruments were based on the USC work, we were committed to and employed their instruments and approach with only a few needed modifications.

Israel: Why did USC use Sunday as a data collection day?

Harkins: They wanted to determine the total activities of physicians.

These questions were for general clarification and lead to the suggestions or comments that I have about the conclusions of Harkins' work.

The conclusion of the paper states that

> the characteristics of the respondents have much less of an effect on the quality of the data that they provide than do the characteristics of the data requested and the formatting of the questions presented to the respondents.

I suggest that a more detailed consideration of the social and medical context of the physician in his work setting is appropriate.

I think *volume, substance,* and *pattern* of physician practice (in a typical week) is dependent on his specialty, type of practice (private, group, clinic, or hospital-based), and the social class, age, and sex of his patient population. For example:

1. The volume, pattern, and nature of medical problems encountered by a pediatrician in private practice on Park Avenue are quite different from that of his counterpart in Harlem.

2. The problems encountered and the volume and flow of patients during a typical week of medical activities are much different for a hospital-based physician at Cornell Medical Center than for a physician at Harlem Hospital.
3. An internist on Park Avenue specializing in geriatrics and another caring for the elderly in the Bedford section of Brooklyn have different patterns of medical practice because of the unique needs, problems, and demands of their patient populations.

Thus, I am suggesting that many of the problems of instrumentation, definition and clarification of terms, and research design in this area of reliability investigation should be addressed by a full consideration of the unique development of the substance (modal medical problems encountered by the physician with his patient population), volume, and pattern of the physician's work week. This might be accomplished by developing reliability studies that consider

1. Specialty,
2. Type of practice (private, clinic setting, hospital-based), and
3. Specific characteristics of the patient population served (age, sex, income group.)

In sum, one way in which future reliability studies will be able to build on the accomplishments of the present study is by emphasizing differences according to physicians' type of practice and the characteristics of their patients.

Economic surveys of medical practice: AMA's Periodic Surveys of Physicians, 1966–1978*

Louis J. Goodman, Center for Health Services Research and Development, American Medical Association

Lynn E. Jensen, Center for Health Services Research and Development, American Medical Association

Introduction

Since 1966, the American Medical Association (AMA) has conducted annual economic surveys of physician work patterns and practice characteristics. This survey series, known as the Periodic Survey of Physicians (PSP), provides the AMA and others with an analytical data base on office-based medical practice in the United States.

The PSP is one of only five major economic surveys of physicians conducted in the U.S. The other four include *Medical Economics'* Continuing Survey of Medical Practice (MEDECON), the National Medical Care Expenditure Survey (NMCES), the Bureau of the Census Survey of Service Industries, and the Health Care Financing Administration/National Opinion Research Center Survey of Physicians' Practice Costs and Incomes (HCFA/NORC). Two of these surveys, PSP and MEDECON, are long standing and conducted on an annual basis. The remaining three surveys are relatively recent and have not been subjected to close scrutiny.

This paper describes the PSP and assesses the progress and problems of the survey over the past decade. Particular attention is devoted to response/nonresponse studies and survey experiments that have been conducted by the AMA Center for Health Services Research and Development. The discussion concludes with an examination of future directions for the PSP and its role in health services research.

The Periodic Survey of Physicians and AMA's data base

The major component of AMA's data base is the Physician Masterfile, which contains current and historical information on every known physician in the United States, including non-members of the AMA, American medical graduates practicing abroad, and foreign medical graduates (FMGs) practicing in this country. This well-defined population allows the Association to select subpopulations for additional data collection and analysis. The PSP is selected from one such subpopulation of physicians. Other subpopulations include group practice physicians, foreign medical graduates, women physicians, and medical school alumni, to name a few.

To maintain the validity and completeness of information on both the Masterfile and sample surveys, each physician's record is updated throughout his career by a triennial census. In addition, a computerized system updates information on each physician continuously. Approximately 5,000 updates—triggered by AMA mailings; physician correspondence; and communications with other organizations such as hospitals, government agencies, speciality boards, and licensing agencies—are processed weekly.

The quality of Masterfile data is essential to the selection of the PSP sample and to subsequent analyses of PSP data. Accordingly, the AMA monitors the inclusion of all variables in the Masterfile, maintains an ongoing quality control program for data updates, and undertakes activities to certify the validity and reliability of the data contained in the file.

Since the quality of certain data elements in the research files constructed from the PSP series depend on the quality of the Masterfile itself, some elaboration is indicated. The quality of Masterfile data has been subjected to several examinations by organizations independent of the AMA and was recently compared to data bases for other professionals (Yett, 1977).

Over the past decade, five separate studies examined the reliability of Physician Masterfile data and found these data to be highly consistent with comparison data samples. All studies,

* The views expressed in this paper are those of the authors and do not necessarily represent the official position or policy of the American Medical Association.

summarized in Goodman and Eisenberg (1977), indicate physician location is accurate within a range of 2.1 to 6.6 percent. As expected, the most fluid variable is the physicians' professional activity.

Quality of data in the Masterfile is important for the PSP series for two reasons. First, the Masterfile provides the universe from which the sample is selected. Second, specific data elements from the Masterfile are merged with questionnaire data.

PSP procedures

In 1966, Theodore and Sutter designed the first PSP to gather information yearly on the practice of medicine:

> The purpose of the first Periodic Survey of Physicians was fourfold: (1) to compare data from the AMA records against speciality and activity information obtained from physicians in the sample; (2) to arrive at a descriptive profile of the physician population in terms of basic characteristics available from both of these sources; (3) to obtain estimates on the utilization of physicians' services in terms of weeks worked per year, number of hours practiced per week, and number of patient visits; and (4) to test the feasibility of obtaining such data through surveys conducted by the Association. [Theodore and Sutter, 1967:516]

Survey design and methodology. Since its inception, the PSP has been based on a probability sample of physicians selected from the AMA Physician Masterfile. Although sample designs have varied somewhat over the years to correspond with specific research objectives, the annual survey has always included a representative sample of office-based physicians engaged in active medical practice in the United States.

Currently the PSP sample is based on a stratified systematic probability sample of approximately 5 percent of the entire U.S. nonfederal, office-based physician population. Specialty representation in the sample is proportional to the size of the specialty category in the PSP physician population. Determination of the PSP sample size, usually around 10,000 physicians, is based primarily on the expected rate of response and the desired precision for the population estimates. For a more detailed description, see Henderson (1978).

The PSP survey instrument is a mail, self-administered, four-page questionnaire. A cover letter from the Executive Vice President of the AMA is printed on the first page of the instrument. The letter explains the purpose of the survey, assures confidentiality of response information, and requests the individual physician's participation. The remaining three pages of the questionnaire are devoted to closed-ended, precoded questions. The physician's name and address are computer-printed on a portion of the last page of the questionnaire. This portion may be removed by the respondent to ensure confidentiality during the processing of the questionnaire. The questionnaire and a prepaid business reply envelope are sent to the physician by first class mail, a procedure associated with high response rates among physicians (Goodman, 1975; Gullen and Garrison, 1973).

Efforts to maintain the confidentiality of physician data include (1) eliminating physicians' names and addresses from all documents and data files containing response data, (2) creating a computer-generated physician identification code meaningful only for file linkage purposes, and (3) destroying data files that facilitate linkage upon completion of data processing.

The PSP questionnaire is designed to include a *core* section that is subject to only minor revisions from year to year and a *research* section that is modified substantially from one annual survey to another. The core section includes questions on physician's practice characteristics, work patterns (hours worked and patient visits), and finances (fees, incomes, and professional expenses). The research section is designed to meet specific objectives in the AMA Center's research program. In the recent past, the research section has dealt with such topics as provider reimbursement and the impact of technological innovation on medical practice. All new questions, as well as modifications to existing questions, are subject to review and approval by a standing committee of the Center.

PSP pilot survey. A pilot test of the survey questionnaire is conducted prior to each annual survey. The primary objective of the pilot survey is to evaluate new or modified questions for clarity, readability, and rate of response. The pilot survey is conducted according to precisely the same procedures as those used for the annual survey, except on a smaller scale.

Pilot questionnaires are mailed to a representative random sample of approximately 500 physicians drawn from the PSP population. Responses to the pilot test are edited, coded, and tabulated; the results are then reviewed to determine the necessity for questionnaire revisions.

Modifications to the instrument are made as needed on the basis of write-in comments and responses that are inconsistent with comparable information or a priori expectations. Pilot results are of substantial use in the wording of unambiguous questions to elicit valid responses (Oppenheim, 1966; Sudman, 1976).

Quality control of PSP data. Selection of a statistically designed representative probability sample of physicians initiates the data collection process. A computerized response log is maintained on each physician in the sample. This log includes the following information:

Physician identification code
Date of initial mailing
Dates of follow-up mailings
Response code
Date of response

The computerized log is updated with appropriate information upon receipt of the returned questionnaire. Follow-up mailings to all nonrespondents are conducted at approximately monthly intervals. As many as five follow-up mailings may be conducted during the course of the survey. Manual and machine editing of returned questionnaires is based on a set of standard procedures described in Henderson (1978).

Steps to prevent disclosure of sensitive information are also extended to the analytical and reporting phases of the survey. For example, tabular information is presented only at major levels of aggregation (i.e., census division, major specialty category). In addition, security is maintained to protect the data file from either inadvertent or unauthorized disclosure.

Uses of the PSP. The current PSP represents the efforts of over a decade of experience. Since the first PSP was conducted in 1966, data collected with this instrument continue to have varied uses.

First, data from the series are published regularly. The AMA's annual *Profile of Medical Practice* contains basic data on medical practice and research reports based on PSP data. Table 1 provides an example of the type of PSP data that are routinely published. The compound rate of growth estimates indicate that substantial changes have occurred during this period with respect to the economics of medical practice.

Next, PSP supports major AMA research activities (Jensen, 1979). Recent AMA research papers utilizing PSP and related data base information have dealt with group practice development (Freshnock and Goodman, 1979a, 1979b; Goodman, Bennett, and Odem, 1977), the evolving role of foreign medical graduates (Way, Jensen, and Goodman, 1978), physician location and specialty choice (Werner, Wendling, and Budde, 1979), and selected aspects of the market for physician services (Cotterill, 1978).

Finally, PSP data have also been used for public policy purposes. For example, during the Economic Stabilization Program (1971–74), the AMA was able to assist the Cost of Living Council in developing fair and workable physician fee guidelines, based in part on socioeconomic data from the PSP.

PSP survey response rates

In recent years, survey response rates to the PSP have declined and do not increase appreciably after the first follow-up attempt. Various organizations, both public and private (e.g., Health Care Financing Administration and California Medical Association), have surveyed extensively the social and economic aspects of physician service delivery. These organizations have found that medical economic data are very difficult to collect. During 1974, the AMA also became concerned that response rates to the PSP were beginning to fall below the exceedingly high levels previously obtained.

Table 2 summarizes AMA's experience with response rates for this survey. As indicated, rates have dropped from 80 percent in 1966 to

Table 1
Physicians' mean net income, expenses, fees,[a] and total hours, by specialty, 1969–1976

Item	General/family practice			Internal medicine			Surgery			Pediatrics		
	1969	1976	Compound rate of growth	1969	1976	Compound rate of growth	1969	1976	Compound rate of growth	1969	1976	Compound rate of growth
Net income	34,734	47,438	4.6	37,630	60,459	7.0	48,848	73,245	6.0	31,812	46,962	5.7
Expenses	24,170	42,407	8.4	21,352	45,567	11.4	25,474	54,779	11.6	18,898	36,624	9.9
Fees (IOV)	7.83	14.80	9.5	16.58	31.2	9.5	12.93	23.29	8.8	9.48	17.45	9.1
Total hours	52.0	51.4	−.2	52.8	55.7	.8	51.5	54.4	.8	52.9	49.8	−.9

[a]Average fee for initial office visit.

Source: Periodic Survey of Physicians, 1972–1976.

Table 2
AMA Periodic Survey of Physicians: sample size, response rate, and number of mailings

PSP no.	Survey year	Sample size	Percent response	Number of mailings
1	1966	3,544	79.9	4
2	1967	5,265	70.4	5
3	1968	5,885	70.1	6
4	1969	5,052	68.3	4
5	1970	7,563	62.0	5
6	1971	7,842	64.0	4
7	1972	9,160	55.5	6
8	1973	7,500	55.9	5
9	1974	10,169	48.7	4
10	1975	10,295	51.4	5
11	1976	10,655	50.0	5
12	1977	10,000	48.7	6
13	1978	10,000 (est).	(Currently in progress)	

the current 50 percent level. Experience shows that after the initial mailing, each successive callback obtains increasingly fewer returns.

Decline in survey response rates, in general, can be attributed to a variety of factors. Those most relevant for the physician population seem to be the changing social climate of the last decade and specific survey design considerations. We shall discuss these topics in turn.

Social climate of economic surveys. In recent years, the American public, including physicians, have become increasingly reticent to provide private and public organizations with personal information. Public demand that information be used only for authorized purposes was reflected in two pieces of federal legislation—the Privacy Act (PL 93-579) and the Freedom of Information Act (PL 94-409). As was true of the general population, physicians were very concerned with the use of their personal, professional, and business information, especially when associated with personal identifiers.

The effects of these conditions were manifested in the exceedingly large numbers of letters and telephone calls that AMA received concerning identification numbers printed on PSP9. The previous two surveys, as well as PSP9, had indicated on the questionnaire that a physician's name would remain "anonymous," when "confidential" would have been more accurate. That is, while the physician's name is not linked to his/her identification code for analysis, the two are indeed linked for follow-up mailings. The term "anonymous" evoked a strong reaction in 1974-75; the term "confidential" was substituted the following year and physicians' expressions of concern subsided.

In addition to a changing social climate, certain specific features of the PSP survey design itself could have reduced physician response rates to the questionnaire. Since its inception, the PSP had been used to collect increasing amounts of information. At this juncture, it was necessary to reassess the PSP's basic survey design to determine how well the survey was faring in light of a changing social environment.

Complexity and anonymity experiment. In 1974, an experiment was devised to assess the response effects of questionnaire complexity and respondent anonymity. A national probability sample of 320 office-based physicians was selected from the AMA Physician Masterfile. Table 3 shows the four experimental groups; 80 physicians were assigned to each group through a randomization procedure. The issues that were tested were whether complexity (write-in versus fixed-alternative questions) and anonymity (identification code versus no code on the questionnaire) were significant factors affecting response to the PSP.

Table 3
Responders and nonresponders to complexity and anonymity survey, 1974

Type of questionnaire	Responders	Nonresponders	Total
Write-in questions, identification code	39	41	80
Fixed alternative, identification code	36	44	80
Write-in question, no identification code	36	44	80
Fixed alternative, no identification code	44	36	80

$\chi^2_{(3)}=2.15$, $p > .20$.

All four groups were mailed initial questionnaires and follow-ups on the same days. The experiment was terminated after approximately two months in the field. The instrument with write-in economic questions and an identification code was mailed as a control, since that was the current PSP survey strategy. The remaining three groups received variants on that survey design.

In general, PSP questionnaires utilize "write-in" questions to collect economic information. For example:

1. What was your 1977 INDIVIDUAL NET INCOME BEFORE TAXES from medical practice to the nearest $1,000?

1977 Individual Net Income
Before Taxes = $_____,000

PROJECTED 1978 Individual
Net Income
Before Taxes = $_____,000

A less complex method of questionnaire item construction is fixed-alternative response categories. For example:

2. Is your practice organized as a professional corporation?
Yes No

3. Which of the following best describes the type and size of your practice arrangement?

a. Solo practice

b. Practice with one or more other physicians

c. Other arrangement

Specify: _____

d. Number of Physicians in the practice INCLUDING yourself:

1. Who are associated FULL-TIME (more than 20 hrs/wk) with this practice: .....

2. Who are associated PART-TIME (less than 20 hrs/wk) with this practice: _____

The dimension of respondent anonymity is assessed by including or not including an identification code on the instrument. The identification code is included on the questionnaire for follow-up mailings and for merging survey data with professional and biographic information extracted from the AMA data base.

Overall, there was no significant difference between the four groups with respect to response. The response was highest, however, for the group receiving the anonymous (no identification code) fixed-alternative questionnaire.

In spite of the findings of this experiment, the PSP was subsequently modified. A concerted effort was made to utilize intervalized fixed-alternative type questions whenever feasible. Furthermore, a guarantee of confidentiality was printed on the questionnaire next to the identification code.[1]

The findings of this experiment corroborate those from another study in the same time period. Fuller (1974) assessed the effect of anonymity and identification on the responses of naval personnel to a mail survey. His findings also indicated group differences in response were not large enough to be of practical importance.

Sources of nonresponse. The catalogue of nonresponse sources (e.g., ineligibles, not-at-homes, refusals) designated by Kish (1965) is generally instructive with respect to the PSP, since questionnaires are mailed to a sample drawn from a well-defined population. Furthermore, the address section of the PSP population file is continuously updated, thereby reducing a potentially large category of ineligible bad addresses. Nevertheless, ineligible returns that do occur must be excluded from the nonresponse category. The ineligible category has ranged from 3 to 5 percent of the total sample, depending on the survey year. Ineligibles include retirees, physicians temporarily not in practice or out of the country, and inactive, deceased, and federally-employed physicians.

In addition, questionnaire returns and follow-up attempts sometimes cross in the mails. Approximately 1 percent of respondents complete more than one questionnaire, causing duplicate returns that must be removed.

In their review of the literature on mail survey response rates, Kanuk and Berenson (1975) found that follow-up mailings are very successful in increasing survey response. However, the unit cost of additional information increases with each successive wave.

Follow-up mailings. Follow-up mailings or callbacks may be an effective method of reducing nonreturns in mail surveys. Follow-ups are used extensively in most mail surveys because each added mailing brings additional returns. However, our experience has shown a monotonic decrease in response with each additional callback. As indicated by Kephart and Bressler (1958), the first follow-up yields a relatively greater percentage of responses. Preliminary response data to our most recent survey (PSP12) were fitted to a model shown in Table 4, suggested by Kish (1965).

Table 4 indicates that a total of 40,554 questionnaires mailed yielded 5,546 responses from the 10,000 possible responses. An average of 7.3 calls per response was needed to achieve an overall response rate of 55.5 percent. Refusals, out-of-the-country, illness, and certain other categories of participants are excluded from this analysis.[2]

Almost two-thirds of the response was achieved in the first two calls. Results for the last follow-up mailing (Wave 6) include returns over a longer time period than the previous waves. These returns may also include late re-

Table 4
Cumulated response with successive follow-ups for PSP12[a]

Wave of mailing (call) i	Responses n_i	Cumulated responses $\sum_{r=1}^{i} n_r$	Calls on possible respondents (no. mailed) t_i	Cumulated calls $\sum_{r=1}^{i} t_r$	Response rate (%) n_i/t_i	Calls per response t_i/n_i	Mean calls of cumulated response $\sum t_r/\sum n_r$
1	2,162	2,162	10,000	10,000	21.6	4.6	4.6
2	1,272	3,434	7,838	17,838	16.2	6.2	5.2
3	847	4,281	6,566	24,404	12.9	7.6	5.7
4	446	4,727	5,729	30,133	7.9	12.9	6.4
5	169	4,896	5,285	35,418	3.2	31.3	7.2
6[b]	650	5,456	5,136	40,554	12.7	7.9	7.3
Total[c]	5,546	—	10,000	40,554	55.5	—	—

[a] For additional discussion of this table format, see Kish (1965).

[b] Responses to Wage 6 were cumulated over a longer time period than preceding waves. In addition, follow-up mailings and late responses crossing in the mails are contained in this category.

[c] Inclusion of refusals, etc., and exclusion of ineligibles would result in an effective survey response of 48.7% to PSP12.

turns from previous waves. The results point out the substantial costs involved in obtaining responses to an economic mail survey utilizing follow-up techniques.

In a mail survey of graduate medical education program directors conducted by the AMA Research Center, a response rate of 51 percent was increased to 65 percent by utilizing telephone reminders (Way et al., 1978). Nonrespondents were called on the telephone and requested to respond as soon as possible. This procedure and others mentioned previously are currently being evaluated for possible use on the PSP. In addition, other remedies for nonresponse, such as substitution, replacement, and subsampling, will receive particular attention in future efforts.

Nonrespondents and representativeness of PSP data. In November 1974, an analysis of nonrespondents to PSP9 was undertaken. A national probability sample of 10,169 office-based physicians was drawn from the AMA Physician Masterfile. An initial mailing with two follow-ups obtained a response rate of 43 percent after three months.

Before attempting a third follow-up mailing, an analysis of respondents and nonrespondents was conducted. Five known key attributes of respondents and nonrespondents were compared to determine if a systematic bias existed with respect to underreporting of certain groups.[3] Ineligibles and refusals were excluded from the analysis. If a systematic response bias were to be uncovered, consideration would be given to (1) substitution of participants for nonresponses or (2) replacement of nonresponse categories with nonresponses from previous PSP surveys, among others.

Nonresponse effects were estimated for the following sample attributes: (1) specialty, (2) geographic location, (3) age, (4) type of practice, and (5) AMA membership. Differential effects of nonresponses were indicated in several attribute categories. These are discussed below.

The results of this analysis show that among the physicians surveyed in the Ninth Periodic Survey of Physicians, the speciality of the physician had some but not a major effect on his willingness to respond to the survey. Physicians in the specialties of general practice, internal medicine, surgery, and obstetrics-gynecology responded to the survey at a somewhat lower rate than did physicians in other specialties.

Similarly, the response rate to the survey varied somewhat by geographic region. Physicians in the North Central States, particularly the West North Central States, had a higher response rate to the survey than other physicians. The South, particularly the West South Central States, had a low response rate compared with other areas of the country.

The major factors affecting response to the survey appear to have been the physician's age and AMA membership. Members of the AMA had the highest rates of response to the survey.

Association sponsorship, as indicated by Marquis (1978b) and Sudman and Bradburn (1974), may account for a large proportion of variation in response. A priori, the differential impact of sponsorship on response in unknown. Ostensibly, AMA sponsorship increases response among members and either decreases response or has no effect on nonmembers of the Association. Nonmembers and physicians in the 46–55 year age range had the lowest response rates. Type of practice did not appear to have an effect on physicians' willingness to respond to the survey.

The nonrespondent analysis indicated a need for additional follow-ups and continued experimentation with the PSP methodology. The findings from the response-nonresponse analysis show that nonresponse as well as other factors may have influenced total survey response.[4]

The questionnaire is initially mailed in March/April of each year. The length of time between surveys is currently fixed. The initial mailing takes place at the end of the first quarter of each successive year. Follow-up mailings occur at monthly intervals thereafter. The physicians sampled now receive the survey questionnaire while federal income tax calculations are presumably fresh in their minds. Therefore, the economic data requested are more likely to be available.

Controlling nonresponse to the PSP. Thus, we have utilized three methods of controlling nonresponse: (1) improving procedures that deal with complexity and anonymity, (2) making successive follow-up mailings to nonrespondents, and (3) estimating the sources and effects of nonresponse and apportioning nonrespondents into subgroups.

Two other methods suggested by Kish (1965) and Sudman (1976) are planned for the next annual PSP. These methods include

1. Substitution of nonrespondents in the current survey with nonrespondents from a previous survey, and
2. Development of a replacement methodology for nonrespondents matched on the basis of known attributes of the entire sample or of the nonrespondent group.

A major problem with replacement schemes, such as that used by the HCFA/NORC survey, is the high probability associated with substituting a "responder" for a "nonresponder," thereby increasing the risk of introducing bias. That is, replacing nonrespondents with participants (similar along a continuum of personal attributes to respondents) could introduce substantial bias into the analysis. The optimal but costly solution involves substituting potential participants with the same vector of characteristics as nonrespondents.

Economic surveys of physicians

Several surveys currently collect economic information from the physician population. Table 5 represents information on the five major economic surveys of physicians. The two established mail surveys of physicians (PSP and MEDECON) differ substantially from the more recent HCFA/NORC telephone survey. Data are not available on the National Medical Care Expenditure Survey, which utilizes in-person interviews, or on the Bureau of the Census Survey of Service Industries. Furthermore, the Census Bureau survey is conducted on an ad hoc basis, the previous one having been conducted in 1905. A survey of this type is also subject to substantial data-handling and time-lag problems. The Bureau of the Census survey may not be directly comparable to the PSP, HCFA/NORC, or MEDECON because its survey population is restricted to physicians with two or more employees.

Although limited in scope, a general assessment of these three surveys may be useful. Aside from the indicated large cost variances between surveys, a major differential is mail versus telephone data collection procedures. Before comparing similar data produced by each survey in question, a short digression on the respective advantages and disadvantages for mail and telephone surveys is in order.

Mail versus telephone surveys. There are several advantages and disadvantages associated with each type of data collection procedure. Telephone interviews as a survey technique offer the following advantages:

1. Immediate response after an appointment has been scheduled;
2. Feeling of personal interest transmitted to the respondent by a trained interviewer;
3. Better response rates than obtainable with a mail instrument;
4. Interviewers trained to "convert" initial refusals;
5. Opportunity for callbacks to elucidate and verify information.

In contrast, several disadvantages have been linked to the telephone strategy:

1. Immediate recall or fingertip access of detailed information is required;
2. Increased opportunity for guesswork and approximations;
3. Greater chance of miscommunication of information (poor connections, interruptions, misunderstanding, etc.);
4. No opportunity for physician verification of transmitted data;
5. Reluctance on the part of physicians to discuss financial information on the telephone;
6. Annoyance on the part of physicians for being interrupted from busy schedules;
7. Potential interviewer bias;
8. Increased costs (personnel, callbacks, etc.).

Inability to locate telephone numbers for some physicians is a major disadvantage of tele-

Table 5
Selected economic surveys of medical practice

Survey	Cost	Timing	Type of survey	Non-response analysis	Sample size	Response (%)
AMA/PSP	$47,000	Annual 1966	Mail	Yes	10,000	50
MEDECON	NA	Annual 1929	Mail	No	25,000	35
HCFA/NORC	$400,000	Annual 1975	Telephone[a]	Yes	1,000–5,000	68–70
Bureau of the Census	NA	One time	Telephone	No	100,000 (est.)	Response mandated by law[b]
NMCES NCHS/NCHSR	$1 million plus	One time 1978/1979	In person and telephone	NA	11,000 households	NA

[a] Although the large majority of interviews were conducted over the telephone, mail surveys and in-person interviews were substituted in certain cases (Sloan, Cromwell, and Mitchell, 1977).

[b] Physicians failing to complete the economic census were subject to a $500 fine.

NA—Not available.

phone interviews. In a recent survey conducted for the AMA, Market Opinion Research (1978) found it necessary to exclude 20 percent of the physicians in the sample owing solely to the inability to find a professional or home phone number for the physician. MOR telephone interviewers found that the acquisition time per completed interview was excessive. Some interviewers were forced to make three or four calls at appointed times to the same physician only to obtain a refusal.

Comparative survey estimates of physicians' average net income. This section examines differences and similarities in income data produced by the PSP, MEDECON, and HCFA/NORC for the same year. Although analysis of this single data element is not sufficient for generalizations on the three surveys, it does provide a vehicle for making some meaningful comparisons.

As indicated in Table 6, the three surveys for which data are available exhibit sizable differences by specialty with respect to physicians' average net income data. These differences result from several sources and are evaluated with respect to the AMA/PSP.

HCFA/NORC survey. The current HCFA/NORC Survey of Physicians' Practice Costs and

Table 6
Comparison of physicians' average net income data, by specialty, 1975
(In Dollars)

Source of data	Total (5 specialties)	General/family practice	Internal medicine	Surgery	Pediatrics	Obstetrics-gynecology
AMA/PSP[11]	56,740	45,410	59,980	68,230	44,250	63,320
HCFA/NORC	53,600	44,800	53,900	61,300	50,000	64,600
MEDECON	58,440	GP: 43,360 FP: 53,440	60,130	67,450	48,330	72,380

NOTE: Data from the three surveys are not strictly comparable for a variety of reasons:

(a) PSP and HCFA/NORC data are means, whereas MEDECON data are medians. Medians for HCFA/NORC are unavailable; PSP medians are as follows:

Total, five specialties	50,400
General practice	41,820
Internal medicine	52,130
Surgery	60,390
Pediatrics	42,280
Obstetrics-gynecology	59,980

(b) The surveys use different methods. The PSP and MEDECON data are collected using a mail survey; the HCFA/NORC survey is conducted primarily over the telephone.

(c) Specialty classifications do not coincide exactly for all three data sources.

(d) The surveys vary to some degree in the type of practice covered. HCFA/NORC surveyed fee-for-service physicians only. The AMA data are for office-based physicians, a broader category than only fee-for-service. MEDECON surveys self-employed physicians.

Source: American Medical Association (1978).

Incomes utilizes telephone interviews with a national sample of approximately 5,000 physicians in 18 specialties. The physicians in this sample were randomly selected from 30 primary sampling units. The initial 1975 survey of 1,000 fee-for-service, office-based physicians in five specialties provided results slightly different from the PSP. HCFA/NORC's explanation of these findings erroneously attributed sampling variance to AMA sampling procedures (U.S. Social Security Administration, 1977). Since the PSP is based on a scientifically designed and executed methodology, variance in findings is probably the result of the following factors:

1. HCFA/NORC data are not accompanied by error variance estimates; the PSP sample is five times larger and for this reason alone should provide more precise estimates.
2. The PSP sample includes salaried physicians and those practicing in large prepaid groups as well as fee-for-service physicians. This may account for shorter hours worked per week in PSP results.
3. The PSP sample includes office-based physicians in all specialty categories as opposed to the five primary care specialties surveyed by HCFA/NORC.

The HCFA/NORC survey did obtain a response rate of 64 percent compared with the current PSP level of approximately 50 percent. However, the major issue in connection with response is whether the survey respondents are representative of the total population. In effect, how close do sample estimates (y) approximate true population values (Y)? The PSP has been subjected to extensive analysis for response and nonresponse bias, including the effects of timing on survey mailings. These analyses have demonstrated a high level of reliability and accuracy of survey responses. For a more complete model of measurement error and its effects on survey response, see Andersen (1975).

MEDECON survey. The MEDECON mail questionnaire survey was conducted every three or four years between 1929 and 1963 and has been repeated annually from 1963 to the present. Prior to 1972, only solo, office-based, self-employed physicians under the age of 65 were eligible for selection. Since 1972, the sample has included nonsolo and incorporated self-employed physicians of all ages. In recent years, samples have consisted of approximately 25,000 physicians, with response rates of about 35 percent.

The MEDECON survey differs from the PSP with respect to sample design and definition, survey response, and nonresponse analysis, as well as the reporting of economic data. First, PSP as well as HCFA/NORC report mean figures, while MEDECON reports only medians. Second, the survey population is not defined as "office-based" physicians as in the other two surveys. Instead, MEDECON defines their sample according to physician's employment status.

MEDECON surveys only self-employed physicians, which excludes a large proportion of the potential "office-based" population. Response/nonresponse analyses have apparently not been conducted for any of the annual surveys, although sample size is substantially larger than for either of the other two economic surveys. Variance estimates do not accompany data reported from this survey. Therefore, more detailed assessment of survey item reliability is not possible at this time.

Dunning and Cahalan (1973–74) show that mail questionnaires are an effective means of collecting data on potentially sensitive issues. The principal uncertainty of the mail technique is the return rate, which is lowest for MEDECON.

Overall, more information on survey design and basic economic survey data is needed to make more definitive statements. At present, the differences described should more than account for data variances between surveys.

Future directions for the Periodic Survey of Physicians

The need for complex socioeconomic information on medical practice will increase in the future. Mail, telephone, and home interview strategies in medical economic surveys face the concomitant problems of maintaining response rates and survey data quality. On the one hand, large-scale surveys such as the NMCES are relatively expensive and require a substantial time investment, from field work to research file construction and data analyses. On the other hand, smaller scale surveys like the PSP are relatively inexpensive but have seen a gradual slippage in response rates. The PSP survey takes approximately nine months to produce an edited research file merged with other relevant research and data files. The timeliness of this type of survey must be evaluated in a cost-effectiveness framework.

The issue of cost-effectiveness in maintaining response rates requires further analysis, including the development of improved strategies for dealing with nonresponse. One strategy is to utilize telephone follow-ups in combination with self-administered mail questionnaires. In a recent study, Siemiatycki (1979) compared mail, telephone, and home interview strategies for

cost and data quality. His findings showed that the combined mail/telephone or telephone/mail strategies were much less costly and just as effective as home interviews. In addition, self-administered mail questionnaires may have a slight advantage in obtaining responses to sensitive questions.

Physicians are an overly burdened respondent group, which may account in part for lowered response rates. Some physicians may be relatively more overburdened than others with respect to certain personal or professional characteristics. Special strategies need to be developed to deal with such problems if valid and timely physician data are to be collected.

Footnotes

[1] Practical requirements preclude the mailing of totally anonymous questionnaires, since follow-up mailings could not be attempted and research files utilizing professional, biographic, and demographic data could not be constructed.

[2] As indicated in Table 2, inclusion of refusals, etc., and exclusion of ineligibles results in an effective response rate to PSP12 of 48.7 percent.

[3] Item response analysis was not attempted in the preliminary study although later evaluations have found some item-by-item differences among the more sensitive financial questions.

[4] Respondents to PSP10, PSP11, and PSP12 are not statistically different from physicians in the total survey population with respect to age, specialty, and geographic location (AMA, 1979).

Discussion: Economic surveys of medical practice

Lynn A. Evans, Department of Community Medicine, Baylor College of Medicine

The major emphasis of the paper by Goodman and Jensen is the response to the Periodic Survey of Physicians (PSP). It is to this area that I will restrict my remarks. The response rate to the PSP as reported began at a high of 79.9 percent in 1966 and has declined to the 50 percent level, where it has hovered for the past four years. In considering the area of nonresponse, three questions come to mind: Does the nonresponse make any difference? If yes, why does it occur, and what can be done about it? Each of these is now examined in turn.

Does it make a difference?

In an attempt to answer this question, Goodman and Jensen described an effort made, using the 1974 PSP (PSP9), to examine the respondent and nonrespondent after two follow-ups were made. The five known attributes examined were (1) specialty, (2) geographic location, (3) age, (4) AMA membership, and (5) type of practice. The results, as described by the authors, were as follows: (1) "specialty of the physician had some but not a major effect on his willingness to respond to the survey"; (2) "the response rate to the survey varied somewhat by geographic region"; (3) and (4) "the major factors affecting response to the survey appear to have been the physician's age and AMA membership"; (5) "type of practice did not appear to have an effect on physicians' willingness to respond to the survey."

In addition, the authors state that some item-by-item differences were found among the more sensitive financial questions. Although the authors did not report specific numbers from the information provided, it would seem that the nonresponse is making a difference. It would have been interesting to examine these differences on a wave-by-wave basis.

Why does it occur?

Goodman and Jensen directly address this issue by stating that

> decline in survey response rates, in general, can be attributed to a variety of factors. Those most relevant for the physician population seem to be the changing social climate of the last decade and specific survey design considerations.

In a report presented at the Second Biennial Conference, Marquis (1978b) refutes the contention that survey response rates have been changing because of changes in society. As for the second reason, "specific survey design considerations," when the authors varied both level of complexity and anonymity, they found that neither significantly influenced response rates.

I would like to offer a third possibility that Goodman and Jensen mention in the very last paragraph of their paper—respondent burden. Consider the following:

1. Every physician in the Masterfile of the AMA is surveyed every three years and completes a short form known as the Physician Professional Activity questionnaire.
2. Yett (1977:195) reported that in 1975, 3,500 updates to the Masterfile were being processed, some of which included physician mailings. These were justified in order to maintain the accuracy and completeness of the Masterfile. By 1978, according to Goodman and Jensen, this figure had risen to 5,000 updates per year. Admittedly, many of these updates are likely to be the same physician's record and do not all require correspondence. However, there does exist a significant level of official contact between physicians and the AMA on a periodic basis.
3. MEDECON annually surveys 10,000–18,000 physicians.
4. In addition, there are a number of federal-level surveys, drug company surveys, state and local medical association requests for information, and numerous medical school mailings requiring information from their graduates.

Hence, a strong case can be made for the possibility of respondent burden. This hypothesis might be tested by asking physicians in the next PSP how many organizations they have recently provided with information. The results can be compared on a wave-by-wave basis. In addition, physicians who have not responded could be asked to explain their reasons for not complying with the request.

What can be done about it?

According to Bradburn (1978), respondent burden is a function of length, respondent effort, respondent stress, and frequency of being interviewed. From the PSP9 experience, it can be concluded that length does not seem to be a problem for this survey. The amount of respondent effort and stress from a surveyor's perspective would not appear to be great when compared with other surveys. However, from the physician's viewpoint, it might be large and is probably influenced by the frequency of being interviewed. This is even more likely when frequency of being interviewed is expanded to include responding to all requests for information.

The AMA is not currently in any position to change the frequency with which physicians are contacted by other organizations. Nor would it be wise for them to decrease their own efforts at maintaining an accurate Masterfile. Hence, the physician's perception of the level of effort required remains the most promising target. The physician must perceive that it is worth his or her time to complete the PSP questionnaire. There are several strategies that might be considered:

1. Having the AMA take the initiative to organize a clearinghouse for physician surveys that would contain representatives from organizations such as the AMA, the APHA, AAMC, NCHSR, and NCHS.
2. Conducting the survey under the joint auspices of selected medical schools in each region, the state medical association, and the AMA.
3. Offering to provide the sampled physician a discount on a new or renewal AMA membership in return for completing the PSP questionnaire.
4. Offering to provide the sampled physician with a personalized feedback report describing how he or she ranks when compared with the remainder of the sample.

In conclusion, the PSP, along with the Masterfile, represents an extraordinarily valuable asset to the health care delivery system. Every effort by the AMA and others to maintain and increase its value should be encouraged.

Methodology of the National Ambulatory Medical Care Survey: Evaluation and extensions

John D. Loft, National Opinion Research Center, University of Chicago

Introduction

The National Ambulatory Medical Care Survey (NAMCS), sponsored by the National Center for Health Statistics and conducted by the National Opinion Research Center, surveys physicians in office-based private practice. Each year since 1973 the survey has collected information about patient visits, including demographic data on the patient, description of the patient's health problem, and the physician's treatment. The survey design is unusual in that it calls for physicians to record data prospectively on a sample of visits occurring during a week-long reporting period. The NAMCS has been successful in maintaining relatively high response rates over the years—response climbed from 76 percent in 1973 to 81 percent in 1977—despite this respondent burden.

This paper is concerned with the methodology of the NAMCS. It focuses on two areas: nonresponse and the extension of the NAMCS procedures to the collection of other kinds of medical data.

To provide a background for the rest of the paper, the first main section briefly describes NAMCS's sampling design, data collection procedures and forms, and data processing.

The following section presents the results of a study, conducted by NORC in 1975, of factors that might affect physicians' decisions about participation in the survey (Minor, Mullan, and Loft, 1976). Response rates for the combined 1973 and 1974 cycles of the survey were compared across various professional, demographic, and geographic variables to evaluate the extent of possible nonresponse bias. Interviewing procedures and physicians' reasons for refusal were also examined for factors that might possibly affect participation in the survey.

The final section discusses several projects that have used data collection methods based on the NAMCS in different settings and for collecting data other than those included on the NAMCS form.

Survey design

The NAMCS design and procedures have been described in detail in several articles (e.g., Minor et al., 1976; U.S. National Center for Health Statistics, 1974). This section highlights the main features of the survey.

A sample of approximately 3,000 physicians is selected each year from the American Medical Association and the American Osteopathic Association masterfiles of physicians in office-based practice. Physicians who are not in office-based practice (i.e., hospital- or university-based, or government employees) and those in subspecialities that do not typically provide direct medical care (i.e., radiologists, pathologists, anesthesiologists) are excluded from the enumeration. Eligible physicians are sorted into the primary sampling units (PSUs)—SMSAs and counties—maintained by NORC[1] and stratified by speciality to ensure coverage of as many specialities as possible. Since the PSUs are chosen with unequal probabilities of selection, the sampling rate for each PSU is set so that the overall probability of sampling any one doctor is about the same. After selection, the sampled physicians are randomly assigned to one of the 52 weeks of the calendar year.

Patient visits are sampled at a rate determined by the number of patients each doctor expects to see each working day of the reporting week. Doctors with small practices are asked to record data for every patient visit; doctors with larger practices sample every other visit, every third visit, or every fifth visit. The rates are calculated so that, on the average, doctors complete no more than ten patient records a day.

Three weeks before the assigned reporting week, NCHS mails a packet introducing the survey to each doctor. The packet contains letters

from NORC and NCHS briefly describing the survey and letters of endorsement from medical societies.

About a week later the interviewer contacts the doctor through a telephone call to his or her office. During this call, the interviewer verifies the doctor's status as an office-based physician providing direct care for ambulatory patients and arranges an appointment for a personal interview.

During the personal induction interview, the interviewer records certain descriptive information about the doctor's practice, determines the appropriate patient sampling rate, and explains the use of the patient log and patient record form. The patient log (Exhibit 1) is a form on which the doctor records in sequence the names of all of the ambulatory patients he or she sees during the reporting week. There are four different forms of the patient log, corresponding to the four different sampling rates.

The patient record form (PRF) is the basic data collection instrument of the survey. The items on the PRF cover the patient's demographic characteristics, the reason for visit (including symptomatic and nonsymptomatic reasons), and the physician's diagnoses and treatment. The form has been modified from year to year (Exhibit 2 shows the PRF for 1978), but the idea of a single page limited to information normally determined during an office visit or readily accessible to the physician on medical records has been basic to the design of the form. The first four or five items can be completed by a receptionist, and the remainder of the form should take only about a minute for the doctor to fill out.

Interviewers are instructed to make two follow-up calls to the doctor during the reporting week: on Monday to remind the doctor to begin the survey and on Wednesday to make sure that everything is going well. The second call also provides an opportunity for the interviewer to answer any questions that the physician may have after using the forms for two days.

After the reporting week, the physican mails the completed PRFs to the interviewer for editing. Missing or inconsistent data are retrieved through either telephone calls or personal visits to the doctor's office.

When the induction interview forms and the PRFs are received at NORC's Chicago office, they are again edited and attempts are again made to retrieve missing or inconsistent information. After this review, both forms are sent to NORC's Data Processing Department for conversion to computer tape.

In data processing, the reason for visit is coded into a Reason for Visit Classification (RVC). The RVC was developed especially for NAMCS by the American Medical Records Association and provides codes for both symptomatic and nonsymptomatic reasons for visits. The physician's diagnoses are coded according to the International Classification of Disease (ICD).

After the induction interviews and PRFs are coded, keypunched, cleaned, and merged into a single file, missing data are imputed for the following items: date of visit; patient's birth date, sex, and race; and whether the patient has been seen before, for either this or another problem. The imputations are based on physician's specialty, the RVC code, and the ICD code. After the imputation, clean data are put into an SPSS system file to facilitate tabulation.

The procedures described above apply to the majority of cases in the NAMCS samples, where the physician agrees to participate in the survey freely or with only a minimum of coaxing from the interviewer. Of course, this is not always the situation. At any time from the initial screening telephone call to the retrieval of missing information, a doctor may decide not to cooperate with the survey procedures. The extent of refusal, factors associated with refusal, and possible bias attributable to nonresponse are discussed in the following section.

Nonresponse in the NAMCS

Table 1 presents the response rates for the NAMCS for each survey year from 1973 through 1977. As the table shows, response rates have increased over the years, especially from 1974 (75 percent) to 1975 (80 percent). Although extensive analysis of this change over time has not yet been done, the results of a study conducted in 1975 using the first two cycles of the NAMCS and examination of other data from field operations records are indicative of some important factors influencing response.

Reasons for refusal. Although the completion rates for the first two cycles of the survey were respectable, one of the major purposes of the 1975 evaluation of the NAMCS was the examination of reasons for refusals in order to develop strategies for increasing the response rate in subsequent cycles.

Reasons for refusal were coded into five categories: the physician said that he or she was too busy to participate; the physician rejected this or all surveys; the physician gave personal reasons for his or her inability to participate;

A

PATIENT LOG

Date _____ 19 ____

As each patient arrives, record his name on the log below. and complete the correspondingly numbered patient record to the right.

PATIENT NUMBER	PATIENT'S NAME
11	
12	
13	
14	
15	
16	
17	
18	
19	
20	

PHYSICIAN'S COPY

B

PATIENT LOG

As each patient arrives. record name and time of visit on the log below. For the patient entered on line 2, also complete the patient record to the right.

PATIENT'S NAME	TIME OF VISIT
1	a.m.
	p.m.
2	a.m.
Record items 1-19 for this patient	p.m.

CONTINUE LISTING PATIENTS ON NEXT PAGE

C

PATIENTLOG

As each patient arrives, record name and time of visit on the log below. For the patient entered on line 3, also complete the patient record to the right.

PATIENT'S NAME	TIME OF VISIT
1	a.m.
	p.m.
2	a.m.
	p.m.
3	a.m.
Record items 1-19 for this patient	p.m.

CONTINUE LISTING PATIENTS ON NEXT PAGE

D

PATIENT LOG

As each patient arrives, record name and time of visit on the log below. For the patient entered on line 5, also complete the patient record to the right.

PATIENT'S NAME	TIME OF VISIT
1	a.m.
	p.m.
2	a.m.
	p.m.
3	a.m.
	p.m.
4	a.m.
	p.m.
5	a.m.
Record items 1-19 for this patient	p.m.

CONTINUE LISTING PATIENTS ON NEXT PAGE

Exhibit 1. Patient Logs: A Form (every patient); B form (every second patient), C form (every third patient); and D form (every fifth patient).

ASSURANCE OF CONFIDENTIALITY—All information which would permit identification of an individual, a practice, or an establishment will be held confidential, will be used only by persons engaged in and for the purposes of the survey and will not be disclosed or released to other persons or used for any other purpose.

D 610427

PATIENT RECORD
NATIONAL AMBULATORY MEDICAL CARE SURVEY

1. DATE OF VISIT
Mo / Day / Yr

2. DATE OF BIRTH
Mo / Day / Yr

3. SEX
1 ☐ FEMALE
2 ☐ MALE

4. COLOR OR RACE
1 ☐ WHITE
2 ☐ NEGRO/BLACK
3 ☐ OTHER
4 ☐ UNKNOWN

5. WAS PATIENT REFERRED FOR THIS VISIT BY ANOTHER PHYSICIAN?
1 ☐ YES
2 ☐ NO

6. PATIENT'S COMPLAINT(S), SYMPTOM(S), OR OTHER REASON(S) FOR THIS VISIT *(In patient's own words)*
a. MOST IMPORTANT _____
b. OTHER _____

7. TIME SINCE ONSET OF COMPLAINT/SYMPTOM IN ITEM 6a *(Check one)*
1 ☐ LESS THAN 1 DAY
2 ☐ 1-6 DAYS
3 ☐ 1-3 WEEKS
4 ☐ 1-3 MONTHS
5 ☐ MORE THAN 3 MONTHS
6 ☐ NOT APPLICABLE

8. PHYSICIAN'S DIAGNOSES
a. PRINCIPAL DIAGNOSIS/PROBLEM ASSOCIATED WITH ITEM 6a
_____
b. OTHER SIGNIFICANT CURRENT DIAGNOSES
_____
_____

9. HAVE YOU SEEN PATIENT BEFORE?
1 ☐ YES 2 ☐ NO
IF YES, FOR THE CONDITION IN ITEM 8a?
1 ☐ YES 2 ☐ NO

10. SERIOUSNESS OF CONDITION IN ITEM 8a *(Check one)*
1 ☐ VERY SERIOUS
2 ☐ SERIOUS
3 ☐ SLIGHTLY SERIOUS
4 ☐ NOT SERIOUS

11. DIAGNOSTIC SERVICES THIS VISIT *(Check all ordered or provided)*
1 ☐ NONE
2 ☐ LIMITED EXAM/HISTORY
3 ☐ GENERAL EXAM/HISTORY
4 ☐ PAP TEST
5 ☐ CLINICAL LAB TEST
6 ☐ X-RAY
7 ☐ EKG
8 ☐ VISION TEST
9 ☐ ENDOSCOPY
10 ☐ BLOOD PRESSURE CHECK
11 ☐ OTHER *(Specify)* _____

12. THERAPEUTIC SERVICES THIS VISIT *(Check all ordered or provided)*
1 ☐ NONE
2 ☐ IMMUNIZATION/DESENSITIZATION
3 ☐ DRUGS (PRESCRIPTION/NONPRESCRIPTION)
4 ☐ DIET COUNSELING
5 ☐ FAMILY PLANNING
6 ☐ MEDICAL COUNSELING
7 ☐ PHYSIOTHERAPY
8 ☐ OFFICE SURGERY
9 ☐ PSYCHOTHERAPY/THERAPEUTIC LISTENING
10 ☐ OTHER *(Specify)* _____

13. DISPOSITION THIS VISIT *(Check all that apply)*
1 ☐ NO FOLLOW-UP PLANNED
2 ☐ RETURN AT SPECIFIED TIME
3 ☐ RETURN IF NEEDED, P.R.N.
4 ☐ TELEPHONE FOLLOW-UP PLANNED
5 ☐ REFERRED TO OTHER PHYSICIAN
6 ☐ RETURNED TO REFERRING PHYSICIAN
7 ☐ ADMIT TO HOSPITAL
8 ☐ OTHER *(Specify)* _____

14. DURATION OF THIS VISIT *(Time actually spent with physician)*
_____ MINUTES

HRA-34-2
REV. 9-76

DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
PUBLIC HEALTH SERVICE
HEALTH RESOURCES ADMINISTRATION
NATIONAL CENTER FOR HEALTH STATISTICS

O.M.B. #68-R1498

Exhibit 2. Patient Record Form, 1978.

Table 1
Response rates for the National Ambulatory Medical Care Survey: 1973–1977

Sample size and outcome	Cycle 1[a]	Cycle 2[b]	Cycle 3[c]	Cycle 4[d]	Cycle 5[e]
Total sample size	1,695	3,029	3,015	3,024	3,000
Out-of-scope physicians	282	458	398	489	494
Net sample of eligible in-scope physicians	1,413	2,571	2,617	2,535	2,506
Completed cases with PRFs	944	1,629	1,770	1,768	1,759
Participant physicians with no PRFs	136	311	312	285	278
Total participant physicians	1,080	1,940	2,082	2,053	2,039
Response rate[f]	76.4%	75.5%	79.6%	81.0%	81.4%

[a] April 1973–1974.
[b] May 1974–December 1974.
[c] January 1975–December 1975.
[d] January 1976–December 1976.
[e] January 1977–December 1977.
[f] Response rate = $\frac{\text{Total participant physicians}}{\text{Net sample}}$.

the interviewer was unable to speak with the doctor; and other reasons. The distribution of these reasons for refusal is presented in Table 2. Over half of the refusing doctors gave "too busy" as their reason for refusal; 25 percent rejected the idea of surveys in general or of this survey in particular; 3 percent of the refusals occurred because the interviewer simply couldn't get past a receptionist or secretary to speak to the doctor; 2 percent of the refusals gave personal reasons, such as illness or death of a relative; and 16 percent of the reasons could not be coded easily into these categories, mainly because of multiple reasons (e.g., "I don't like government surveys and I'm too busy anyway").

Table 2
Reasons for refusals and breakoffs, Cycles 1 and 2

Reason	Percent
Too busy	52.2
Reject survey	25.6
Personal reasons	2.6
No contact with doctor	3.5
Other	16.1

NOTE: N = 940. This excludes 19 refusals for which no specific reason was recorded.

Reasons for refusal appear to be related to the time at which refusals occur. Table 3 shows the timing of refusals and breakoffs. As might be expected, the majority of refusals—70 percent—occur at the initial screening telephone call. At this time, the physicians have not

Table 3
Timing of refusals and breakoffs, Cycles 1 and 2

Timing	Percent
During telephone screening	70.4
During induction interview	14.9
Prior to recording week	1.5
During recording week	13.2

NOTE: N= 858. This excludes 114 cases that were refusals or breakoffs but for which there is no specific information about timing.

seen the survey forms and probably have only a vague idea of the survey tasks. It is therefore likely that refusals at this point are based more on negative reactions to surveys in general—for example, because they are a waste of valuable time—rather than to the features of this particular survey. The remaining 30 percent are equally divided between refusals during the induction interview and breakoffs after an initial agreement to participate. Note that almost all of the breakoffs occur during the reporting week rather than during the week or so between the interview and the beginning of the recording period. This suggests that breakoffs probably occur because of specific survey tasks rather than because of general objections to the survey.

Overall, these findings indicate the importance of the telephone screening call as the point at which physicians are most likely to decide about participating in the survey. As discussed below, endorsing letters from medical organizations help to "break ground" for the interviewers by pointing out the value of the survey to the medical profession. Also, inter-

Table 4
Completion rates by number of letters received from specialty societies

Number of letters	Completion rates[a]					
	Cycle 1		Cycle 2		Cycles 1 and 2	
	Percent	N	Percent	N	Percent	N
0 letters	77.2	669	73.6	1,164	74.9	1,833
1 letters	75.9	635	76.5	1,170	76.3	1,805
2 or 3 letters	75.3	109	79.3	237	78.0	348

[a]The equation for completion rates is $CR = \frac{\text{Cases with PRFs + unavailables}}{\text{Cases with PRFs + unavailables + refusals and breakoffs}}$.

viewer training notes the importance of the "too busy" refusal by alerting interviewers to the fact that the induction interview should last only about 20 minutes and that completion of the patient log and patient records should add only a minimal amount of time to the physician's workday.

Endorsement by medical societies. Feasibility studies conducted in 1970 through 1971 (U.S. NCHS, 1974) indicated that letters of endorsement from medical societies such as the AMA, the AOA, and speciality societies were an effective means of increasing participation in the survey. These letters noted the importance of information about ambulatory care to the medical profession and to medical education programs. Post-survey evaluation interviews with doctors who participated in these feasibility studies showed that more than half of the respondents said that they were influenced by these letters.

In the 1975 evaluation, we examined the impact of such endorsements more directly. In the 1973 and 1974 cycles, all of the doctors in the samples received a letter of endorsement from the AMA or the AOA; some also received letters from the speciality societies of which they were members. Table 4 presents the distribution of response rates across the number of speciality society letters received. The number of letters appears to have had a small positive effect on participation: 75 percent of the doctors who received only the AMA or AOA letter participated in the survey, whereas 78 percent of the doctors who received two or more additional letters participated. Athough the percentage difference is small, it represents about 120 doctors and, if each of these doctors completed patient records for the average 35 visits a week, about 4,200 patient records.

Conversion attempts. If at any time interviewers encounter a refusal that they cannot convert, they complete a report documenting the reason for refusal; the case is then transferred to a "converter," an interviewer specially trained at handling refusals. After reviewing the reported reason for refusal, the converter contacts the doctor, usually by telephone, and attempts to respond to whatever problem or objection the doctor has with the survey.

Table 5 displays the results of these efforts over the five years from 1973 through 1977. The conversion rates have increased from about a fifth of the initial refusals in 1973 to a little over a third in 1977. In 1977, the conversion efforts added nearly 10 percent to the final response rate of 81 percent.

We have found that the most effective conversion tactic is a long distance telephone call to the physician from an NORC supervisor in Chicago or in another large city. Nearly all of the conversions are accomplished through this means. Many physicians are impressed with the importance of the survey when they receive a personal

Table 5
Conversion rates, Cycles 1–5

Year	Initial refusals	Attempted conversions	Successful conversions	Conversion rate based on	
				Refusals	Attempts
1973	424	406	85	20.0%	20.9%
1974	719	690	117	16.3%	17.0%
1975	817	772	200	24.5%	25.1%
1976	721	697	237	32.9%	34.0%
1977	724	720	255	35.2%	35.4%

telephone call from Chicago, and they are then more willing to listen to arguments for their participation.

The nature of these arguments, of course, is determined by the reasons for refusal originally given to the interviewer. Often, it is a matter of cleaning up a misunderstanding that the physician has about the purpose of the survey or reemphasizing the usefulness of the results and the importance of the doctor's contribution to the data set. In some cases, the converter may offer to send in an interviewer to provide clerical help if the doctor is very busy or understaffed.

Interviewer training and experience. One factor about which we have very little quantitative information is interviewing training and experience. However, since it is the interviewers who first contact the physicians and convince the majority of them to participate, who must explain the survey tasks and adapt the procedures to individual practices without jeopardizing comparability across practices, and who have the most direct control over the procedures in the field, their role should not be overlooked in a discussion of factors contributing to the success of the survey. In 1977, approximately two-thirds of the interviewers employed on the NAMCS had at least one year's experience on the survey. This experience is invaluable for developing the rapport, flexibility, and judgment required of the interviewers throughout the survey.

Response rates by physician characteristics. In the 1975 evaluation of the NAMCS, we computed response rates for eligible doctors and compared these rates across the categories of several professional, demographic, and geographic variables. The concern was to identify physicians who might present particularly difficult field problems and also to identify biases that might arise from nonresponse among certain types of physicians. Table 6 displays the response rates across the professional, demographic, and geographic variables available from the sampling frame.

The overall response rates for the various physician groups are clustered around the overall average of 76 percent except for two relatively high percentages (91 percent for "other" specialties in Cycle 1 and 84 percent for psychiatrists in Cycle 2) and four relatively low percentages (68 percent for the East South Central region and 70 percent for physicians 70 years or older in Cycle 1 and 69 percent for general and family practitioners and 70 percent for the Mountain region in Cycle 2). These distributions suggest that the effect of nonresponse bias is minimal. However, since data recorded in the induction interview and on the patient records are not available from nonparticipating doctors, the actual effect of nonresponse bias cannot be determined.

Summary. It appears that response rates are not strongly affected by any single factor; instead, a number of different features of the survey are involved. Endorsement of the survey by medical societies, efforts to minimize respondent burden, effective conversion tactics, and interviewer training and experience all contribute to the relatively high response rates enjoyed by the NAMCS. Nonresponse does not appear to be concentrated in any one type of physician, except for physicians in general and family practice, who are likely to have large, busy practices.

These response rates, it should be noted, are based on all responding physicians, including "poor" respondents—that is, doctors who participated in the induction interview but saw few patients and doctors who reported on fewer patients than they should have. We count as respondents all physicians who agree to participate because NCHS wants the option of deciding when to discard a doctor's records because he does not have a reasonable number of patients to describe his practice. The NCHS figures for response rates, after deleting these doctors, are slightly lower. The figures for comparison are those in the last three columns of Table 1: the NCHS response rate figures are 80 percent for Cycle 3, about the same as ours; 79 percent for Cycle 4; and 78 percent for Cycle 5. These figures indicate that although doctors continue to agree to participate in the survey, the percentage who return usable data is dropping slightly and should be weighed against the increase in response rates.

As a final caveat, I should point out that NAMCS's high participation rates do not necessarily mean that the data collected are valid descriptions of ambulatory patient care provided in doctors' offices. Although extensive efforts are made to ensure that physicians understand the procedures and definitions of terms associated with the patient log and the patient record, it is not possible to monitor the physicians while they record the data, nor is it feasible, given the present procedures, to validate data directly with comparable information from an independent source.

After the 1975 evaluation of the NAMCS increased our understanding of the survey's procedures and affirmed their basic soundness, we were able to consider, in collaboration with interested parties, the extension of these proce-

Table 6
Completion rates by professional, demographic, and geographic variables[a]

Variable	Cycle 1		Cycle 2		Cycles 1 and 2	
	Percent	N	Percent	N	Percent	N
Total	76.4	1,413	75.5	2,571	75.8	3,984
Specialty groups:						
General and family practice	72.7	422	69.3	644	70.6	1,086
Medical	77.3	361	75.8	689	76.3	1,050
Surgical	76.1	503	78.3	969	77.6	1,472
Other	87.4	127	79.9	249	82.4	376
Specialty:						
General and family practice	72.7	422	69.3	664	70.6	1,086
Internal medicine	73.2	190	71.5	354	72.1	544
Pediatrics	81.8	77	82.6	161	82.4	238
Other medical	81.9	94	78.2	174	79.5	262
General surgery	73.7	156	78.7	268	76.9	424
Obstetrics and gynecology	75.8	120	79.6	226	78.3	346
Other surgical	78.0	227	77.5	475	77.6	702
Psychiatry	86.2	94	83.5	176	84.4	270
Other	90.9	33	71.2	73	77.4	106
Type of degree:						
Medicine	76.0	1,319	75.3	2,440	75.6	3,759
Osteopathy	81.9	94	77.9	131	79.6	225
Board qualification:[b]						
Board qualified	77.0	635	77.4	1,245	77.2	1,880
Not board qualified	76.0	778	73.7	1,326	74.5	2,104
Years since graduation from medical school:[b]						
Less than 10 years	83.1	83	80.0	185	81.0	268
10-19 years	78.7	399	75.1	814	76.3	1,213
20-29 years	75.2	379	77.2	666	76.5	1,045
30-39 years	72.0	314	72.6	552	72.4	866
40 or more years	75.7	144	73.5	223	74.4	367
Age:[b]						
20-39 years	84.0	213	77.3	440	79.5	653
40-49 years	75.3	450	76.3	778	76.0	1,178
50-59 years	75.8	376	75.5	693	75.6	1,069
60-69 years	72.9	251	70.6	398	71.5	649
70 or more years	69.6	79	76.3	131	73.8	210
Sex:						
Male	76.5	1,371	75.6	2,476	75.9	3,847
Female	73.8	42	72.6	95	73.0	210
Census region:						
Northeast	75.3	369	75.4	732	75.4	685
North Central	80.3	401	77.8	595	78.8	675
South	73.5	388	75.2	703	74.6	703
West	76.5	255	73.2	541	74.2	510
Census geographic division:						
New England	77.9	86	73.9	157	75.3	243
Middle Atlantic	74.6	283	75.8	575	75.4	858
East North Central	80.3	320	79.1	468	79.6	788
West North Central	80.2	81	73.2	127	76.0	208
South Atlantic	75.6	225	75.1	373	75.3	598
East South Central	68.4	57	78.1	114	74.9	171
West South Central	71.7	106	74.1	216	73.3	322
Mountain	82.9	76	69.9	143	74.4	219
Pacific	73.7	179	74.4	398	74.2	577
Area:						
SMSA	76.3	1,139	74.9	2,211	75.3	3,350
Non-SMSA	77.0	274	79.2	360	78.2	634

[a] The equation for completion rates is $CR = \frac{\text{Completed cases}}{\text{All in-scope cases}}$.

[b] D.O.s are excluded from these rates because information on these variables was not available from AOA Masterfile.

dures to other kinds of medical data collection efforts. The next section of this paper describes three such surveys.

NAMCS procedures in other settings

In 1976, NORC conducted a pretest of the NAMCS procedures to collect data on ambulatory patients seen in hospital outpatient departments. In 1977, we examined the feasibility of adding special supplements to the basic patient record. Currently, the NAMCS methods are a part of a survey of physician participation in Medicaid programs. These three projects are examples of the way in which the NAMCS might be extended to other types of data collection requirements.

Extension of the NAMCS to hospital outpatient clinic visits. As the name implies, the National Ambulatory Medical Care Survey was designed to collect data on all encounters between physicians and ambulatory patients. Although the majority of such encounters occur in physicians' offices, a substantial number occur in the outpatient clinics and emergency rooms of hospitals. Especially in large urban areas, these settings are an important component of ambulatory care and may be the exclusive or primary source of medical services for some families. Moreover, there is reason to believe that patients treated in hospital clinics and emergency rooms may differ from those treated in a doctor's private office in some important ways, including demographic characteristics, symptoms presented, and the severity of their problems.

In 1977, NORC investigated the feasibility of extending the procedures developed for the survey of office-based ambulatory care to hospital outpatient clinics (Sheatsley, Scharf, and Loft, 1977). The specific goals of the study were a preliminary evaluation of the applicability of the NAMCS methods to hospital clinics, revisions of the forms and procedures, and a field test of the revised procedures in a sample of 20 hospitals.

The data collection problems encountered in a hospital are, of course, quite different from those faced in office practice. First, there are the sheer number and diversity of the personnel who must be convinced to participate and who must be trained in the survey tasks. Even though a hospital director may agree to cooperate with the survey, the success of the NAMCS procedures in a hospital setting depends on the cooperation of clinic heads, staff doctors, nurses, clerical help, and anyone else who may participate in either direct patient care or the management of patient visits. Hospital employment of part-time personnel, in both medical and nonmedical positions, further hinders the training of some people who may be expected to complete survey forms.

Furthermore, patients are processed through hospitals by clinic association rather than by doctor association. This added another level to the sampling of patient visits; that is, the secondary sampling unit was the clinic. Information that would enable the sampling of clinics had to be collected from each hospital.

Finally, in addition to these differences between hospitals and doctors' offices, hospitals may differ from one another. The procedures had to be adapted to large hospitals with many clinics and to small hospitals with only one or two clinics in a way that would ensure comparability of the data collected from both types.

In spite of the differences between office practices and hospital clinics, we attempted to keep changes to a minimum in order to preserve comparability between the data collected in both settings. The survey forms, the patient log, and the PRF used in the hospitals were identical to the NAMCS forms. The only departure was that there was no patient visit sampling within the clinics because it was decided that the complexity of sampling visits would offset the benefits of reduced respondent burden.

A sample of 20 hospitals was selected with stratification by geographical region, size of community, type of hospital ownership, and size of hospital. Each hospital was assigned to one of NORC's Area Supervisors, who was responsible for inducting the hospital administrator and other key personnel, collecting data on the clinics in each hospital, and, after selection of the clinics, training clinic personnel and monitoring data collection. In hospitals with fewer than ten clinics, all were included in the clinic sample; in larger hospitals, a systematic sample of ten clinics was selected.

Whenever possible, the people who were to fill out the survey forms received personal training, usually in a group. In the larger hospitals, however, this proved unwieldy. The survey procedures were outlined in memos distributed to physicians and other personnel. Clerks completed Items 1 to 4 of the patient record form and attached the forms to patients' folders as these were distributed to physicians. Physicians then filled out the remainder of the PRF at the end of the patient visit. The Area Supervisor was available at all times for consultation.

Cooperation was assessed at the three major stages of the survey—initial contact, induction

of clinics, and data collection. The objective of the initial contact procedures was the hospital administrator's approval to conduct the survey in his or her hospital. Of the 20 hospitals approached, 17 (85 percent) agreed to participate. Clinic induction procedures were directed at persuading the heads of each clinic in the sample to participate in the survey. Of the 118 clinics sampled in participating hospitals, 117 (99 percent) participated in the survey. Finally, the success of the data collection depended on the cooperation of the doctors, nurses, and clerks within each clinic who were asked to complete a patient record for each patient treated. Success also depended on the ability of the Area Supervisor in each hospital to retrieve missing data. Unmonitored, the clinic personnel completed PRFs for 67 percent of the patients logged; after standard data retrieval by the Area Supervisors, this figure rose to 91 percent.

The feasibility study uncovered a number of problems that would require resolution before a full-scale survey could be initiated—notably, revisions in the forms, the problems of training the hospital staff to perform the survey tasks, and some problems with retrieving missing or incomplete information. The overall conclusion of the study, however, was that the procedures were workable in hospitals and could provide data comparable to those collected in the survey of office-based ambulatory care.

Supplement to the NAMCS patient record form. It has been suggested several times that a supplement could be used to collect special-purpose data in addition to "core" items on the PRF. A major concern about the use of supplements was whether the additional length of the PRF would significantly lower the response rate. An important factor in the success of the NAMCS in office-based practice has been the effort to make the survey as unobtrusive as possible, and it seemed possible that additional items could add to respondent burden.

In 1977, an opportunity arose to investigate this concern (Sheatsley and Loft, 1977). The Consumer Product Safety Commission (CPSC) asked NCHS to add a special supplement to the PRF to collect data on product-related injuries and illnesses seen by doctors in office-based practice. Original plans called for four CPSC questions to appear either on a small form separate from the PRF or on an addition to the PRF; in either case, it was to be clearly labeled as a "Special Purpose Supplement." There was, however, concern that a separate form might tend to be overlooked or mislaid and that there would be many opportunities for error in matching the supplements to the correct PRFs. Also, it was suggested that some doctors might view a "Special Purpose Supplement" as an extra burden and consequently be less willing to complete a separate form labeled as such. For these reasons, it was decided to present the supplement as a continuous part of the PRF. Thus, the four questions were added to the 14 items already on the PRF.

These revised forms were administered to a sample of 302 physicians who had not previously participated in the NAMCS. These physicians were selected through the same sampling procedures used in the NAMCS and randomly assigned to one of six reporting weeks during May and June of 1977. Contact and induction procedures were identical to those employed in the NAMCS. Doctors were not told that they were participating in a special pretest. The response rate among these physicians was then compared with the rate among the 345 doctors in the regular NAMCS assigned to the same six reporting weeks.

In the pretest sample, 235 physicians were categorized as eligible for the survey, and 80 percent (N = 187) of these in-scope doctors participated. In the ongoing NAMCS sample, 286 physicians were categorized as in-scope and 79 percent (N = 277) participated. The 1 percent difference between the response rates in the two samples is not at all significant, and we concluded that the longer form had no effect on the response rate.

Survey of Pediatrician Participation in Medicaid. The third project that extends the NAMCS procedures is a Survey of Pediatrician Participation in Medicaid, sponsored by the American Academy of Pediatrics and currently in the planning stage. The data collected in this survey will enable the estimation of the levels at which pediatricians in different states participate in Medicaid and the examination of the impact that this participation has on utilization of physician services.

The survey involves an induction interview somewhat longer than NAMCS's, during which information concerning the physician's opinions about and experience with Medicaid will be collected. In addition, data about the doctor's practice and the types of patients treated will be obtained.

Demographic data about each patient and information about each patient's health problem, mode of payment, and services rendered during the office visit will be recorded by the physician for a sample of his or her patients on a form similiar to the NAMCS Patient Record Form.

Six months after this main survey, we will collect data on the doctor's charges for each of the sample visits, the amount actually paid, the source of payment, and the length of time elapsed before collection. This information will be collected through a mail survey, with telephone and personal follow-ups where necessary.

Summary. These three projects demonstrate ways in which the NAMCS procedures can be extended to other types of health research problems. These procedures are particularly well suited to physician surveys in which the basic data collection unit is the physician-patient encounter. The Outpatient Feasibility Study and the ongoing survey of office-based practice show that physicians in both settings are generally cooperative, as long as the goals of the survey appear to justify the time that it takes to participate.

Although time is important in justifying the study to physicians, we learned in the CPSC pretest that additional items per se do not affect response rates and that it is feasible to supplement the basic NAMCS items with questions of more particular interest. It is, however, important to limit the amount of information collected about each encounter to what can be recorded on a single page.

Finally, another way to extend the survey is to expand the induction interview to obtain more information about providers of health care and the settings in which they practice. This information might then be related to the data about encounters. For example, in the Survey of Pediatrician Participation in Medicaid, data about the physician's experiences with and opinions of Medicaid and extensive data about the physician's practice are collected in the induction interview.

Footnote

[1] In order to reduce respondent burden, doctors in each year's sample are assured that they will not be asked to participate again in the NAMCS for at least two years. Because of this promise, since 1974 smaller counties in the NORC master probability sample have had to be replaced because all or almost all of the doctors in the counties fell into the sample for a prior year.

Discussion: Methodology of the National Ambulatory Medical Care Survey

William D. Kalsbeek, School of Public Health, University of North Carolina

Introduction

Collecting good-quality health data from sources of ambulatory care presents survey researchers with a severe challenge to their collective ingenuity. The principal source of difficulty in this setting is that one must deal with the "elusive" physician—"elusive" not in a derogatory sense of the word but because our request for his or her participation must be cast against other priorities: excessive yet still-growing patient needs, increased paperwork, even other research endeavors. This then is the setting within which the National Ambulatory Medical Care Survey (NAMCS) must operate and to which the Loft paper is addressed.

I congratulate the author on presenting an interesting discussion of the design, realities, and potential of this important survey. The paper leads me to several comments that I will categorize into three broad subject areas: sampling design, survey design not directly related to sampling, and the problem of nonresponse.

Sampling design

The design for sampling patient ambulatory visits apparently consists of three stages: primary sampling units (i.e., relatively small county aggregate area units), office-based physicians, and patient visits. The design as described leads to an equal probability sample of physicians. Rates for systematic sampling within a physician's office are, however, set at one in 1, 2, 3, or 5 depending on an estimate of the number of patients seen by that physician. This action, taken to establish a more constant caseload among physicians, is done at the expense of retaining the statistical advantages of a self-weighting design. I wonder if we might not "have our cake and eat it too" by a two-phase design in which a large sample of physicians would be chosen with equal probabilities in the first phase. Measures of size reflecting patient volume would be collected from the first-phase sample and used for selecting a second-phase sample with probabilities proportional to size. Assuming that the size measures were of reasonably good accuracy, selecting patient visits with probabilities inversely proportional to size would yield very close to a self-weighting design with nearly equal size clusters in the final stage.

Nonsampling component of survey design

The prospective approach used in the NAMCS is probably well chosen in view of the large inherent difficulties with a retrospective approach. The major drawback of the latter approach is its heavy reliance on the physician's medical records as a source of data. Physician medical records, as found by this discussant and others, are not particularly conducive to survey research. Specifically, these record systems are almost invariably developed to meet the physician's individual needs, both for accounting purposes and for establishing patient histories to provide a continuity of care. Standardization among these individualized record-keeping systems is therefore almost nonexistent, thus presenting both measurement and logistical problems. There are also, however, some difficulties with the prospective approach; for example, some degree of *survey control* must be sacrificed by placing the overwhelming burden of data collection in the hands of those being surveyed. One cannot expect to feel completely comfortable with the deployment of data collectors for whom the survey has some direct effect but little direct benefit either monetarily or professionally.

Another major concern is the diligence with which data collection methods are carried out. One wonders, for example, whether all patient visits are logged by the physician. My experience from a recent study has led me to speculate that there is perhaps some underrecording in this kind of prospective operation (Kalsbeek et

al., 1975). In this study, office-based physicians were randomly allocated to one of two groups: one in which patients with head or spinal cord trauma were identified prospectively and the other in which this same type of patient was identified retrospectively. Results indicated that the number of reported patients with head or spinal cord trauma was somewhat lower in the prospective group than in the retrospective group. This might lead one to speculate that lesser control contributed to this result.

Methods to minimize problems with the prospective approach have been established and are correctly emphasized in the NAMCS. Thus, one method is having the data edited for consistency and completeness by both the interviewer and the central office staff. Another method calls for intervention in the form of two telephone calls by the interviewer during the week of data collection.

Given the importance of control in a prospective data collection operation, I wonder if perhaps even closer intervention by the interviewer might be worth considering. For example, the interviewer might make additional calls during the week of data collection and also check the NAMCS patient log against the receptionist's weekly logbook in the physician's office. There are, of course, practical limitations to the extent that one can expect to intervene without having negative effects on survey participation. Perhaps, as a specific suggestion, it would be worthwhile to test the effect on response rates of increasing interviewer involvement.

The author mentions that the NAMCS data collection forms were modified each year. Without knowing any details, I wonder if any of these changes might have contributed to a loss in the between-year comparability of certain estimates. Finally, procedures call for the demographic items of the patient record form (PRF) to be completed by a receptionist and for the rest of the form to be completed by the physician. In this regard, I was curious whether there was any evidence that completion of the remaining items on the PRF might also have been relegated to some assistant (e.g., nurse or receptionist).

Nonresponse

I move finally to a discussion of nonresponse, which, considering the aforementioned difficulty with surveying physicians, is probably the most significant aspect of the paper. The findings are an important contribution to knowledge about the size and implications of survey nonresponse in the NAMCS. Since most of the paper's discussion centers on the contents of several tables, I will organize my comments around each table individually.

Table 1 of Loft's paper summarizes response rates for cycles of the NAMCS from 1973 through 1977. The author includes in the numerator of his rates the category "Participant physicians with no PRFs"; yet I wonder if all physicians in this category should be viewed as survey respondents. Although it is not specifically stated in the paper, I am led to infer that this category includes those physicians who verbally agreed to participate but who, for one reason or another, did not complete data collection.[1] Excluding this class from the numerator reduces the response rate by 10–14 percent, although (as with the reported response rate) this modified response rate also generally increases with time.

Table 2 presents reasons for refusals and breakoffs and clearly illustrates physician "elusiveness," as mentioned earlier. I wonder, however, if *given* reasons completely reflect *actual* reasons. For example, might it be that an answer of "too busy" is really a diplomatic way of conveying "I don't like government surveys"? This unsubstantiated contention, if true, would suggest that the reason "too busy" is somewhat overstated.

Tables 3 and 4 (particularly the latter) deal with the role of endorsements in improving survey response rates. Table 3 points to the potential importance of endorsements and other strategies, particularly during the initial telephone contact. Response rates are presented in Table 4 by the number of specialty society endorsement letters sent in addition to the AMA or AOA letter sent to all physicians. Although effects of the apparently unrandomized assignment to reporting categories are unknown, intensifying endorsement efforts apparently produced slightly higher response rates in Cycle 2 and in Cycles 1 and 2 combined. Curiously, however, the effect of greater intensity appears to have had the reverse effect in Cycle 1, despite earlier indications in the paper that endorsements influenced over half of all respondents in their decision to participate. Perhaps, however, the real answer lies in how the *nonrespondents* reacted to these letters.

In all instances, differences in Table 4 appear to be nonsignificant statistically, although the author seems to suggest that the practical significance is real. These results, incidentally, are consistent with our own findings in a survey involving similar solicitation attempts on hospitals (Kalsbeek and Hartwell, 1977). Although the Loft paper correctly states the increase in re-

sponse that would have been realized if the maximum level of endorsements had been used, I wonder if we can generally expect that our often time-consuming and costly efforts to lobby professional associations for support will always be cost-effective. This issue needs to be resolved by further study; however, to speculate on a solution, perhaps the "best" approach is to use a single AMA/AOA endorsement.

Assuming that the staff of "converters" remained largely intact during 1973–1977, Table 5 nicely demonstrates an expected improvement in conversion rates as the converters accumulate experience in these matters. It would have been interesting, however, to know the background characteristics of this specialized group of survey workers and to see an analysis of conversion success by these characteristics. Some characteristics to consider would have been the amount of survey experience and the level of familiarity with the medical profession. We have observed, for example, that a converter with close professional similarities to the sample physician is the most successful. This, I would surmise, is due to the converter's ability to "speak the language" of the physician and thereby gain his or her confidence. Examples of this type of converter would include medical students, employees with some medical training, physician consultants, or physicians practicing in the same general area as the sample physician but sympathetic to the survey.

Comparison of response rates by amount of interviewer experience was mentioned but not included in the paper. Although this was an unfortunate omission, one might speculate that the number of years of survey experience and response rates are directly related. This relationship has been well established elsewhere in the literature (Durbin and Stuart, 1951; Heneman and Paterson, 1949).

One final set of results, illustrating response rates for Cycles 1 and 2 by various physician characteristics, is presented in Table 6. These findings are designed to indicate any relationships between a physician's inclination to respond and his or her other characteristics (e.g., age, region, etc.). Relatively constant response rates among categories of a physician characteristic imply that the characteristic is distributed similarly in both the responding and nonresponding subpopulations. Systematic differences in response rates, on the other hand, point to differences between respondents and nonrespondents. Furthermore, if these differences are related to survey variables, some level of nonresponse bias may be implicated. Most comparisons in Table 6 reveal small differences with little indication of predominating patterns except for some notable differences among specialty groups, number of years since graduation from medical school, and region.

Comparison of response rates in this manner is, however, only an indirect means of assessing the impact of nonresponse bias. I believe that more direct means are needed to establish conclusively that nonresponse bias is not significant, my principal argument being that we can only speculate on the completeness and real importance of the list of characteristics that have been investigated. For example, a paper presented by this discussant at the last Biennial Conference illustrated that large nonresponse biases may even emerge in surveys with 75 percent response rates and relatively small differences between respondents and nonrespondents (Kalsbeek and Lessler, 1978). I submit that the evidence presented in the Loft paper does not allow us to put to rest our concerns about nonresponse bias in the NAMCS.

Extensions of the NAMCS procedures

The paper concludes with a brief discussion of extensions of the NAMCS procedures. One interesting but difficult extension is to hospital clinics and emergency rooms. The feasibility study for extending the NAMCS procedures to clinics was apparently successful. Results clearly indicate that participation by clinics is almost completely assured once the hospital has agreed to participate. However, entrance into this area promises additional complexity in the logistics of data collection. In multiple-physician clinics, data collection will have to be centralized (with some inconvenience to physicians) or divided and coordinated among physicians. These difficulties would be compounded even further in dealing with emergency rooms where similar complexity and a greater transitory environment lurk, waiting to sidetrack any efforts at collecting good-quality survey data. As partial dispensation, dealing with patients seen in clinics and emergency rooms does give us recourse to quality checks of the survey data against secondary sources such as inpatient hospital records.

Another of the extensions discussed in the paper deals with the addition of a small supplement of four questions on product-related injuries. Results of a field experiment, which compared response rates for physicians approached with a 14- or 18-question PRF, led the author to conclude that added questions will not reduce the response rate as long as the information can be limited to the same length (one page in the NAMCS). We might speculate,

however, that lower response rates would have occurred if the form had been doubled or tripled in length.

To summarize, the Loft paper has presented us with an interesting view of the methods and some accompanying implications of the NAMCS. We conclude that the methods used in this difficult survey setting are apparently sound, although some possibilities for further improvement remain. Finally, we caution against any premature dismissal of the effects of nonresponse bias on survey estimates.

Footnote

[1] Clarification from Loft identified this class of physicians as those who completed the interview about practice characteristics but failed to complete the prospective log on physician activities. In every case, the doctor was given the patient records to complete, but for reasons such as personal illness, illness in the family, vacation, or leaving town for conferences during the preorganized reporting week, the doctor saw no patients during that week. Doctors who completed the induction interview but then refused to complete the patient records were handled like other refusals or breakoffs. The ambiguity is, of course, that some "hidden" refusals may have occurred when physicians who said that they would participate saw patients but did not complete the patient records and *said* that they saw no patients. The figures give the benefit of doubt to the physicians.

Open discussion: Session 1

Reliability of data from USC study

The open discussion on Harkins' paper was in large part devoted to clarification of the USC study (the primary research design) for which this study was a reliability check. In response to a series of questions, the following points of description about the USC study were made:

1. Each specialty board was asked to encourage participation from its membership. The response rate varied by specialty and by level of support from the organization. General and family practitioners had a lower response rate than did some other specialties.
2. The cooperation of the physicians and professional organizations included telephone calls by colleagues to sample members urging them to participate.
3. Battelle made no attempt to describe the final nonrespondents in the USC study. High-volume practices may be found more frequently among nonrespondents than among respondents. If so, productivity estimates made from the study may be understated.
4. Publications from the USC study have been released; more results are forthcoming. (See references in Harkins' paper.)

Other questions were directed toward clarification of the evaluation study that Harkins presented. In response she noted the following:

1. Physicians could be paid $50 for participation in the follow-up log-diary; 40 percent of the physicians were paid.
2. Results of the follow-up indicate that approximately one-half of the physicians worked on the diary as the day progressed; similarly, about 50 percent said that they waited until the end of the day to fill it out with the help of the day sheet. A few even admitted that they filled out at least part of the diary during the following weeks.

Comments on the evaluation methodology and the resultant conclusions about the research study were varied but centered around the impact of specialty practice on all of the other variables (whether or not respondent burden was interrelated with specialty). Evidence suggests, however, that response was more related to variability by follow-up procedures used by specialty groups to encourage participation than to the nature of the specialty practice types per se. There was more variation within than between specialties in response burden. Also, there were no problems associated with response burden and reliability once the physicians became respondents.

The problems associated with physicians delaying rather than filling in the diary at the time of service were discussed at length. Delay was acknowledged as a troublesome issue in using the physician service-diary methodology.

It was pointed out that more events (visits, services, etc.) were reported in the USC study than in the follow-up reliability study by Battelle. A question was raised about the effect that this differential in reporting rate had on reliability. Did this automatically reduce reliability? The paper presents two measures of reliability—kappa and the index of reliability. Kappa values are affected by this rate differential, whereas the index of reliability is not. Thus, minor differences in the reported reliability do occur, with the index of reliability being higher.

Concluding discussions addressed problems of identifying specialty practices through existing lists and ways of increasing physician responses through motivation, feedback, and encouragement of staff in the physicians' offices to organize responses.

Economic surveys of physicians

The general discussion on the paper by Jensen and Goodman and the remarks by Evans centered around respondent burden, efforts to in-

crease response rates, and justification of sample size. In some ways, these concerns overlap. Whether or not it was necessary to maintain the sample size of 5 percent of the physician universe for statistical purposes was addressed, with the comment that a reduction in sample size would be one way of reducing exposure to questionnaires by physicians over time. Reduction of sample size would have the added advantage of allowing the extra dollars to be spent on follow-up procedures, which should increase the response rate. A clearinghouse for surveys of physicians could also be used to reduce respondent burden.

It was noted that the American Dental Association uses telephone follow-ups and maintains a high response rate. Goodman noted that reasons for the large sample size were the necessity for doing regional analysis, for analyzing small subsections of the respondents based on practice characteristics not known prior to sampling (i.e., poststratification requirements), and for having an adequate number of respondents in regression analyses of subsections of the sample.

National Ambulatory Medical Care Survey Methodology

As in the general discussion of the Jensen/Goodman paper, the comments on the Loft paper turned to the problem of the physicians' being surveyed too much and the impact that this had on response rates and ultimately on the quality of the data. Perceived respondent burden may increase because the physician has been in a similar study before and also because a physician may be asked to participate in more than one survey running concurrently. Questions were directed toward the values of motivating respondents and of providing feedback to them. A recent study on breast cancer was cited in which a higher response rate was obtained by making the questionnaire more topic-oriented than routine. Responses to the concept of individualized feedback to respondents addressed the problems of cost and the relative advantages of indirect feedback.

Recommendations

This section summarizes suggestions made in Session 1 for further research and improved methodology for conducting surveys of physicians and other health care providers. It includes ideas from the three papers and from the discussants, as well as comments from the audience.

1. Reliability of information obtained from provider log-diaries and questionnaires may be improved by furnishing detailed instructions and definitions to respondents. Apparently even frequently utilized concepts such as "inpatient" versus "outpatient," "practice arrangement," and "specialty" are often not uniformly interpreted. However, long and explicit definitions may discourage respondents from thoroughly reading the instructions. Response rates may also be reduced by elaborate directions. These possible tradeoffs need to be systematically evaluated.
2. A purported advantage of the log-diary approach is that experiences are recorded soon after they happen. This minimizes recall problems and inaccurate reporting. If, however, the recording is done later, these potential benefits may be lost. The effects of provider characteristics and of the techniques used to elicit log-diary information on the time when the information is recorded and the completeness of that information require further investigation. In addition, a better understanding is needed of the effect that a delay in log completion has on the reliability and validity of the information provided.
3. It is plausible that physician characteristics such as specialty, type of practice, supporting facilities and personnel, and the kinds of patients seen will influence responses to requests for survey information. Additional efforts are needed to document such relationships. This work can lead to a better understanding of the differential approaches needed to obtain valid and reliable information from different kinds of practitioners.
4. Provider participation in social surveys may be influenced by (a) efforts to convince them of the importance of the study; (b) gains that they will derive from participation, such as compensation, feedback that might be useful for practice management, association and interaction with sponsoring professional, academic, and other esteemed organizations, and public service; and (c) encouragement of the providers' staffs, who often assume major responsibility for completing the required forms. Additional studies should investigate the relative importance of these and other incentives to increase participation.
5. A major concern about provider surveys, as well as other social surveys, is the effort needed to maintain acceptable response rates and factors that might be associated with the increasing difficulties of obtaining provider cooperation. Even data to document such a trend are somewhat anomalous, as evidenced by the declining rates shown in the Goodman/Jensen paper and the increasing rate

shown in the Loft paper. More should be done to establish what are the trends in cooperation on provider surveys. It is important in such studies to take into account the effort necessary to obtain a given response rate.

6. It is not clear what factors are critical in determining level of provider response. A number of factors are suggested as important: (a) auspices under which the study is conducted; (b) respondent burden in a particular survey; and (c) number of surveys to which a provider is asked to respond in a given time period. Efforts need to be undertaken to develop a general framework for considering these various factors in order to establish their relative importance in determining the current pattern of provider cooperation.
7. It was suggested by Evans that a clearinghouse for physician surveys be organized. Such a clearinghouse might contain (a) a listing of organizations that have in the past conducted, or are currently conducting, physician surveys, with descriptions of those surveys; (b) published and unpublished reports of the findings of these studies; and (c) data from these studies that might be used for secondary analyses. Such a clearinghouse would have the potential advantage of reducing physician burdens in filling out forms by offering an opportunity for coordinating the demands made on physicians by organizations and researchers. It would also reduce duplicate efforts and might thereby result in more cost-effective research. Finally, it has the potential to improve the quality of research by making it much easier to learn the current state of the art in the field.

Although appealing, this proposal would need to overcome a number of administrative and political barriers: (a) developing auspices under which the clearinghouse would be set up; (b) establishing a continuing financial and administrative base to support it; (c) convincing data-gathering organizations to participate and contribute to a data bank; and (d) ensuring the privacy of provider respondents. Despite obvious problems, the idea of a clearinghouse does seem worth pursuing so that more definitive recommendations could be made.

SESSION 2: Health interviews by telephone and the reinforcement and feedback to respondents by interviewers

Chair: Charles F. Cannell, Survey Research Center, University of Michigan

Recorder: Floyd J. Fowler, Jr., Center for Survey Research, University of Massachusetts/ Boston

A researcher's view of the SRC computer-based interviewing system: Measurement of some sources of error in telephone survey data

Robert M. Groves, Survey Research Center,
University of Michigan

We at the Survey Research Center, with support from NSF, are now in the midst of our first use of a computer-based interviewing system, in which telephone interviewers use CRT terminals to administer a questionnaire and record results. This is not a new technology; several other systems now exist, some having been in use for years. The designs of these systems are generally not part of the survey literature, yet one quickly discovers that several design features affect the cost and quality of the data produced. We approach initial efforts at design with a focus on increasing measurement and control of survey errors, and I want to address features of the system that so affect the research. My hope is to stimulate discussion of design alternatives and to offer some guidance to those currently making choices on system designs. I will ignore the processing of interviews by the interviewer; the movement from question to question is a part of all systems and is about the simplest part to implement.

Instead, I will concentrate my remarks on four aspects:

1. An overview of the system design from the researcher's vantage;
2. Presentation of putting up a questionnaire on the system;
3. Description of the sample control system; and
4. Procedures for monitoring interviewers' performance.

System design

We begin with Exhibit 1, which attempts to give an overview of the sets of activities within the system. The circles in Exhibit 1 can be viewed as separate programs linked together. In general, they are used sequentially. An interviewer signs on the system and then moves to MA to record the time (unfortunately, the system in its current edition does not have a clock function). This act signals the program SP to locate the group of sample numbers judged most likely to be answered. In SP the interviewer obtains a sample number in one of two ways: The interviewers may choose a number from their individual assignment list, or they may choose to have the machine give them a number.

After a sample number is obtained, the interviewer moves automatically to CL, where the number is dialed and a respondent selected. In our current research, one of three different versions of the questionnaire is randomly assigned to each sample number. These are called EX, PX, and CX. After respondent selection and agreement to the interview, the interviewer moves to the questionnaire program itself. After each dialing, the interviewer is returned to SP to record the result of the dialing.

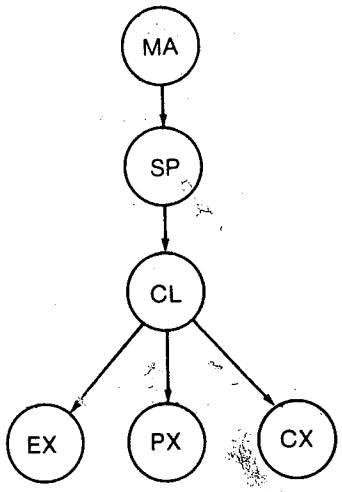
The advantages of this design are the following:

1. Since the system will most often be used for surveys on randomly generated sample numbers, not all numbers will require an interview data record. A hierarchical design permits some savings on disk space because records are stored only when needed.
2. The researcher needs to program only the lower level of the system; the higher levels are generally constant.

Our present implementation of the design also has disadvantages:

1. The movement between programs is time-consuming in its present state, slowing response time between screens.
2. Parameters must be passed between programs—for example, numbers relating to the selection of a respondent. This causes unnecessary work for the interviewer, who must enter them at the beginning of a program.

Exhibit 1
Program levels in SRC computer-based interviewing system



	Activities	*Function*
MA	1. Record day and time	Direct machine to appropriate queue of sample numbers
SP	1. Display cases assigned to interviewer, choice of cases 2. Permit machine assignment of numbers 3. Record results of calls	Storage of call records on all sample cases
CL	1. Choice of questionnaire type 2. Dialing of number 3. Introductory comments 4. Selection of respondent	Identification of working/nonworking numbers
EX, PX, CX	1. Questioning of respondent 2. Interviewer observations	Record of interview data

Putting up a questionnaire on the system

Next I will describe what steps the researcher must perform in constructing a questionnaire on the system. The system has been designed to permit the research staff to do this work. Exhibit 2 shows part of a questionnaire that we are currently using. It looks different from most hard-copy questionnaires:

1. There are lines around questions, indicating screens. The system that we are using is screen-based rather than line-based. The display does not scroll through lines in a questionnaire. Instead, entire screens are flashed on the terminal. The boxes in Exhibit 2 are meant to describe screens that the interviewer would see. Thus, the first act of the researcher is to conceive of the questionnaire as a set of screens.
2. There are no "Go To" or "Skip To" instructions in the questions. These are internal to the program.
3. There are comments made by the interviewer after some responses. For example, in A4 if the respondent immediately mentions a bad reaction to a medicine, the interviewer provides the feedback, "Thanks, that's useful information." This feature exemplifies the programmed interviewer behavior that is part of the current study.

We have found two forms of documenting a questionnaire helpful. One is a flow chart as in Exhibit 3, where boxes are question displays and directed lines indicate movement between screens. Numbers on lines indicate that the path is taken only if the previous response is the given value. (For example, if on A4 the answer is 1, "Yes, I have had a bad reaction to a medicine," the next question is A4A, "What medicine was it?")

The material below the flow chart describes the sequencing of displays. Each screen is listed with the default screen that would be displayed next, the screen to return to if the interviewer wants to back up, and several other pieces of information. The research staff is responsible for the material in Exhibits 2 and 3.

After these characteristics of the program are defined, we are ready to place the questionnaire into the machine. Exhibit 4 demonstrates how this is done. The typist who is putting up the questionnaire indicates what portion of the job to do for a particular screen labeled STRT. If this person chooses 1 on the first screen in Exhibit 4, he/she will be taken to the second screen in the exhibit. This display describes what portion of the screen is free for display of material. All places that are enclosed by asterisks cannot be used. The top line is used throughout the program to display information

Exhibit 2

A3F. Ever broken any bones?
1. YES **F**
5. NO **F**

A3G. Ever had diabetes?
1. YES***
** Thanks.**
5. NO ****

A4. For the next few questions you may have to think especially hard to remember everything.

Have you ever had a BAD REACTION to any MEDICINE?
1. YES **F**
2. SPONTANEOUS MENTION **Thanks, that's useful information.**

4. QUICK NO
5. THOUGHTFUL NO

90

A4A. What medicine was it?

(IF NAMED): **Thanks, that's useful information.**

A4B. As I mentioned, sometimes it's hard for people to remember everything. Perhaps if you think about it a little more you will remember some things you have missed. Was there anything at all, even some minor reaction to any medicine?

(IF YES): Uh-huh, I see, What medicine?

(IF NAMED): **Thanks, that's useful information.**

A4C. Was there anything at all, even something minor?

(IF NAMED): **Thanks, that's useful information.**

A4D. Have you ever had a bad reaction to any OTHER medicine?

(IF YES): What medicine was it?

A5. Next, we want to find out what you have felt or noticed in different parts of your body at any time during the last two weeks. These may be things you noticed for the first time or things you also noticed before. First, thinking about your head, neck, shoulders, and arms; please tell me ALL the things which you have felt or noticed there in the last TWO WEEKS.

1. ONE MENTION **F**
2. TWO OR MORE MENTIONS **Thanks, that's useful information.**
5. NO MENTIONS

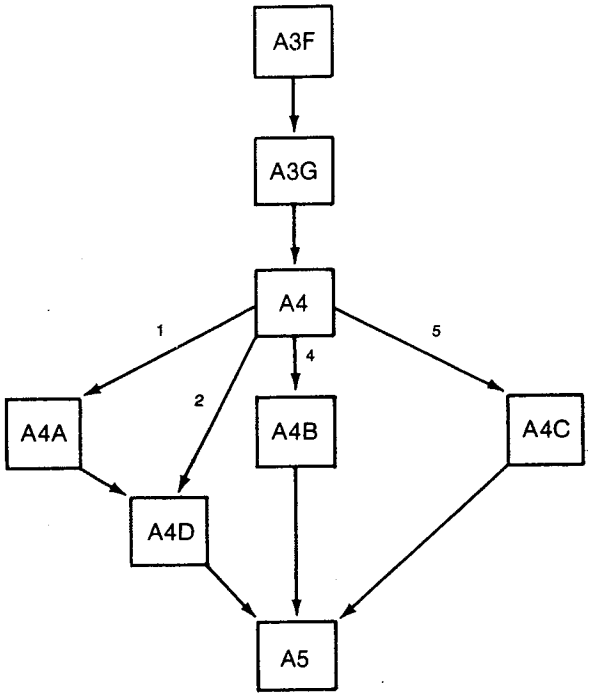
to the interviewer. The ten asterisks on the left of the middle of the screen are spaces for the numeric responses. The six lines at the bottom are reserved for text of open-ended responses. The typist places on this screen the question wording and response categories of each question, one screen per question.

The typist is then returned to M005 and can choose to enter the default sequencing. The next three screens in Exhibit 4 describe what the typist does. The last screen in Exhibit 4 is the format for the entry of conditional sequencing, movement to some screen other than the default next screen, based on the value of some previous screen.

When these steps have been completed for

Exhibit 3

Flow chart for question series A3F–A5



Defaulting sequencing, and screen description

Screen Name	Question	Next Screen	Last Screen	Screen Type	Field Width	Valid Codes
A3F	Ever broken bones?	A3G	A3E	4	1	1,5,9
A3G	Ever had diabetes?	A4	A3F	4	1	1,5,9
A4	Bad reaction to medicine?	A5	A3G	1	1	1,2,4,5,9
A4A	What medicine?	A4D	A4	3		
A4D	Any other?	A5	A4	3		
A4B	Think . . . anything?	A5	A4	3		
A4C	Something minor?	A5	A4	3		
A5	Head, neck, etc.?	A6	A4	4	1	1,2,5,9

Conditional sequencing

Screen	Question	Condition No.	Condition Logic
A4	Bad reaction to medicine?	01	IF A4 EQ 1 THEN A4A
		02	IF A4 EQ 2 THEN A4D
		03	IF A4 EQ 4 THEN A4B
		04	IF A4 EQ 5 THEN A4C

each screen, the program for the questionnaire is complete and can be run immediately.

The advantages of this design are the following:

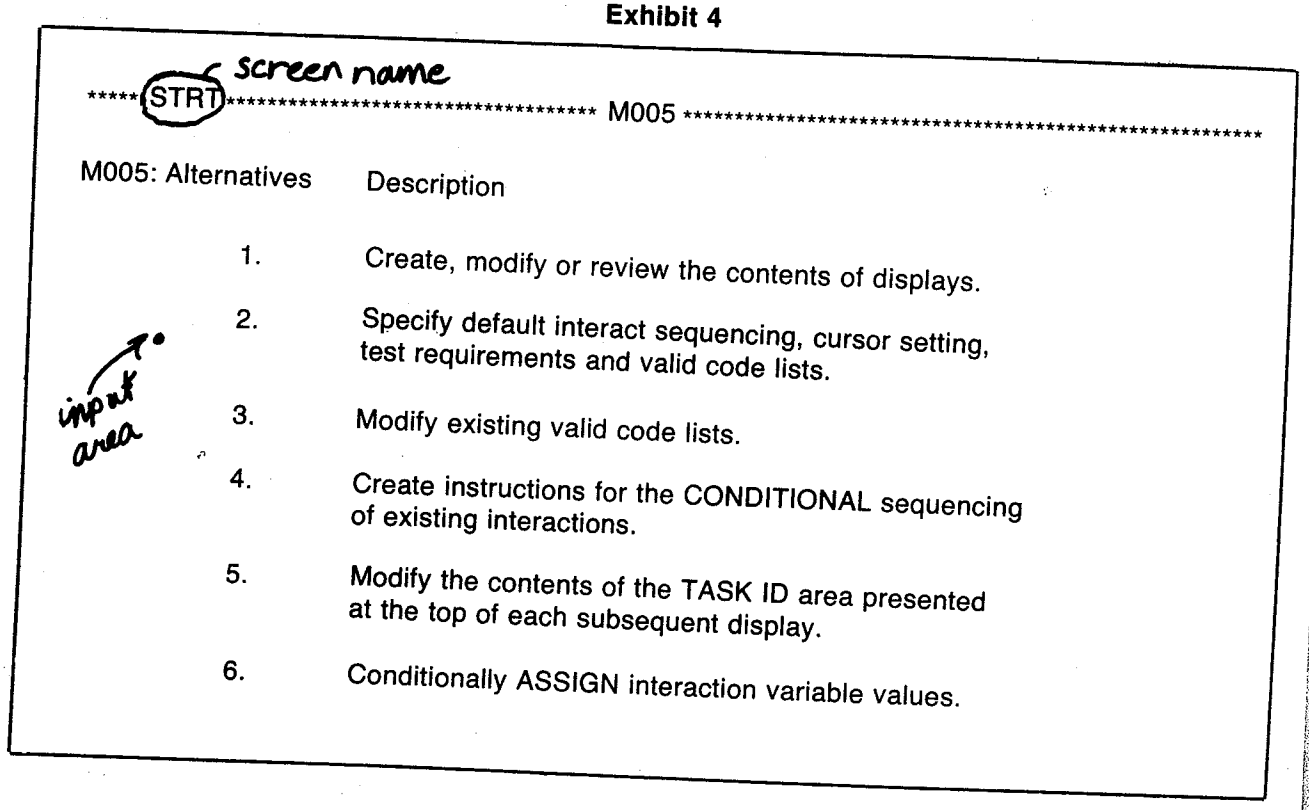
1. Flexibility—the procedures successfully reflect what happens as questionnaires receive revision during pretesting. A change can be made in question wording; within seconds the revised program can be used.
2. Interactive, simple procedures—clericals can construct the final instrument, much as they might format and type the final form of a hard-copy questionnaire.

The disadvantage is that the costs of the ease of alteration are offset by the machine having to interpret more code at the time of interviewing. Response time becomes an issue. At the current time we are experiencing median response times of 3 seconds on a 12-terminal system.

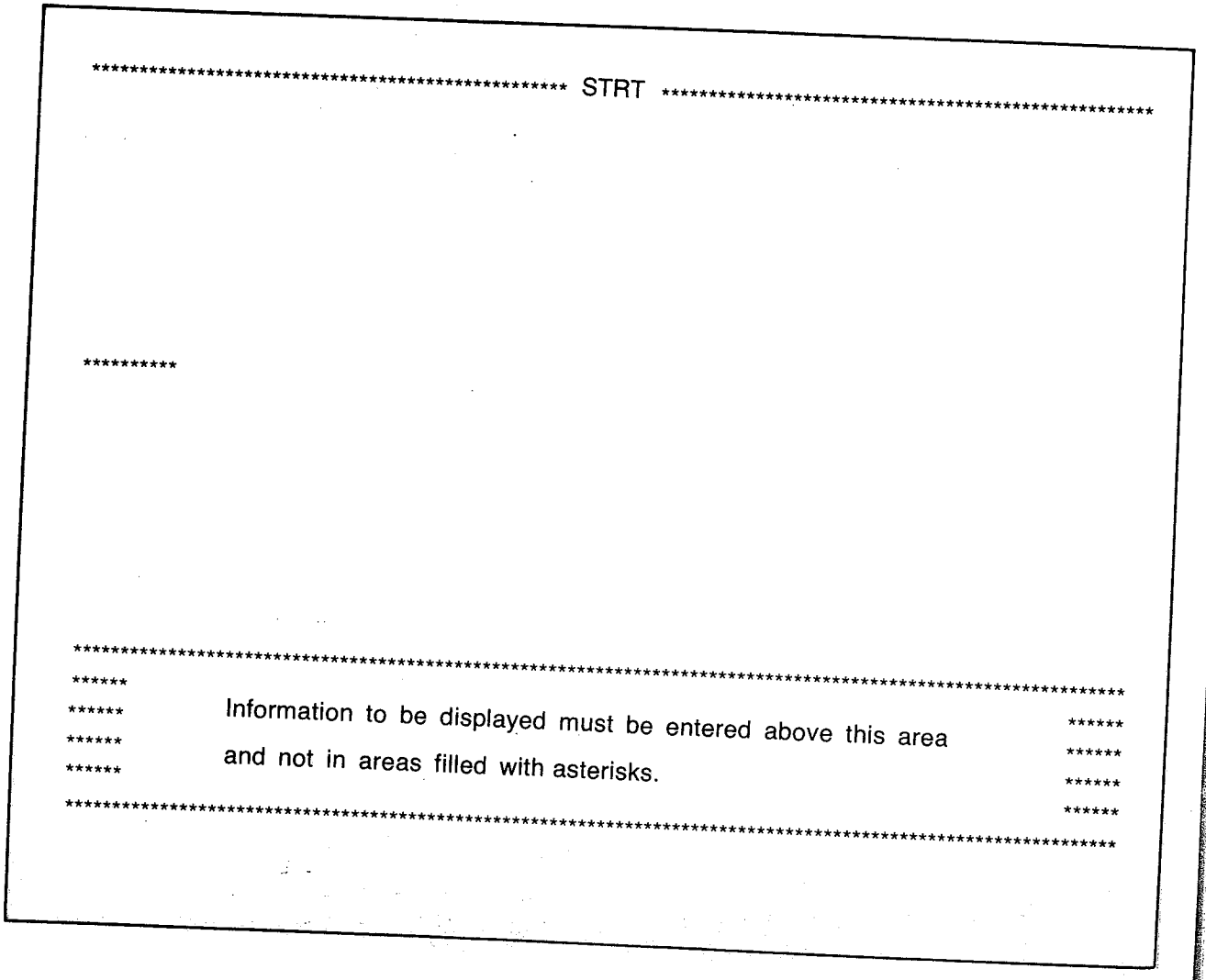
Sample control system

Exhibit 5 presents two screens that the interviewer sees after recording the result of each dialing. The first lists all cases assigned to that particular interviewer. For the example shown, the case number 9028 is assigned to the inter-

Exhibit 4



OPTION 1 CHOSEN ON M005



OPTION 2 CHOSEN ON M005

STRT

T003:
Please enter the name of the NEXT display to be presented. NEXT is the name of the display to be presented if no conditional sequencing instructions override this default value.

****

STRT

T004:
Please enter the name of the display to be returned to if the PF2 key is depressed. This is normally the LAST display which must have been viewed no matter which path through the program has been followed to this point by the user.

****

STRT**_____ TASK DESCRIPTIONS _____

0. Information is only to be displayed. No data will be entered.
1. Both text and numeric responses are expected. The text response is expected to be entered first followed by the numeric response. The curso is initialized to the text area. Text responses redisplayable as appropriate.
2. Only a numeric response is expected. If text is entered, it is treated as comments, i.e., it cannot be redisplayed. The cursor is initialized to the numeric response area.
3. Only a text response is expected. The numeric response area is ignored. Text is redisplayable and the cursor is intialized to the text area.
4. No. 4 is the same as No. 1 except the numeric response is expected first and the cursor is initialized to the numeric area.

OPTION 4 CHOSEN ON M005

STRT Please enter the following information to effect the conditional sequencing of interactions

SKIP # . .
Enter 1 to create or modify a sequencing instruction or
2 to delete an existing instruction.

IF
NAMExxyyyy
. . . . . .WHERE:
NAME is any existing display name and,
xx is one of the relations EQ (equal to), NE (not equal to), GT (greater than), LT (less than), LE (LT or EQ), GE (GT or EQ) and
yyyy is a legitimate value of a NAME and must be entered right justified with leading zeros.

THEN SKIP TO
. . . .

ELSE GO TO . . . .

Exhibit 5

```
                    MA 002

          Listed below are the cases currently assigned to
          MA 002 for the task SP to be invoked next

CASE State
9028  1001 APPT

               To process an assigned case enter its Case ID.
               To scroll through assigned cases depress ENTER repeatedly.
               To obtain another case assignment scroll to the end of the current assignments.
               To restart scrolling at the beginning of the assignment list depress PA2.

                                  Enter ASSIGNED Case ID . . . .
```

```
                    MA 002

               Listed below are the functions currently assigned to
               MA 0002 for the task SP to be invoked next.

STATE

   01 1 STRT Never been answered numbers
   02 1 STRT Six Call Numbers
   03 1 STRT Noninterview Callback Numbers
   04 0 SPRV Supervisor Review Numbers

               To perform an assigned function enter the appropriate State ID.
               To scroll through assigned states depress ENTER repeatedly.
               To restart scrolling at the beginning of the assignment list depress PA2.

                                  Enter ASSIGNED State ID . .
```


viewer, and the designation to the right of the case ID shows that it is an appointment. If the interviewer entered 9028 at the bottom of the screen, that would become the active case.

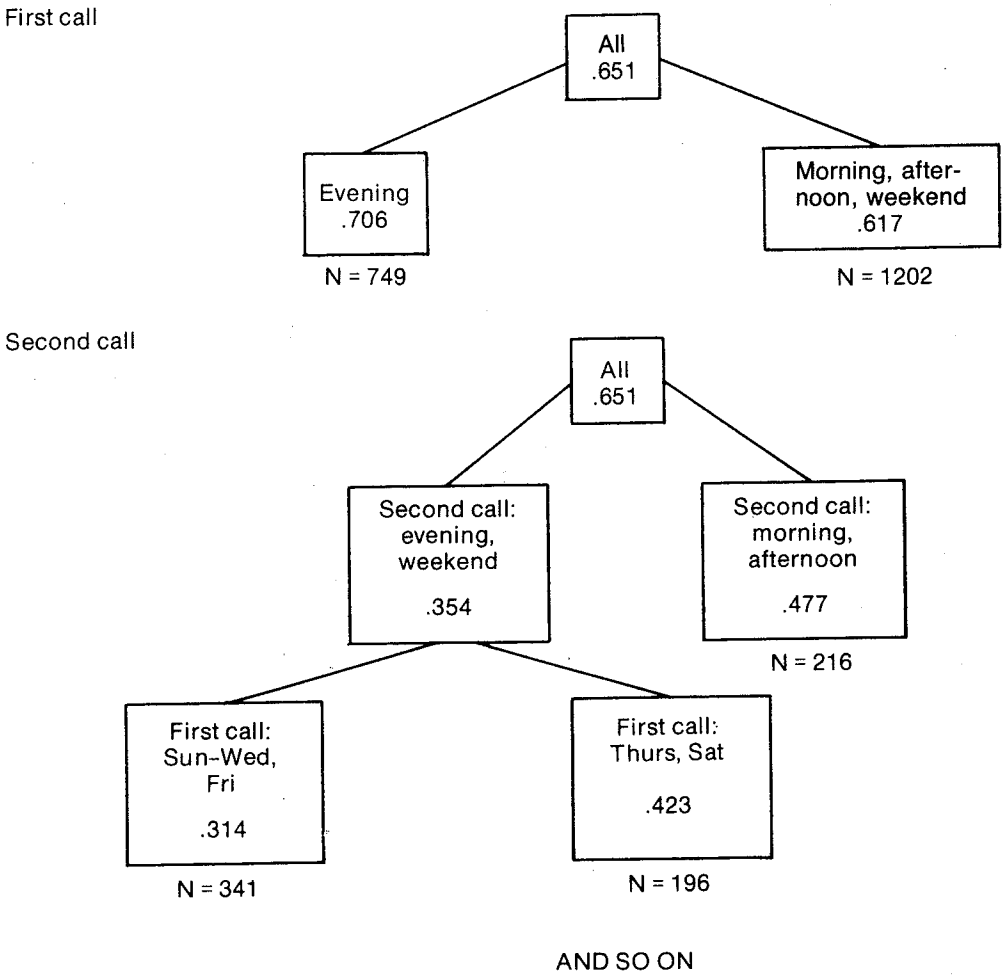
If the appointment time had not yet arrived, the machine could assign a case to the interviewer. The next screen lists alternative sets of numbers that are candidates for machine assignment to interviewers. In most cases, interviewers will concentrate on State 01, those sample numbers that have not yet been answered.

We have found that the software design and implementation for the questionnaire programming is straightforward relative to the programming required for the sample control and administration.

The strategy that we employed analyzed call records for a year's worth of surveys conducted at SRC. Exhibit 6 presents AID trees on one-quarter of the data. We used this analysis to calculate probabilities of a number being answered, given a past history of calling. The AID algorithm defines groups so as to maximize the between sum of squares for the groups. The trees can be read by looking at the contents of the boxes. For example, 70.6 percent of the numbers called in the evening on a weekday received an answer on the first call; 31.4 percent of the numbers with a first call on Sunday–Wednesday or Friday and a second call on the evening or weekend were answered.

We performed analyses like these for numbers that were answered on call 1, call 2, call 3, etc. After that, we defined priority groups of

Exhibit 6
AID trees for callback algorithm



numbers based on the likelihood that they would be answered on the next call on a particular day and time. These groups of numbers are termed *queues*. Exhibit 7 shows how the sample numbers are distributed across the queues; at any particular moment, each number is in one and only one queue.

The numbers in these queues have had the same calling patterns. Part B of Exhibit 7 shows how the calling on the sample numbers is directed by the empirical results of the AID analysis. On Monday mornings, numbers in queue number 03 are called first; when those have all been called, those in queue 01 are called and so on. Monday afternoon, a different priority pattern obtains.

The intent of this design is to ensure that at any one moment interviewers are dialing numbers that have the highest probability of being answered. We cannot yet evaluate whether we are reducing the number of calls required on each sample number for disposition. We have discovered, however, that late in a survey the sample administration needs to take on a different strategy. The supervisor needs to scan the contents of each queue of numbers and make individual decisions on future calls.

Exhibit 8 merely notes what functions supervisors may perform as part of their job. In addition, a supervisor can review the calls on each number and any interviewer notes describing the results of the calls.

Procedures for monitoring

We are now performing two kinds of monitoring of interviewers' performance. First, supervisors can see a copy of each interviewer's display and can monitor parts of the interview that are causing problems for a particular interviewer. After the interaction is complete, the supervisor provides feedback to the interviewer regarding his or her performance. Another type of monitoring that is ongoing we have labeled "statistical monitoring." The purpose of this work is to collect empirical data measuring

Exhibit 7
Graphic representation of queuing system

A. States and queues

State	Queue	Case ID's
01	01	2412, 0976, 0844, 1122, . . .
	02	1144, 2314, 1523, 0753, 0932, . . .
	.	
	.	
	.	
	99	0234, 1125, 1342, . . .
02	01	1223, 0846, 1634, 0764, . . .
	.	
	.	
	.	
03	01	1223, 1332, . . .
	.	
	.	
	.	

B. Queue priority lists for State 01

1. Monday, 9–12	03, 01, 19, 04, 05, . . .
2. Monday, 12–5	01, 04, 02, 03, . . .
3. Monday, 5–9	03, 02, 01, 04, 10, . . .
4. Monday, 9–12	01, 03, 04, 02, . . .
5. Tuesday, 9–12	03, 01, 04, 05, . . .

Exhibit 8
Supervisory functions within the system

1. Review contents of queues of sample numbers
2. Review contents of individual interviewer's assignment lists
3. Move a sample number from one queue to another
4. Activate priority lists for sample numbers

interviewer behavior on a subset of interviews chosen using randomization rules. We want to use these data in analyses to provide insights into the correlates of interviewer variance.

At present, this work is done with paper-and-pencil coding using the categories listed in Exhibit 9. The monitor scans the screens to locate interviewers in sections of their work eligible for monitoring and uses the forms shown in Exhibits 10 and 11 to record interviewer behavior.

We are taking crude first steps, but this technique may hold promise for investigating sources of error related to interviewer behavior. What results from this procedure is a vector of interviewer behaviors associated with each piece of data in a data record.

Both survey costs and survey errors are proper concerns of methodologists; reducing costs frees money to measure and control errors. We have discovered very close relationships between two pairs of variables:

Magnitude of tasks performed by machine

Magnitude of tasks performed by interviewer

Decisions on these affect response time, *or*

Exhibit 9
Codes for monitoring interviewer behavior

Question-asking	11	Reads question exactly as printed
	12	Reads question incorrectly—minor changes
	16	Reads question incorrectly—major changes
	17	Fails to read a question
	18	Reads inappropriate question (due to prior miscode)
Repeating questions	21	Repeats question correctly
	25	Repeats question—unnecessarily
	26	Repeats question—incorrectly
	27	Fails to repeat question
Defining/clarifying	31	Clarifies or defines correctly
	35	Defines or clarifies—unnecessarily
	36	Defines or clarifies—incorrectly
	37	Fails to define or clarify
Short feedback	41	Delivers short feedback—correctly
	45	Delivers short feedback—inappropriately
	46	Delivers short feedback—incorrectly
	47	Fails to deliver short feedback
Long feedback	51	Delivers long feedback—correctly
	55	Delivers long feedback—inappropriately
	56	Delivers long feedback—incorrectly
	57	Fails to deliver long feedback
Pace/timing	65	Reads items too fast or too slow
	66	Timing between items—too fast
	67	Timing between items—too slow
Overall clarity	75	"Unnatural" manner of reading item (poor inflection, exaggerated or inadequate emphasis, "wooden" or monotone expression)
	76	Mispronunciation leading to (possible) misinterpretation
Machine-related	96	Lag due to backing up in questionnaire
	97	Slow machine response time

hardware (computing power needed), *or* number of terminals supported.

Least sensitive of the machine functions in affecting response time seem to be valid code checks and conditional logic for skipping questions; the functions needed for most surveys do not present large burdens. Most sensitive are decisions regarding data base management—whether the sample is administered by machine, whether interview data are constructed to be immediately accessible, whether monitoring data are recorded on-line.

Some measurements of survey error are simple, inexpensive by-products of such a system (e.g., question-wording experiments); other measurements may create significant increases in the load on the machine (e.g., on-line reliability monitoring of interviews with machine storage of monitor data).

Those involved in these developments must spend the coming years evaluating alternative approaches to computer-assisted data collection. The promise of measurement and control of survey errors through this technology will only be fulfilled with simple and cost-efficient procedures.

Exhibit 10

Monitoring form

Interviewer's number _____

Case number _____

Questionnaire type _____

Monitor _____

Date _____

Time _____

Check which section monitored: section one _____

section two _____

section three _____

section four _____

Did you give feedback after this monitoring? yes _____

no _____

Was there a lag time for you between this and your last monitoring?

yes _____ no _____

IF YES, was it due to:
waiting to listen to a specific segment? _____

waiting to hear a specific interviewer? _____

no one interviewing at the time? _____

unable to find an eligible segment? _____

other? _____

How much time did you spend waiting to monitor this segment?

less than 5 minutes _____

between 5 and 10 minutes _____

between 10 and 15 minutes _____

more than 15 minutes _____

Exhibit 11

P468222

Case Number_____

Section I

Screen Name	missed screen?	CODES	Extra Comments
A2A			
A2B			
A2C			
A2D			
A2E			
A2F			
A2G			
A2H			
A3A			
A3B			
A3C			
A3D			
A3E			
A3F			
A3G			
A4			
A4A			
A4B			
A4C			
A4D			
A5			
A5A			
A5B			
A5C			
A6			
A6A			
A6B			
A6C			

Observations on the behavior of automated telephone interviewing

D. Garth Taylor, National Opinion Research Center, University of Chicago

I. Observations on Interviewer Behavior
   A. There are no consistent rules on what makes a good telephone interviewer.
   B. There are large interviewer differences in response and refusal rates.
   C. There is high turnover among telephone interviewers.
   D. Attrition may be due to exhaustion. Because of the way they are fielded, it is cost-effective to increase the number of interviews done by each interviewer.
   E. Interviewers develop their own speed and style of interviewing.

II. Observations on the Advantages of Computer-Assisted Telephone Interviewing (CATI)

Advantage #1: CATI can reduce the time required for questionnaire preparation and for data processing.

> CANON #1 of CATI Data Processing: The visibility of a task is inversely proportional to the level of effort involved in solving the problem.

Advantage #2: CATI can lead to improvements in survey management.

> CANON #2: It will often be advisable to break out some of these tasks as batch computer operations. This will involve a separate staff of computer operators with the associated errors in communication caused by building one more step into the critical path.
>
> LEMMA 2A: In budgeting a CATI project, it is wisest to have a senior computer operator available during all times that you are dependent on machine performance.

Advantage #3: CATI can lead to improvements in the quality of data collected.

(a) Supervisors can directly monitor the interviewers.
(b) Interviewers can display the question-by-question instructions at any time.
(c) Interviewers can display the names, face sheet information, or any other kinds of respondent information at any time.
(d) Because of the automated checking procedures, there are fewer imputations and callbacks required for missing, out-of-range, or inconsistent information.

> CANON #3: CATI does not guarantee clean data. Rather, there are tradeoffs in how clean you want the data and at what cost in terms of programmer time, interview waiting time, and the risks of jeopardizing other aspects of the quality of the interview.
>
> LEMMA 3A: CATI does not eliminate coding costs.

Advantage #4: CATI enhances the ability to do methodological research.

(a) Randomize question forms, order of responses within questions, the order of questions.
(b) Randomize and otherwise control the verbal behavior of the interviewer.
(c) Use the timing features of the computer to regulate the pace of the interview, the amount of time allowed for probing, etc.

Advantage #5: CATI allows for greater flexibility in questionnaire construction and in questionnaire administration.

(a) More flexibility and creativity in dealing with open-ended information.
(b) Can tailor wordings to respondents.
(c) Can achieve a fuller integration of measurement theory with data collection procedures by (1) administering a short scale first and then using a branching procedure to question further in the region of the true score; (2) using a

strategy of overlapping measurements and missing data routines so that a long questionnaire can be administered to a *population* but each member of the population is asked to respond to a much shorter block of questions.

CANON #4 (also known as the Golden Rule of Data Processing):

(a) Do not underestimate hardware requirements.
(b) Do not underestimate the time required to develop programs.
(c) Simplify your requests.
(d) Start with small projects, small innovations, and/or small experiments.

Applying health interview techniques to mass media research

Peter V. Miller, Institute of Communications Research, University of Illinois at Urbana-Champaign

Introduction

This paper reports some results of the first application of the interviewing procedures developed in health studies by Cannell and his colleagues—instructions, feedback, and commitment—to information gathering on nonhealth topics. The subject of this study is use of the mass media, an area that presents some interesting opportunities for methodological research, as well as some formidable problems. We will explore both sides of the matter.

The implications of trying to apply the experimental interviewing procedures to a new substantive area are wider ranging and more profound than they might seem at first. For the study director, the overall import of the attempt is that he or she must assume much more responsibility for the measurement that goes on in the field than is normally the case. The new procedures essentially make the questionnaire a script, which interviewers read with little or no "ad-libbing." Embodied in the script are the directions, feedback, and exhortations to good reporting that are intended to increase the validity of the information collected. By adopting these procedures, one eschews the practice of relying on the interviewer to provide the "contextual communication" in the interview that surrounds the questions. The interviewer, by following the scripted questionnaire, becomes much more an extension of the study director and puts into practice his or her conceptions of good measurement.

The question of what is good measurement is not always easy to answer, however. In many ways, the health area was ideal for formulating the new interviewing techniques because ideas of what constitutes valid reporting are more firmly grounded here than in other substantive areas. The practice of conducting "validity studies" to identify response error in major health variables created the conditions within which the experimental interviewing procedures developed. External information or good assumptions about reporting error are needed to effect any "improvements" in data collection techniques; validity studies greatly helped to provide this key information. As we shall see, things become much more complicated when we try to measure behaviors and attitudes that present largely unknown reporting problems.

An overview of the techniques

Cognitive and motivational problems beset the respondent in any survey interview because of the demands of the question-answering task. Some of the manifestations of these underlying difficulties are under- and overreporting of events and experiences and social desirability bias in attitude reports. The interviewing procedures developed by Cannell and his colleagues are designed to attack the underlying problems and are based on the notion that "typical" interviewing practices exacerbate rather than ease the respondent's difficulties. For example, standard interview procedures do not inform respondents well about the goals of the survey, nor about their responsibilities in connection with particular questions (Cannell, Oksenberg, and Converse, 1977). The feedback provided respondents by interviewers is just as likely to reinforce poor response behavior—e.g., cursory answers or refusals to answer—as it is to reward the respondent's hard work (Marquis and Cannell, 1969). Tape recordings of interviews reveal that the interviewers' reinforcement utterances and probes could be much more appropriately utilized to increase response validity. Another problem is that respondents often tolerate the interview but do not have enough psychological investment in it to report embarrassing events or things requiring considerable memory work. Interviewers vary considerably in their ability and willingness to obtain hard work from respondents, so that a standard

technique to increase respondent motivation seems desirable.

In summary, the interviewing procedures discussed in this paper are intended to give the respondent better information on how to perform adequately in the interview, to motivate him to expend the effort required to reduce the effects of reporting biases, and to provide positive feedback when accurate and complete information is obtained. Following is a brief description of each of the techniques. For more information, see Cannell et al. (1977), Miller and Cannell (1977), or U.S. National Center for Health Services Research (1977).

Instructions. Despite the oft-discussed proliferation of surveys, being interviewed is still an unusual experience for most people, and some orientation to the kind of behavior required is apt to be useful for the naive respondent. *General* instructions (addressed to reporting behavior in the entire interview) that have been employed in the research reviewed here emphasize that the information sought should be as accurate and complete as possible. They also suggest approaches to the interview that will aid in fulfilling the demands of the respondent role, such as telling respondents to think carefully, take their time, give exact answers, and give as much information as possible.

Question-specific instructions seek to alleviate, or at least make sensible, the response demands of given items. For example, when the objective is a report of the frequency of a particular experience (say, doctor visits or movie attendance), we might emphasize in an instruction just prior to the question that exact information is needed on the following item. In other words, ranges or approximations are not sufficient to fulfill the task posed by the question.

On open questions, instructions have emphasized the specificity of the desired information, or its completeness. The information has proved helpful in avoiding abbreviated, general answers to questions on symptoms of illness and the names of television programs viewed the day before the interview, for example. Behavioral instructions complement these goals by telling the respondent to take time in answering and to think carefully.

It is also possible, and often desirable, to instruct respondents *not* to think carefully—that their first impression is all that is required. Whatever the approach, the point is to tell the respondent *something* about the response demands in the question and the sort of mental process required to deal with them. The instructions then form the basis for reinforcement of appropriate behavior.

Feedback. When response objectives are clearly stated in the instructions, it becomes possible to standardize the kinds of feedback that interviewers provide to respondents and to make the feedback contingent on how well the respondent meets the question objectives. As we pointed out earlier, interviewers' communication about respondent performance is apt not to be dependent on how the respondents perform. In part, this reflects a lack of clear objectives for respondents in the typical survey interview; also, interviewers often are not trained in how to differentiate between poor and adequate response behavior. The result is that an interviewer expression of approval (e.g., saying "OK" or "Uh-huh" or making positive references to responses) is nearly as likely to follow a refusal to answer, or a hasty, incoherent, or incomplete answer, as it is to follow a more considered response. The idea of maintaining "rapport" with respondents also encourages the practice of noncontingent feedback, since, following this dictate, one does not want to alienate the respondents by, in effect, challenging their performance. Secondary or "probe" questions—another sort of feedback—are apt to be used in very different ways by different interviewers and also on a noncontingent basis.

In the research discussed in this paper, interviewer performance was controlled in the experimental interviewing conditions by providing alternative feedback responses for different types of respondent behavior. When the respondents fulfilled the requirements set out in the question and instructions, the interviewer would provide a positive comment—anything from "OK, I see" to "Thanks, this is the sort of information we want to get." For a response that did not meet question demands, the interviewer would repeat the original question. In order to make this work, the study director must have a good idea of the range of possible answers that respondents can render for a given question, so that the appropriate feedback responses can be built into the instrument. Further, interviewers must be trained to distinguish between good and bad response behavior and to follow the script without deviation. "Diagnostic probes" have been utilized by interviewers in cases where it is not clear initially whether the respondent was giving an appropriate response. For example, the interviewer might ask, "How do you mean that?" or "Could you tell me more?" in cases where the response is possibly irrelevant or incomplete. Interviewers, then, do not become automatons when using the scripted questionnaire; they must accurately perceive responses and utilize the correct parts of the script. The difference between this procedure

and a typical survey interview is that the interviewer does not have to devise the rules for communicating in the interview as well as follow them. Exhibit 1 features examples of the instruction-feedback linkages, as employed in health and media use studies.

Commitment. Apart from understanding aspects of the respondent role, the person being interviewed must have sufficient motivation to meet the demands of the role. It is very easy for respondents to fall into a pattern of "just giving answers" to questions and not treating the interview as an event worthy of attention. In order to combat this tendency, a technique has been developed by Cannell and his colleagues that involves asking respondents early in the interview to commit themselves publicly to rendering considered, accurate answers. See Exhibit 2 for an example of the commitment procedure and form, which both respondent and interviewer sign. The anonymity of the respondent is preserved by giving the form to the respondent to keep after it is signed. Respondents are informed that the interview cannot continue unless he or she signs or initials the form. In the several studies done using the procedure, almost no one has refused to sign.

As we mentioned earlier, all of these efforts to standardize and improve the interviewing experience rest on the assumption that we can discern when "better" answers are provided by respondents. We also must assume that considerable work has gone into making the questions as problem-free as possible. It makes no sense to employ new interviewing techniques if the questions pose hopeless response problems. The strategy of employing new interviewing techniques, instead, recognizes that even good questions are subject to various sorts of measurement error and that the error can be reduced by motivating the respondent, explaining the question demands, and reinforcing good response behavior.

Issues of response validity in media use data

Cannell and his colleagues have focused on the impact of interviewing techniques on under- and overreporting biases that are attributable to particular types of respondent burden. Questions that ask about nonsalient or embarrassing matters, or that present difficult memory tasks, are likely to be characterized by underreporting, while socially desirable behaviors are likely to be overreported. See Cannell et al. (1977) or U.S. NCHSR (1977) for some examples of findings on these sorts of biases in health data.

What hypotheses should we entertain concerning response behavior about media use?

Exhibit 1
Examples of instruction-feedback procedure

Sickness experiences

Q1. *Let me just mention that to be most accurate you may need to take your time to think carefully before you answer.* (PAUSE) Have you been sick in any way within the last two weeks?

IF R SAYS YES	1a. In what ways were you sick? 1b. *Uh-huh, I see. This is the kind of information we want.* Were you sick in any other ways within the last two weeks?
IF R SAYS NO WITHIN 5 SECONDS	1c. You answered that quickly. Were you sick in any way at all in the last two weeks? 1d. (ANY MENTION) *Thanks, this is the kind of information we want.*
IF R SAYS NO AFTER 5 SECONDS	1e. Were you sick in any way at all within the last two weeks? 1f. (ANY MENTION) *Thanks, this is the kind of information we want.*

Television watching

Q2. *On this next question, we'd like to get numbers as exact as possible.* How many hours did you personally spend watching television yesterday?

EXACT NUMBER	2a. *I see . . . thanks.*
APPROXIMATION	2b. Can you be any more exact about the number of hours? 2c. (EXACT NUMBER) *I see . . . thanks.*
NO RESPONSE, DON'T KNOW	2d. Let me repeat the question. How many hours did you personally spend watching television yesterday? 2e. (EXACT NUMBER) *I see . . . thanks.*

Exhibit 2
Commitment procedure example

That's the last of this set of questions. The rest of the questions are on how media like newspapers, TV and radio fit into your daily life. We are asking people we interview to give us extra cooperation and that they try hard to answer accurately so we can get complete and accurate information about this topic. You are one of the people we hope is willing to make the extra effort.

Here is an agreement which explains what we are asking you to do. (HAND AGREEMENT) As you can see, it says, "I understand that the information from this interview must be very accurate in order to be useful. This means that I must do my best to give accurate and complete answers. I agree to do this."

We are asking people to sign this agreement and keep it for themselves so that we can be sure that they understand what we are asking them to do. It is up to you to decide—if you are willing to agree to do this, we'd like you to sign your name here (POINT OUT LINE). Down below there is a statement about confidentiality and I will sign my name here (POINT OUT LINE).

(IF R HAS NOT ALREADY SIGNED) Are you willing to make the extra effort to continue the interview?

AGREEMENT

I understand that the information from this interview must be very accurate in order to be useful. This means that I must do my best to give accurate and complete answers. I agree to do this.

_____
Signature of Respondent

All information which would permit identification of the people being interviewed as a part of this project will be held in strict confidence. No information that would allow identification will be disclosed or released to others for any purpose.

_____
Signature of Interviewer

Some intuitive predictions come readily to mind. It seems clear that much of mass media use would be difficult to remember, since it is often a secondary activity, performed while other actions (housework, conversation, or traveling, for example) are occurring. In other words, we may have a problem of salience for some types of media use, which may lead to underreporting. Such is likely to be the case for radio listening and, often, for television viewing. For other kinds of media contact, we encounter social desirability problems, which may be expressed in either under- *or* overreporting. People may be unwilling to admit that they have attended a pornographic movie, and they are likely to overestimate the number of books that they have read in the recent past or their reading of the newspaper editorial page.

These predictions rest solely on assumptions, since there is little or no "validity information" on these behaviors, and collecting such information would present extraordinarily difficult problems. Further, there is some dispute about my guesses on underreporting bias, at least with respect to television viewing. It is instructive to review this case, for it raises salient issues involved in employing the experimental interviewing techniques in this substantive area.

Probably the most common type of measure of television contact is the amount of time spent viewing. Robinson (1977), comparing data from the Roper (1971) recall measure on time spent viewing TV with estimates from reports of viewing elicited in time-budget diaries, has argued that recall measures overestimate the actual amount of viewing. The diary measures, which require the respondent to write down all activities during the course of one day in hourly segments, reveal considerably less time spent viewing than does the Roper measure, which asks respondents to estimate how much time they spend viewing TV on a "typical" day.

Robinson points out that "Roper-type" time estimates for all activities would sum to over 24 hours for the single day. Generally speaking, he argues that respondents will overestimate how much time they spend on *any* activity, if asked to provide an estimate in recall fashion. As reasons for the overestimation tendency, he cites the fact that respondents have difficulty in constructing accurate estimates of time spent on a particular activity or in understanding what behavior is to be included under the heading of a given activity.

It is important to note for our purposes, however, that one major reason recall estimates of time spent on activities for a day total over 24 hours is that people often engage in more than one activity at a time. When they are asked to recall how much time they spent on any one of these behaviors, the estimates are bound to overlap with those for other activities that were

undertaken jointly. For example, there is a large amount of "secondary" activity time devoted to various sorts of mass media use (viz., reading while eating, listening to the radio while commuting, watching TV while ironing). Therefore, the "inflated" estimate for television viewing in the recall measure could include a goodly amount of "secondary viewing." It seems clear that when Robinson speaks about television viewing time being overestimated, he is referring to what might be called "attentive" or "primary" viewing. For example, he buttresses his argument with evidence from a small "validity study" in which people were videotaped as they watched TV and then were asked to estimate how much time they watched. The findings from this study showed that people were often not watching the TV set when it was turned on and that they reported watching an entire program when they only watched part of it (Robinson, 1972).

If we are interested in only "primary" viewing of TV, we would expect, following Robinson, that a recall time-spent measure is subject to an *over*report. If we want to include both primary and sizable amounts of secondary viewing, however, the prediction would be for an *under*report, since the secondary activity is less salient and would be difficult to remember. (Further, there is a case to be made that the diary method—which Robinson treats as the more valid measure—is likely to *under*estimate secondary viewing, because it asks respondents to list other things that they were doing while engaged in a "major" activity, without adequate instruction on what to include under the heading of "other things.") Ultimately, the assumption about whether the time allocations are under- or overreported depends on the manner in which the measure is to be used. If we wish to predict information gain from watching TV, we want a measure of "attentive" viewing, and this is probably subject to overreporting. When we want to describe the extent to which TV "permeates" everyday activity, it is crucial to pick up "secondary" or "wallpaper" contact with the medium, and this is likely to be underreported. The problem is that most investigators simply measure "time spent" without instructions to the respondent (or a clear idea themselves) on whether attentive or secondary viewing is desired. This is precisely the sort of problem that the experimental interviewing techniques described in this paper are designed to alleviate, since they demand that the investigator make a careful, deliberate judgment about what it is he or she wants to find out and then communicate this to the respondent.

In this study (see Exhibit 1), we specified in the instructions for the questions on TV watching that an *exact* amount of time was desired (that the respondent should consider *all* contact with the medium, whether primary or secondary). This action was intended to combat the assumed tendency to underreport the time when TV was serving as "background noise" for other activities. Similar procedures were employed for radio-listening time estimates, following the same assumption.

For open questions that posed memory or salience problems, such as asking for details of newspaper articles, the titles of books read in the last few months, or the names and viewing times for TV programs watched the preceding day, the procedure was to ask for all information possible and to reinforce several mentions on the assumption that this sort of question is also subject to underreporting.

For the time-spent questions and the open questions, the hypothesis was that we would show higher mean reporting levels for the experimental condition. Similarly, we expected more reporting of X-rated movie attendance in the experimental condition, if the techniques worked to alleviate social desirability bias in this question.

Other questions in the study featured the possibility of *over*reporting, and for these we hypothesized lower levels of reporting in the experimental condition. This is the case for such questions as the number of books read in the last three months and a report of reading the editorial page of the newspaper "yesterday." Both of these questions are assumed to present a positive social desirability bias.

The study: design characteristics and limitations

Like earlier studies testing the effectiveness of the experimental interviewing procedures, the study described here involved an experimental design using a homogeneous population, in which some respondents experienced the new methods, while others had "control" interviewing procedures (all respondents were asked the same questions). This study departs from the earlier work, however, in the simplicity of its design and in the nature of the "control." Rather than dividing the subject population into separate groups to test the impact of each of the experimental procedures, there was only one experimental group, which received all of the new interviewing techniques in combination. Therefore, it is impossible to distinguish the individual contributions of each of the procedures or to tell which combination of them accounts for differences between the control and ex-

perimental conditions. To justify this design decision, we rely on the work of Cannell and his colleagues, who have found that each technique has a salutory impact on reporting and argue that another evaluation of the individual procedures is not as important as a test of their overall effectiveness as an integrated set of tools for improving communication in the interview. According to this view, instructions are not optimally effective without contingent feedback, and the commitment procedure should provide a motivational advantage in the respondent learning task.

The appropriate control group to compare with the treatment condition presents an interesting choice, since it involves a tradeoff between an optimal view of the effectiveness of the experimental techniques and a comparison of the techniques with "typical" interviewing practice. The health studies by Cannell and his colleagues featured a control condition in which the interviewers were instructed to say *nothing* besides reading the questions exactly as written. This approach affords the opportunity to view easily the effects of the various interview innovations, but the control condition certainly does not simulate the "typical" interview performance that the techniques are designed to improve. In the study reported here, the control condition was defined by giving the interviewers a modicum of basic training and telling them to "do their best" to get accurate information. In other words, we tried to set up conditions that were similar to those experienced by the majority of interviewers; they were instructed to read the questions exactly and slowly, to use appropriate voice inflections, to record accurately, and so forth, but they were allowed to provide whatever contextual communication (establishing "rapport," probing, etc.) that they saw fit to get good information. The comparison of this procedure with the experimental condition, which emphasized "sticking to the instrument," seems to offer a more realistic comparison of the new techniques and "typical" interviewing behavior than was featured in the health study designs.

The sample consisted entirely of adult white women from middle-class neighborhoods in Lafayette, Indiana. This homogeneous population is desirable for the interviewing study, since variation in actual experience is reduced, making differences between treatment conditions more attributable to *reporting* of experience. We further standardized experience by cluster sampling houses in groups of four and randomly assigning the dwellings to control and experimental groups, thus having two experimental and two control interviews possible in each cluster. A selection table was employed to select respondents in houses in which more than one adult woman lived. A total of 209 interviews were taken, with 102 in the experimental group.

Interviewers were assigned to specific addresses and were assigned to equal numbers of experimental and control interviews. The order of interviews in the different treatments was determined by respondent availability; the interviewers did not take all control or all experimental interviews before attempting the other type.

The interviewers for the study were graduate students participating in a class on survey research methods and graduate students and upper-division undergraduates studying interviewing methods. Because of limitations on the sample size and the atypical nature of the interviewer corps, we viewed this study as a pilot effort to understand how the experimental interviewing procedures would work in a nonhealth substantive area.

Findings

Differences in reporting between experimental conditions. Table 1 presents the basic differences found in reporting between interviewing conditions for the media use survey. Consider first the items for which the reporting task involves problems of recall. In reports of watching TV and the time spent with that medium on the day preceding the interview, we find that there was substantially more reporting of this experience in the experimental condition. Similarly, more time spent listening to the radio was reported by respondents interviewed with the experimental procedures (although the percentage reporting any radio listening did not differ between conditions). On the assumption that TV and radio exposure is often "background noise" and likely to be underreported, the differences between conditions suggest that the experimental interviewing procedures produce more valid reporting of these behaviors.

Some of our questions asked respondents to recall details of programs or newspaper articles, the names of TV programs viewed on different days in the week preceding the interview, details of news stories about different topics, and so forth. As with the time-spent estimates, we anticipated an underreporting bias for these questions and expected that the experimental techniques would produce higher levels of reporting than would the control condition. The results in Table 1, summarizing reporting differences between groups for several open questions,

Table 1
Effect of instructions, feedback, and commitment on various media use reporting tasks

Problems	Experimental condition	N	Control condition	N
Recall problems:				
Percent reporting TV watching "yesterday"*	86%	102	66%	107
Time spent (mean)*	197 minutes	88	157 minutes	71
Percent reporting radio listening "yesterday"	67%	102	65%	107
Time spent (mean)*	157 minutes	68	89 minutes	69
Number of mentions, 10 open questions (mean)*	30.71	102	22.39	107
Social desirability problems:				
Percent reporting editorial page reading "yesterday"*	38%	84	55%	70
Number of books read in last 3 months (mean)*	2.9	99	5.3	104
Number of book titles recalled (mean)	2.8	61	2.14	65
Percent reporting ever attending X- rated movie	61%	101	51%	106

*$p < .05$.

support our hypothesis.

The efficacy of the experimental procedures in producing more valid reporting would be somewhat suspect if they only produced higher mean scores for different items, given the fact that no validity information is available. One must also show that assumed *over*reporting biases are reduced by the techniques in order to claim that the overall product of the interviewing innovations is more valid reporting. Two tests of the effects of the experimental procedures on overreporting are presented in Table 1. Consistent with our expectation, the experimental group reported less reading of the newspaper editorial page and fewer books read in the last three months. Turning to a socially *undesirable* behavior, the experimental group reported more attendance of X-rated (pornographic) movies, demonstrating again the impact of the techniques on a probable underreporting problem.

Finally, notice in Table 1 that while the mean number of book *titles* recalled does not differ between conditions, the ratio of titles to the number of *books* reported read is substantially higher in the experimental group than in the control (.97 to .40). This supports the notion that the lower level of reported book reading in the experimental condition is due to the reduction in social desirability bias by the experimental procedures. When asked to support their claim about the number of books read, the respondents in the experimental condition more readily produced the titles of the books. Respondents in the control condition appear to have exaggerated their reading behavior; at least, they could not as often cite specific titles to back up their claims.

A closer look at the media usage differences. A variety of alternative explanations for the findings just presented suggest themselves. First, we must be certain that the interviewing manipulation is really responsible for the differences noted, and then we must add further support for the argument that the differences really reflect more valid reporting for those in the experimental condition.

The study design attempted to match the actual experience of the respondents who were randomly assigned to the different interviewing conditions by selecting a homogeneous population and clustering to gain even more homogeneity. Differences between the groups could then be more readily attributed to reporting rather than to actual behavior. But there is always the possibility that, despite the sample design and randomization, the experimental and control groups were very different sets of people to begin with and that the findings reflect these differences more than the effects of the experimental treatment. For example, the reports on time spent with television on the day preceding the interview would likely depend on how much time respondents had to spend at home. Women who work, attend

school, or are otherwise occupied outside the home would be expected to report less time watching TV than their counterparts who leave the house less often. If those occupied outside the home were disproportionately represented in one experimental condition or another, the estimates for TV time might be due to this factor rather than to the interviewing techniques.

Examining the demographic characteristics of the experimental and control groups, we found only slight differences in years of education and age. Respondents in both conditions were then classified according to their self-reported occupation, with those working full time, students, and those occupying more than one role (e.g., housewife *and* student) in one category, and housewives, retirees, and disabled persons in the other group. This variable crudely measures the likelihood that the respondents would be spending time at home rather than other places. Approximately 7 percent more of the respondents in the control condition were classified in the "outside occupation" category. It could be, then, that the difference between the control and experimental groups in TV time reported was due to the fact that more respondents in the control group were absent from the home and could not have viewed television. (Remember that the experimental group showed more time spent with TV.)

To assess this possibility, we performed a multiple classification analysis (MCA) (Andrews, Morgan, and Sonquist, 1969) in which the TV time estimate for the day preceding the interview was predicted by the type of interview procedure, occupation as coded above, and years of education (dichotomized at 12 years and below and 13 years and above). The results of this analysis are presented in Table 2. The adjusted deviations and betas in the second column indicate that the impact of the interview technique on TV time reported does not diminish when controlling for the occupation and education of the respondents. When the experience of the respondents is even further standardized in analytic controls, the effect of the improved interviewing techniques remains strong, giving us added confidence that the interviewing procedures produced the reporting differences between the two groups.

But what is the nature of the differences? Is it possible that our assumptions about reporting bias are faulty and that the experimental procedures are really producing "pseudo-data" resulting from the respondents' desire to please the interviewer? Let's focus on the TV time example again. Respondents could be making up answers to this question in response to the urging in the experimental condition that *all* time in contact with the medium be reported. To test this possibility, another multiple classification analysis was performed on time spent watching TV with the type of interview and occupation as independent variables and the number of TV programs reported watched as the covariate. The logic of this test is that the time estimate difference between the groups should depend on recall and that controlling for the number of programs reported should "wash out" the difference between the conditions if it is due to this mechanism. The experimental techniques, in other words, should lead to better

Table 2
The effect of interviewing techniques, occupation, and education on report of time spent watching television[a] (Multiple classification analysis)

Item	*N*	*Deviation*	*Eta*	*Deviation*	*Beta*
Interview type:					
Control condition	105	−31.56		−30.96	
Experimental	100	33.14	.23	32.51	.23
Education:					
0–12 years	129	8.53		11.52	
13+ years	76	−14.47	.08	−19.52	.11
Occupation:					
Working, student, other	117	−24.63		−22.79	
Housewife, retired, disabled	88	32.8	.20	30.3	.19
Grand mean = 140.66 minutes					

Multiple R^2 = .10

Multiple R = .32

[a] N includes both those who reported no TV watching and those who reported some viewing.

Table 3
The effect of interviewing techniques and occupation on report of time spent watching television, controlling for number of TV programs reported (covariate)[a] (Multiple classification analysis)

Item	N	Unadjusted deviation	Eta	Adjusted for other independents and covariates	Beta
Interview type:					
Control condition	105	−31.56		−10.31	
Experimental	100	33.14	.23	10.82	.08
Occupation:					
Working, student, other	117	−24.63		−15.74	
Housewife, retired, disabled	88	32.8	.20	20.93	.13
Grand mean = 140.66 minutes					

Multiple R^2 = .52

Multiple R = .72

[a]N includes both those who reported no TV watching and those who reported some viewing.

recall of the content viewed and hence to higher reported time spent. If this is not the case, we would have more reason to suspect that the difference in time estimates between the two conditions is the result of a bias produced by the experimental techniques.

The results of the MCA, in Table 3, support the notion that the difference in TV time reported between the interview conditions is the result of better recall in the experimental group. The effect of the experimental manipulation is markedly reduced by the control for number of programs recalled, as one can see by looking at the adjusted deviation and beta for interview type. Meanwhile, the impact of occupation on TV time reported, which should not be so dependent on the number of programs recalled, is only slightly reduced when the covariate is taken into account. It appears, in sum, that people exposed to the experimental interviewing procedures are not simply making up information to please the interviewer; instead, the TV time differences seem to reflect more effort on the part of respondents in the experimental condition.

Taking a look at a reporting task that presents a social desirability problem adds further support for the contention that the experimental procedures set up the conditions for more valid reporting. Consider the X-rated movie attendance item. As was seen in Table 1, this behavior—which seems likely to be underreported—is reported more often by those in the experimental condition than by the control group respondents. To further elucidate this finding, we have partitioned the sample into two groups: those with 12 years of education or less and those with 13+ years. The analysis, presented in Table 4, focuses on the people for whom reporting X-rated movie attendance is likely to be

Table 4
The effect of interviewing techniques on report of X-rated movie-going, by education

Lo-Ed (0–12 years) Interview type

Ever attended X-rated movie?	Control	Experimental	
Yes	36 (50%)	40 (68%)	76
No	36 (50%)	19 (32%)	55
	72	59	131

Phi=.18
χ^2=3.5

Hi-Ed (13+ years) Interview type

Ever attended X-rated movie?	Control	Experimental	
Yes	18 (53%)	21 (51%)	39
No	16 (47%)	20 (49%)	36
	34	41	75

Phi=.02
χ^2=.01

more threatening or embarrassing. The hypothesis is that those with high education are less likely than those with low education to be embarrassed reporting "pornographic" movie attendance. The "Hi-Ed" group is likely to be more cosmopolitan and less self-conscious about sexual taboos. The "Lo-Ed" people, by contrast, are more likely to find this type of event embarrassing to admit—particularly to a younger, college student interviewer.

The implication of this hypothesis is that the experimental interviewing techniques should have more impact in the "Lo-Ed" group than in the other; we should find greater differences in reporting between the two experimental conditions for the "Lo-Ed" than for the "Hi-Ed" people, since the latter are likely to report this behavior anyway. The cross-tabulations in Table 4 bear us out on this point, since the relationship between reporting and interview type is much stronger in the "Lo-Ed" group than in the other. Specifying the general finding in this way, we have more confidence in the inference that the experimental procedures reduce social desirability bias.

A similar analysis presented in Table 5 focuses on a social desirability problem that may be more acute for "Hi-Ed" than for "Lo-Ed" respondents. Considering a report of the number of books read in the last few months, the assumed tendency is for the behavior to be overstated, paticularly by the "Hi-Ed" group, who may place more value on appearing to be well read. If this is the case, and the experimental treatment is working to reduce this tendency, we should find less difference in reporting of the number of books between the education groups in the experimental than in the control condition. As can be seen in Table 5, in the control condition there is a marked disparity between the education groups in the number of books reported, while the experimental group shows only a moderate difference. When we consider the fact that many respondents in the control group were unable to support their book-reading estimate with a report of the titles of the books, the difference between "Hi-Ed" respondents in the different experimental treatments in this analysis indicates that the new interviewing procedures are reducing the pronounced tendency for those with higher education levels to overreport their book-reading behavior.

In summary, it appears from our look at the data in Tables 2–5 that the experimental interviewing procedures do produce differences in reporting of media use behaviors and that these differences are consistent with our best assumptions about the direction of error in various measures. An interesting problem that we have not yet considered, however, is what implications the better measurement of individual variables has for estimates of the relationships between measures. We take up this question in the next part of the paper.

Table 5
Relationship between education and number of books reported read, by interviewing condition

Education	*Mean number of books reported*	
	Experimental condition	
0–12 years	2.40	(s=5.2)
13+ years	3.66	(s=5.3)
Grand mean	2.9	
	*Control condition**	
0–12 years	3.85	(s=8.5)
13+ years	8.49	(s=13.3)
Grand mean	5.3	

* "Difference between "HI-Ed" and "Lo-Ed" means significant, $p < .05$.

The price of error in relationships. It is readily understood that random measurement error attenuates our estimates of relationships between variables, and there are standard correction procedures for dealing with this problem, although some observers have suggested that the procedures are underutilized (viz., Bohrnstedt and Carter, 1971). The impact of nonrandom error—the sort that we have considered in this paper—on relationships seems to be less predictable, although such writers as Costner (1969), Alwin (1977), O'Muircheartaigh (1977), and Schuman and Presser (1977) have discussed various facets of the issue. It is clear at the outset that the implications of nonrandom response errors for relationships are different from the implications for mean scores, or differences. It is not necessarily the case, then, that correcting a reporting bias in a given variable will alter its relationships with other variables. But, as Schuman has demonstrated in his experiments with question wording, nonrandom error *can* have pronounced effects on relationships, contrary to the heuristic rule commonly adopted in public opinion research that wording differences produce different univariate distributions but do not affect relationships. The problem can be seen to arise in those instances when the bias in one measure is itself correlated with the score on another measure, or when the biases in two measures are correlated with each other. Costner (1969) has termed this problem "differential bias." Manifestations of differential bias are cases in which a relationship between two

variables is either inflated or reduced by the correlation involving the error terms. In such instances, the experimental interviewing procedures can have a salutory effect, in theory, to the extent that they reduce the biases distorting the correlations.

One example is the relationship between occupation and time spent watching television. We noted earlier that our measure of occupation is related to estimates of the time spent with TV "yesterday," but what does the relationship look like when we examine it *within* each experimental condition? The assumption guiding this analysis is that the underreporting bias involved in reports of "secondary viewing" of television is correlated with the amount of time that the respondent has available to watch (measured by the respondent's occupation). Those most likely to spend a great deal of time watching TV are also most likely to underreport the experience, according to this view. The result would be that the difference in TV time estimates between those for people who have lots of opportunity to watch and those for people who are often absent from home would be attenuated. If the underreporting bias were reduced, one would find a greater difference in TV time reported for people with different occupations.

A comparison of the relationship between occupation and TV time reported "yesterday" for the control and experimental groups is presented in Table 6. The relationship does not reach the level of statistical significance in the control group, and the eta is about half the magnitude of that observed for the experimental group. Assuming that our surmises about reporting bias are correct, the results of this analysis indicate that in certain cases the experimental interviewing procedures, by reducing nonrandom error, help to elucidate the nature of relationships between variables.

Of course, this is only one example, and we need to do much more of this sort of work for at least two reasons. First, although the experimental interviewing procedures have a demonstrated impact on reporting biases of various sorts, many investigators who are primarily interested in relationships among variables need to know if the extra effort involved in using the new techniques will pay off in improved estimates of relationships. As mentioned earlier, reduction in response bias in a variable does not necessarily have implications for correlations between that measure and others.

Secondly, there are many cases in which assessing the efficacy of the experimental techniques in reducing response bias for individual variables involves examining their relationships with other measures. We saw some of these instances earlier when considering the TV time, X-rated movie, and book-reading estimates. In each case, we supported the inference of more response validity in the experimental group by checking our expectations about how the variables should relate to other measures. Although this was helpful in these cases, it is essential when we consider attitude or "quality of life" measures, for which it is not often possible to make a reasonable assumption about reporting bias. Assessing measurement validity in these cases is strictly a matter of seeing whether the items correlate with measures with which they should correlate and do not relate to theoretically unrelated variables. To examine through a technique such as confirmatory factor analysis whether several measures are tapping the same underlying construct, or whether they share "method variance" instead, involves dissecting the pattern of correlations among the items. To the extent that the experimental interviewing procedures help to reduce reporting bias that affects the relationships, we would obtain a different perspective on the validity of such measures. A considerable but very important task for future research is to assess whether the experimental interviewing techniques alter reporting of attitudinal variables in a way that provides this perspective.

Table 6
Effect of interview type on relationship between occupation and time spent watching television (Multiple classification analysis)

Occupation	*N*	*Deviation*
	*Experimental condition**	
Working, students, other	53	−35.59
Housewife, retired, disabled	47	40.14
Eta	.24	
Grand mean = 173.80 minutes		
R^2 = .06		
R = .24		
	Control condition	
Working, student, other	64	−11.44
Housewife, retired, disabled	41	17.86
Eta	.13	
Grand mean = 109.1 minutes		
R^2 = .02		
R = .13		

* Analysis of variance significant for this group, $F = 6.09$, $p < .02$.

The process of interviewing effects. To gain supplementary evidence on how the experimental interviewing methods work, we asked respondents and interviewers to complete an

evaluation form after the interview, in which they rated their own performance and perceptions of the interview. Included in the list of ratings were questions on the effort expended by the respondents, their accuracy in answering, their honesty, and the degree of favorability toward the interview. Also included were perceptions of how accurate the respondents thought they had to be when answering and how difficult the questions were. The ratings were used in a discriminant function analysis to identify the ones that best distinguished between the interviewing conditions. Similar to stepwise multiple regression, this analysis seeks a linear combination of variables that best predict membership in a given category—in this case, the experimental or the control group. "Discriminant function coefficients," analogous to regression coefficients, are computed for each rating, and the scores on the ratings are used to classify respondents into groups. The predicted classifications can then be checked against the actual group membership (experimental or control, in this case) to judge the efficacy of the analysis prediction.

The ratings that best distinguished between the control and the experimental group were the respondents' perceptions of how accurate they thought they were supposed to be when answering and how difficult they thought the questions were. That is, respondents in the experimental condition said they thought that more accuracy was required in responding than did the control group respondents, and those exposed to the experimental procedures rated the questions as more difficult than did those in the control condition—even though they were the same questions. This "manipulation check" indicates that some primary messages of the experimental procedures were communicated during the course of the interview. The conclusion is buttressed by the fact that the interviewers rated respondents in the experimental group as expending more effort than those in the control condition.

Some suggestions for further research

Considerable work remains to be done in applying the experimental interviewing techniques to new substantive areas and new measurement problems. At Michigan's Survey Research Center some research is currently under way that utilizes the interviewing procedures on the telephone and extends the work presented here on media use. Certainly, it is necessary for replication studies to be done to support or reject the conclusions stated here.

Besides replicating the findings on reduction of bias, we need to expend much more energy assessing the impact of the techniques on relationships. This requires large N within-condition analysis, to see if relationships among variables show different patterns for those respondents exposed to the experimental treatments. One particularly interesting possibility is to employ a variant of the Campbell and Fiske (1959) multitrait-multimethod matrix in designing the comparison of relationships between the control and experimental conditions. Different question types could be employed to assess several different "traits" within each condition, and the assessment of the efficacy of the techniques would involve seeing whether the impact of "method variance" on relationships is reduced in the experimental group. As stated earlier, this sort of procedure is the only way to deal with questions of validity for many of the measures that we commonly use in public opinion research. Interviewing experiments must be designed to accommodate this kind of test.

Another area of interest is the possible impact of the techniques on random measurement error, or simple reliability of measures. In published studies to date, repeated measures of the same variables have not been part of the design considerations, since the effort has focused on reporting bias. It seems encumbent on researchers utilizing the experimental methods to explore their effects on reliability as well. Of course, the ultimate goal of all such work is to produce a set of data collection procedures that may be used widely in survey research and that consistently show valid reporting across a wide range of response problems. Therefore, research should also focus on the potential difficulties of applying the techniques in new areas and on the costs of doing so. For example, we have seen in this research that it is often difficult to make reasoned guesses about the extent and direction of error in measures, but this exercise is crucial to applying the experimental techniques. Also, the procedures may demand more time in interviews (the experimental interviews in this study took about ten minutes longer, on the average, than did the control interviews) and may reduce the number of questions included in any questionnaire so that those included can be measured well. Investigators who might employ the procedures must be convinced that the quality of the information obtained compensates for any costs in money and time that the techniques involve. Further research could focus on cost/benefit ratios so that survey practitioners can make a reasoned decision.

Finally, it seems that we should be investigating more closely how the experimental procedures work their effects. Careful monitoring of the respondent role-learning procedure is needed to discover how instructions and feedback effects generalize from one question to another, and for which types of questions these techniques are most effective. The work of psychologists such as Einhorn (forthcoming) on learning of heuristics provides some seminal perspectives on how to investigate the cognitive effects of the techniques.

Based on past experience, it seems that the experimental interviewing work can have some real benefits. It remains to us to delimit these advantages more carefully and to explore the ramifications of applying the new techniques in survey research generally. Eventually, one would like to use data collected via this approach as benchmarks against which other survey methods could be judged.

Discussion: Applying health interview techniques to mass media research

Lu Ann Aday, Center for Health Administration Studies, University of Chicago

This research appears to be a useful effort to extend some of the approaches for evaluating interview technique biases, developed in connection with health care studies, to another substantive interest area—mass media research. The analysis documents the applicability of social survey methodology advances in the health care field—such as those being reported and subjected to critical review at this conference—to many other substantive fields of interest. The questions that occurred to me as I reviewed Miller's paper, however, primarily concerned the implications that the experiment and the resulting conclusions have for the design and conduct of health-related surveys. It is this focus that will tend to guide my comments.

The magnitude and determinants of over- and underreporting biases are of continuing concern to health survey researchers and methodologists. A variety of hypotheses are used to account for these outcomes. The reinforcement strategies that are employed in the interview situation represent one set of hypothesized determinants of such biases. It is this hypothesis with which Miller's experiment is primarily concerned. His data tend to confirm that modes of interviewer reinforcement may well impact on the number and kind of events that respondents report. A model of Total Survey Error (TSE), however, points to a number of factors that may give rise to positive and/or negative biases in the data—some of which may relate to the behavior of the interviewer; others to the salience or threatening nature of the events to be reported, the recall period about which information is to be provided, the availability of records or other memory aids, nonfield-related errors from estimating missing data, etc. Work by Ronald Andersen et al. (1979) on the sources of bias in health survey estimates suggests that the direction and magnitude of the bias may differ considerably for different estimates and for different population subgroups for which they are reported. Although interviewer reinforcement strategies are undoubtedly important contributors to some of these estimation errors, they may represent only one small part of the TSE problem.

Marquis (1978a), in a reassessment of bias in reports of hospitalizations, argues that the particular design strategy for evaluating biases may, in fact, itself determine the direction of the bias. Retrospective designs that check to see if information from records is captured in social surveys tend to yield underreports of expected events, whereas prospective designs that check social survey responses against records tend to show overreporting of health care encounters. Comprehensive design strategies incorporating both prospective and retrospective approaches tend to show that the under- and overreporting observed when applying the respective approaches may, in fact, cancel one another out. Miller's work does not provide an "objective" nonsurvey criterion against which survey responses can be evaluated. Comparisons are made between the experimental and control groups only. The direction of the shifts of responses in the respective groups intuitively suggests that the interviewer reinforcement strategies contribute to reducing biases. We cannot know with certainty that the estimates obtained in the experimental situation were closer to "true" population parameters. We need, therefore, to be conscientious critics of the paradigms that we use to detect biases and of how the methods that we employ may enable us to see one facet of a problem but not necessarily the whole.

A central and important question that Miller raises is the impact that interviewer reinforcement (or any biases, for that matter) in social survey estimates may have on the relationship of these estimates to other variables. Correlation and prediction are, it is probably fair to say, higher-order research aims than simple estima-

tion or description of observed events. To the extent that we understand the association between variables of interest, there is perhaps a greater possibility for altering certain factors to give rise to agreed-on outcomes. These kinds of considerations are of particular interest in policy-oriented health services research that is concerned with those mutable factors in the health care system that can be altered to bring about changes in that system, such as improved access, lower costs, or higher-quality medical care. Miller points out that his current analyses do not probe deeply into these issues, but it seems a significant next step to examine how systematic errors in estimating outcomes of interest may affect what one concludes about the correlates or causes of these outcomes.

Other observations on Miller's paper concern the difficulties that he points out of operationalizing a true experiment to test the hypotheses of interest. One aspect of the design—the assignment of the same interviewers to both the experimental and control situations—caused me to wonder if "maturation" on the part of the interviewers might tend to mitigate the experimental and control group differences. Thus, might the interviewers over time begin (perhaps unconsciously) to apply some of the reinforcement strategies from the experimental protocol to their "regular" (nonexperimental) interviews? There is no evidence of this effect, however, in the data presented. I would also be interested in more detail describing how the experimental and control group interview protocols differed—particularly on the variables reported in the analyses—and how a comparable methodology might be designed for application in a telephone interview setting.

Response styles in telephone and household interviewing: A field experiment from the Los Angeles Health Survey*

Lawrence A. Jordan, System Development Corporation

Alfred C. Marcus, School of Public Health, University of California at Los Angeles

Leo G. Reeder, School of Public Health, University of California at Los Angeles

In the past few years considerable interest has developed in using telephone surveys to assess public opinion. As Klecka and Tuchfarber (1978:105) observed, "a minor revolution has taken place in survey research. Telephone surveys based on some form of random digit dialing (RDD) have gained general acceptance among much of the survey community as a legitimate method for sampling public opinion." Although the telephone may have "come of age," as Dillman and Frey (1974) suggest, more research is needed to determine the comparability of telephone interviewing with conventional face-to-face interviewing techniques. For example, studies have examined the advantages of using the telephone for making appointments, locating the hard to reach, and screening for proper respondents (Sudman, 1966); for brief reinterviews on health matters (Kegeles, Fink, and Kirscht, 1969; U.S. Public Health Service, 1962; Walden, 1975); for gathering data on fertility (Coombs and Freedman, 1964); and for sampling (Cooper, 1964; Hauck and Cox, 1974). Although few of these studies have actually compared telephone and household interviewing techniques using a controlled experimental design, several recent experimental comparisons of household and telephone surveys have been undertaken, including the Los Angeles Health Survey (LAHS) experiment, the 1976 University of Michigan study (Groves, 1977), the University of Cincinnati study (Klecka and Tuchfarber, 1978), the Medical Economics Survey-Methods Study (Yaffe et al., 1978), and the National Crime Survey and Current Medicare Survey experiments of the Census Bureau (Bushery, Cowan, and Murphy, 1978).

After the infamous 1936 *Literary Digest* debacle, telephone surveys were generally distrusted. The prediction of an Alf Landon landslide was a grave embarrassment for survey research, directly attributable to the socioeconomic bias in the *Literary Digest*'s sample of telephone owners. Many more households in the United States now have telephones, however, so that socioeconomic biases are much less likely in telephone surveys (Dillman, 1978). Since 1963, the National Center for Health Statistics has asked about telephone availability in the national Health Interview Survey. These data show that telephone availability has increased steadily from 80 percent of households in 1963 to over 90 percent in 1977 (Thornberry and Massey, 1978). However, these data also reveal variations in access to telephones among certain population subgroups. Groups with lower rates of telephone access include people living in the South or in rural areas, the unemployed, those with lower education and income, and those who are separated or divorced. Consequently, while social class and life-style differences in telephone availability have diminished greatly in the past decade, subgroup variations still exist and should be considered when deciding to conduct telephone surveys.

There are many good reasons for wanting to use telephone surveys:

1. They are usually less expensive than household surveys, depending to some extent on the kind of information and sample required. Costs of telephone surveys have ranged from 20 to 66 percent of the costs for comparable household surveys—45–66 percent (Groves, 1977); 50–66 percent (Hochstim, 1967); 20–25 percent (Klecka and Tuchfarber, 1978).
2. It is now feasible to conduct nationwide surveys by long-distance telephone, using a cen-

* This paper is an expanded version of one presented at the annual meeting of the American Statistical Association in San Diego, August 1978. The research was supported by Grant Number 5-R18-CA-18451, "Processes in Health Behavior and Cancer Control," awarded by the National Cancer Institute, DHEW, to Leo G. Reeder, Principal Investigator. Dr. Reeder died in a plane crash on September 25, 1978.

trally located and supervised staff of interviewers. A central location leads to more control over the interviewing process, with nearly immediate detection and correction of interviewer or respondent communication problems; it also allows computer-assisted telephone interviewing (Shure and Meeker, 1978).

3. Time in the field can be greatly reduced with telephone methods. The NBC–Associated Press and CBS–*New York Times* polling organizations can deliver the marginal results from short questionnaires almost overnight.

For all these reasons, telephone surveys are an attractive alternative to household interviews. Consequently, it is important to determine whether the data collected by telephone methods are comparable to those collected by household interviews. Initial results have been very promising, with recent comparisons of the two methods revealing no large systematic differences (Bushery et al., 1978; Colombotos, 1969; Groves, 1977; Hochstim, 1967; Klecka and Tuchfarber, 1978; Reeder, 1976, 1977; Rogers, 1976; Schmiedeskamp, 1962; Wiseman, 1972). We turn now to a comparison of the two methods in our data.

Procedures

The Los Angeles Health Survey. The LAHS is an ongoing, annual or biennial survey of health behavior and health attitudes in the Los Angeles metropolitan area. In 1976, the LAHS research team conducted household (HH: N = 1,210) and telephone (TC: N = 303) surveys, at the same time and with essentially the same instrument. In 1977, further independent household (N = 931) and telephone (N = 381) surveys were conducted.

We attempted to keep the HH and TC interviewing arrangements as nearly alike as possible, especially by using the same interviewers. Although the interviewers had many months and often years of experience at UCLA conducting household surveys, several of them did not adapt well to telephone interviewing. For the 1977 surveys, several improvements in the telephone inteviewing arrangements were introduced, the major one being that the interviews were made from a central location with on-line monitoring instead of made from the interviewers' homes.

Sampling. For the HH samples, the LAHS used the sampling frame developed by the UCLA Survey Research Center for its Los Angeles Metropolitan Area Sample (LAMAS). This frame contains approximately 20,000 computer-readable addresses sampled on an area probability basis. Samples drawn from the frame may be characterized as three-stage cluster samples, with probabilities proportional to size. Each housing unit (HU) in Los Angeles County has an equal chance of being selected (Sumner, 1978).

For the TC sample, there was an attempt at matching the cluster design of the LAMAS frame. Using reverse telephone directories, telephone exchanges were sampled from the addresses in the block groups representing the secondary sampling units from the frame; then calls were made to randomly selected telephone numbers from the obtained exchanges. The exchanges cover much larger areas than the original census tracts, however, and only 20.8 percent of the TC respondents lived in the targeted census tracts.

Table 1 lists design effects for standard errors of means, which indicate the extent of clustering in the samples. The design effects were computed as

$$DE = \text{S.E. (BRR)} / \text{S.E. (SRS)},$$

where S.E. (BRR) is the standard error of the mean obtained using balanced repeated replicates, and S.E. (SRS) is the ordinary standard error obtained with simple random sample assumptions (Kish and Frankel, 1970). Values for DE that exceed 1 indicate that the respondents within PSUs are more homogeneous than expected for a simple random sample.

For the HH sample, design effects ranged from 1.09 to 2.07, with design effects of about 2.0 for the demographic variables Percent Anglo, Education, Income, and SEI (socioeconomic index). For the TC sample, however, the design effects ranged from .89 to 1.20, indicating that there was very little clustering for the TC sample. Software for statistics more complicated than standard errors of means is not widely available, and there is no direct relationship between DEs for standard errors of means and DEs for standard errors of mean differences (Frankel, 1971). However, using the approximate relation

$$\text{S.E.}\,(\bar{X}_1 - \bar{X}_2) = \sqrt{\text{S.E.}^2\,(\bar{X}_1) + \text{S.E.}^2\,(\bar{X}_2)},$$

we can estimate the standard error of a difference when cluster effects are present as

$$\text{S.E.}\,(\bar{X}_1 - \bar{X}_2)$$

$$= \sqrt{\text{S.E.}^2\,(\bar{X}_1)\,DE_1^2 + \text{S.E.}^2\,(\bar{X}_2)\,DE_2^2}$$

$$= \text{S.E.}\,(\bar{X})\sqrt{DE_1^2 + DE_2^2}, \quad \text{where}$$

both standard errors of means are equal. The

usual estimate for S.E. $(\bar{X}_1 - \bar{X}_2)$ is S.E. $(\bar{X})\ \sqrt{2}$ where the standard errors of means are equal. Thus, a very conservative way of handling the design effects would be to estimate DE_1 and DE_2 by the most extreme values in Table 1 (for Percent Anglo) and treat the obtained t values for mean differences between samples as if they were inflated by a factor of $\sqrt{(2.07^2 + 1.20^2)/2} = 1.69$. By this rule, one would need a t of 3.32 instead of 1.96 for significance at the .05 level. Not all variables have DEs as large as Percent Anglo, however, so t-test factors have been computed for each variable in Table 1.

Table 1
Design effects for selected variables: 1976 household (HH) and telephone comparison (TC) samples

Variable	Design effects[a]		t-test factors[b]
	HH	TC	
Respondent age	1.27	.89	1.10
Percent male	1.04	.99	1.02
Number of adults	1.21	.96	1.09
Number of children	1.61	1.08	1.37
Percent married	1.50	1.17	1.35
Percent Anglo	2.07	1.20	1.69
Mid-education	2.04	1.18	1.67
Total income (missing=mdn)	1.90	1.07	1.54
Total income (regression est.)	1.96	1.03	1.57
SEI	1.99	1.04	1.59
Health status (self-report)	1.19	1.13	1.16
Health status (composite)	1.19	1.03	1.11
Agreeing	1.30	1.01	1.16
Extremeness	1.46	1.04	1.27
Evasiveness	1.40	1.00	1.22
Acceptability	1.12	1.06	1.09
Accessibility	1.40	1.19	1.30
Susceptibility	1.18	1.13	1.16
Motivation	1.09	1.18	1.14
Efficacy of care	1.15	.97	1.06
Seriousness	1.14	1.06	1.10
Cost concern	1.36	1.00	1.19

[a] The design effect is a ratio of the balanced-repeated-replicates estimate of standard error to the simple random estimate.

[b] The t-test factors were computed as $\sqrt{(DE_1^2 + DE_2^2)/2}$.

Weighting. For both the HH and TC samples, individual adult respondents within HUs were selected using the Kish (1965:400) procedure. This raises another methodological issue, which is that while HUs have equal probabilities of selection, individuals do not (persons in two-adult HUs have only half as much chance of being in the sample as persons in one-adult HUs, etc.). The unweighted sample statistics are estimates for the population of HUs in Los Angeles. To make inferences about the population of adult persons, it is necessary to weight the sample observations inversely by the number of adults in each HU. Now, despite the clear technical differences between weighted and unweighted analyses of the data, we would seldom want to make a strong inferential distinction between them. That is, for most variables of interest, we would expect an analysis of data for random respondents from randomly selected HUs to yield essentially the same results as an analysis of data for a simple random sample of respondents. Accordingly, we adopted the conservative practice of interpreting only the results that were *robust under weighting*, or statistically significant in both the weighted and unweighted analyses. In the present study, we found that the weighted and unweighted differences were almost invariably in the same direction. However, there were a few borderline results that were significant in one analysis but not in the other. We will call attention to such results in later sections of this paper. Since the sample sizes in this study are rather large, results that are barely significant at the .05 level are likely to have little practical value, and little will be lost by ignoring results that are not robust using multiple criteria.

Response rates. A final methodological topic is that of response rates. The initial listings consisted of 2,020 addresses for the HH sample and 1,114 telephone numbers for the TC sample. For the HH sample, 137 units (6.8 percent) were considered ineligible (vacancy, no dwelling at address, language barrier); for the TC sample, 401 units (36.0 percent) were considered ineligible (discontinued, business number, not in service, language barrier). The only common category of ineligible units was a small category for language barriers (English or Spanish not spoken). Business addresses were eliminated from the HH sample during field listing, except for a few addresses listed in error or with changed occupancy since the field listing. Business numbers, on the other hand, were the second largest category of ineligible TC units. Discontinued numbers represented the largest single category of ineligible TC units and had no clear counterpart in the HH sample. For the HH sample, 1,210 of 1,883 (64.3 percent) of the eligible units provided completed interviews; for the TC sample, 303 of 613 (49.4 percent)

eligible units provided completed interviews. Thus, with response rate defined as the ratio of completions to eligible units, the response rates were lower for the telephone interviews than for the household interviews. There is, however, no evidence that these different response rates contributed to demographic or other biases in the two samples, as shown below.

Results

Sociodemographic variables. First, we will look at the sociodemographic variables, as shown in Table 2. The most striking difference between the HH and TC samples is the amount of missing data for income. Respondents were asked for total family income on a 15-point scale. For the HH sample, hand cards were used, but for the TC sample, the interviewer had to read the appropriate response categories. The income question is one of the most sensitive sociodemographic questions, and some respondents either refuse or claim not to know the answer. We found that 12 percent of the HH respondents and 21 percent, or nearly twice as many, of the TC respondents did not answer.

In the University of Michigan study, Groves (1977) also reported more income data missing in the telephone survey. He found, however, that the missing data rate declined for subsequent telephone surveys, which he attributed to improved interviewing techniques.

Our own experience has been similar. The missing-data rate dropped from 21 percent in 1976 to 15 percent in 1977 for the TC sample, which is still slightly higher than the 12 percent rate obtained with the two household samples. It remains to be seen whether, as more experience is gained with telephone methods, the missing-data rate can be reduced to the 12 percent obtained on face-to-face interviews. We have received more "don't know" than "refused" responses for income, and we assume that many of these "don't know" responses are indirect refusals from reluctant respondents (Robins, 1963). A rather large proportion of the missing income data for the 1976 telephone survey was in the "don't know" category (14.5 percent), suggesting that respondents were being evasive rather than refusing the information outright.

In view of the large amounts of missing data for income, the obtained mean difference on income is difficult to interpret. The groups are not significantly different on other indicators of socioeconomic status, such as education and Duncan's socioeconomic index (SEI). Therefore, in the absence of confirming evidence of a socioeconomic difference between the samples, we conclude that the obtained income difference is probably spurious or, if real, is rather small in size.

Another finding in Table 2 is that the TC sample has a significantly larger proportion of Anglos. Since this comparison was not signifi-

Table 2
Selected demographic variables: means for household (HH) and telephone comparison (TC) samples, unweighted, 1976

Demographic variable	HH sample (N = 1,210)		TC sample (N = 303)		t for mean difference
	$\bar{X}$	s	$\bar{X}$	s	
Age	42.8	17.4	44.01	18.1	−1.16
Percent male	0.43	0.50	0.37	0.48	1.188
Percent married	0.55	0.50	0.50	0.50	1.56
Number of adults	1.84	0.75	1.82	0.78	0.41
Number of children	0.88	1.37	0.76	1.23	1.39
Income	8.21	3.82	7.61	3.48	2.49*[b]
Income—percent missing	0.12	0.32	0.21	0.41	−4.58***
Mid-education	12.46	3.57	12.25	3.36	0.93
SEI	48.80	22.50	47.83	22.74	0.67
Percent Anglo	0.67	0.47	0.73	0.44	2.01*[a,b]
Percent Hispanic	0.06	0.24	0.08	0.26	−1.27

[a] Significance criterion was not met for weighted analysis.

[b] Significance criterion was not met for conservative analysis allowing for cluster effects, using t-test factors in Table 1.

Significance levels for unweighted analysis:

* $p < .05$.

*** $p < .001$.

cant in a weighted analysis of the same data, we recommend caution in accepting the finding. Note also that if we adopt a conservative critical value of 3.32 for t values, as suggested earlier, neither the Income mean nor Percent Anglo difference is significant.

Our final observation about Table 2 is that there are very few demographic differences between the HH and TC samples, which confirms the findings of Groves (1977) and Klecka and Tuchfarber (1978). With over 1,500 respondents, we have adequate power to detect quite small differences between the samples; and our results suggest that, apart from the greater amount of income data missing in the TC sample, the two samples are essentially the same on all sociodemographic variables. This lack of demographic differences is a fortunate result, since telephone interviewing will probably soon be the dominant method for obtaining many kinds of survey data.

Attitude variables. The main focus of our paper is on the HH and TC differences obtained for the attitude variables in Table 3. The first seven variables in the table are 2-to-6 item content scales for measuring aspects of respondents' health beliefs (Berkanovic, Marcus, and Jordan, 1978; Kirscht, Becker, and Eveland, 1976) and are composed of 4-point Likert items (strongly agree, agree, disagree, strongly disagree). There were 32 items in all, administered at three points in the questionnaire and then scored to form the seven attitude scales. As in the case of the income question, administration of the attitude items was supplemented by a hand card for the HH sample, with the four response categories displayed for the respondent's use. The instruction read:

> I'm going to read some statements related to health and illness. I'd like you to tell me the extent to which you agree or disagree with each statement. People have many different opinions about health—so there are no right or wrong answers. We are simply interested in *your* opinions. (HAND CARD). Please tell me if you *strongly agree, agree, disagree,* or *strongly disagree* with the following statements.

For the TC sample, the instructions were identical, but since hand cards could not be used, the interviewer instruction read:

> Repeat response categories as you read items.

An attempt was made to balance the scales for keying in the agree-disagree direction. For example, the Cost Concern scale contains the following pair of items:

(A) When I think about going to the doctor I'm concerned about how much it will cost.

Table 3
Selected attitude variables: means and standard deviations for household (HH) and telephone comparison (TC) samples, unweighted, 1976

Attitude variable	Items in scale: Agree keyed	Items in scale: Disagree keyed	HH sample (N = 1,210): $\bar{x}$	HH sample (N = 1,210): s	TC sample (N = 303): $\bar{x}$	TC sample (N = 303): s	t for mean difference	χ^2 for standard deviation difference
Susceptibility	3	3	−1.79	2.34	−2.08	2.74	1.86	12.82***
Acceptability	3	3	0.92	2.76	0.64	3.58	1.48	36.14***
Accessibility	3	3	1.97	2.34	2.31	2.84	−2.16*[a]	19.56***
Cost concern	1	1	−0.45	1.42	−0.28	1.67	−1.79	13.56***
Seriousness	2	3	2.72	1.91	2.61	2.07	0.88	3.23
Efficacy of care	1	2	3.39	1.46	3.47	1.74	−0.82	15.95***
Motivation	3	1	−4.13	1.94	−3.85	2.02	−2.23*[a]	0.80
Agreement bias	—	—	12.53	3.93	13.88	4.26	−5.25***	3.25
Evasiveness (ln)	—	—	0.48	0.67	0.61	0.66	−3.03**	0.11
Extremeness (ln)	—	—	1.91	1.07	2.51	0.90	−8.99***	13.58***

[a] Significance criterion was not met for weighted analysis.

Significance levels for unweighted analysis:

* $p < .05$.

** $p < .01$.

*** $p < .001$.

(B) I don't worry much about the cost when I think about going to the doctor.

Persons concerned about the cost of medical care will tend to agree with the first statement and disagree with the second, so that the two items can be considered balanced for agreement. Balancing is an important precaution if the items are affected by "acquiescence" or agreement response bias (Jordan, 1977). As shown in Table 3, the first four scales (from Susceptibility through Cost Concern) are balanced; the next two (Seriousness and Efficacy of Care) are slightly unbalanced, with one more disagree than agree item; and the seventh scale (Motivation) is unbalanced, with three agree-keyed items and only one disagree-keyed item.

The last three variables in Table 3 are response bias measures, which were also obtained from the health belief items. The agreeing bias variable was obtained by scoring all 32 items in the "agree" direction regardless of content keying. The evasiveness variable was computed as a count of "don't know" and "no answer" responses. The extremeness variable was computed as the difference between the number of extreme or "strong" responses and the number of middle-category responses. In raw score form, the evasiveness and extremeness variables were markedly skewed, so the variables were normalized by adding a constant and taking logarithms.

The main message in Table 3 is that the TC sample is higher on all three response-style measures. The TC sample has a greater tendency to agree, to omit responses, and to use more extreme categories. The extremeness effect is the strongest of the three response-style effects and is significant at well beyond the .001 level, with a mean difference on the order of one-half of a standard deviation. By contrast, there are only minor mean differences for the seven content measures. There are slight mean differences between the HH and TC samples for the Accessibility and Motivation scales, but they are fairly small and are not robust under weighting. In addition, the TC sample is more variable on all seven content measures, significantly so for five of the seven. The larger variances are attributable to the extremeness response bias, since the extreme categories for the item responses contribute more than the middle categories to variances for the content scales.

Open-ended questions and checklist. Another important issue in assessing the comparability of the two modes of interviewing concerns the quantity of information obtained for each method. Telephone interaction is typically faster paced than face-to-face interaction and usually requires more verbal patter to maintain the interest and attention of respondents. Our interviewers report, for example, that routine pauses in the conversation, which can be tolerated in a face-to-face situation, often seem interminable on the telephone. They also miss visual cues for deciding when to probe or pause for a more complete response. Such differences in the modes of interaction suggest that detailed information may be more difficult to obtain in telephone interviews. For example, Groves (1977) found a greater frequency of multiple responses to open-ended questions for the HH group, suggesting a greater facilitation of communication for household interviews. Accordingly, we compared the groups on number of responses to several open-ended questions, including (1) the number of reasons mentioned by the respondent for not seeing a doctor when a need was perceived (in the past 12 months); (2) the number of acute problems that the respondent reported having within two months of the interview; (3) the number of chronic health conditions reported by the respondent; and (4) the number of medications and treatments currently prescribed for the respondent by a physician. As shown in Table 4, none of these differences was significant, nor was there a significant mean difference on a composite measure created by summing across all open-ended responses to the above items (t = 1.39, n.s.).

A different picture emerges, however, when we examine counts of responses to checklist items. Questions of this kind included a list of ten preventive health behavior activities and six types of health advice given by family members and friends. In the HH administration, respondents were shown a card and asked to indicate which of the statements or activities in the list applied to them. For the TC sample, each response was read separately, after which respondents were asked if that particular response applied to them. In every case, the TC respondents gave significantly more responses to each checklist item than did the HH respondents. For the health advice items, we had intended that some of the items would be contradictory and that respondents would report on the usual or modal advice that they received. For example, respondents were asked if the advice that they *usually get* included "get medical advice as soon as possible" and "wait and see what happens," and we wanted to compare respondents receiving "act now" and "delay" advice from their social networks. In the TC sample, many more respondents endorsed both kinds of items, and we had respondents indi-

Table 4
Household-telephone comparisons for open-ended and checklist questions, 1976

Item	HH sample mean (N = 1,210)	TC sample mean (N = 303)	K-S test[a] D	t for mean difference
	Open-ended questions			
Number of reasons for not seeing doctor	.346	.314	.021	—
Number of acute problems mentioned	.346	.287	.045	—
Number of chronic problems mentioned	.800	.746	.029	—
Number of medications	.834	.838	.024	—
Sum of open-ended questions	2.326	2.185	.034	1.39
	Checklist questions			
Voluntary preventive health behavior	1.037	1.399	.118**	—
Kinds of family advice	.916	1.307	.238***	—
Kinds of friend advice	.607	1.059	.219***	—
Sum of checklist questions	2.560	3.766	.240***	8.30***

[a]D is the maximum absolute discrepancy between the two cumulative proportion distributions. Significance levels are for a two-tailed, two-sample Kolmogorov-Smirnov test of differences between the two distributions.

**$p < .01$.

***$p < .001$.

cating as many as five kinds of advice. As in the analysis of the attitude items reported above, this response difference between the two modes of interviewing suggests greater acquiescence in the telephone interviewing condition, in the sense that the respondents tended to endorse or agree with more of the checklist items. On the other hand, it may be that the aim of eliciting the usual or dominant kind of advice is communicated more successfully in the face-to-face situation.

Discussion

As shown above, we found that the telephone comparison sample had more missing data on family income; more acquiescence, evasiveness, and extremeness response bias; and more and somewhat contradictory answers to checklist questions. These are differences in response style rather than in item content (Jackson and Messick, 1958). For several reasons, we can probably rule out sampling fluctuations as an explanation for these response-style differences between the samples. Except for the mean difference on family income (which is equivocal in view of the missing-data difference on the same question), we found no demographic differences between the samples. Also, although the attitude items yielded ample evidence for response-style differences between samples, observed differences in attitude content were of marginal significance and were not robust under weighting.

These differences seem to indicate that the quality of the data obtained over the telephone was not as good as that obtained in the face-to-face situation, perhaps because TC respondents were insufficiently motivated or were not working hard enough during the interview. As Cannell, Oksenberg, and Converse observed in a recent National Center for Health Services Research report on interviewing:

> First, answering a question accurately and completely requires the respondent to use cognitive skills in comprehending a question, recalling or organizing and processing the relevant information, and finally in formulating an answer. Second, these cognitive activities often require considerable effort, which the respondent must be willing to exert. [U.S. NCHSR, 1977:14]

Exerting such effort, of course, is part of the respondent's role in survey research. Perhaps these differences pinpoint a special problem in telephone interviewing—that of motivating people to play the respondent role over the telephone. Support for this interpretation can be found in the literature. For example, Groves (1977) found that in telephone interviews there is more respondent suspicion, higher refusal rates, less detailed information in response to open-ended questions, and more evasiveness to sensitive questions. Yaffe et al. (1978) also found lower accuracy of reporting in telephone interviews, but only for an urban Baltimore sample and not for their more rural Washington

County sample.

Many of the specific field procedures and interviewer techniques that are traditionally used to stimulate respondent motivation (a personal style of interviewing, positive reinforcement through visual contact, introductory letters, etc.) are often more difficult to implement in telephone surveys. There are, of course, other possible explanations for the differences that were found between the two modes of interviewing. Thus, the greater tendency to select extreme categories, acquiesce, and provide evasive answers in the telephone condition might be related to the more technical side of telephone interviewing—to the fact that voice transmission over the telephone is imperfect, which could result in greater respondent confusion and misunderstanding; to the inability of the interviewer to use visual cues to discern when respondents need clarification or probing; to the lack of interviewer flash cards; to the typically faster pace of telephone interviews; and so on. These considerations call attention to the need for further research on telephone interviewing. Clearly, more research is needed to clarify the differences between the two modes and to explore the dynamics that account for these differences.

Telephone interviewing as a black box — Discussion: Response styles in telephone and household interviewing

Eleanor Singer, Center for the Social Sciences, Columbia University

The paper that we have just heard falls within a large class of studies that ask the question, "What differences in response are associated with the method of asking the question?" What I will try to do in the next few minutes is to (1) locate telephone/personal interview differences among other methods differences that have been investigated; (2) summarize briefly the results of comparisons between telephone and personal interviews; and (3) suggest that perhaps theory and research about at least some methods differences should take a different approach in the future.

Investigations of method-linked response differences

Comparisons of differences in response produced by interviews on the telephone and in person are only a recent instance of a long line of similar investigations. For example, we have examined differences associated with mail versus personal interviews; with the effect of interviewers' age, race, sex, social status, and expectations; with different sorts of interview introductions; and with differences in question wording, question order, and the like. Most of these methods, like the telephone, are "black boxes," in the sense that we know *that* they are associated with differences in response, at least some of the time, but we do not know *why* or *how* they bring about such differences. We have no clues to the process and therefore cannot control it.

There are a few notable exceptions, and they tend to come from small, controlled laboratory experiments. For example, 30-odd years ago Hyman (1954) and his colleagues at NORC showed that some interviewers tend to make respondents' ambiguous replies consistent with a series of related attitude questions, thus introducing a spurious consistency into the data by virtue of their expectation for consistency. In this case, there is no black box: we see the process quite clearly. Among some interviewers, the expectation for consistency leads to recording errors when responses are ambiguous. But most research on interviewer expectations—my own, for example—tells us nothing about how such expectations are translated into self-fulfilling prophecies. I submit that much methods research, and certainly most research on telephone interviewing, is of the black box variety; therefore, the conclusions that we draw from it are based on shifting sands, liable to reversal in the next study. I will come back to the implications of this after summarizing some of the conclusions that have been drawn about response differences associated with personal and telephone interviews.

Comparisons of telephone and personal interviews

Most of the recent reviews of the literature have stressed the comparability of telephone and personal interviews—the absence of large and significant methods differences between them. Yet, when I looked over the major articles cited, it seemed to me that not only are there some documented differences but these tend to show a consistent pattern. I will summarize the studies very briefly in chronological order and then categorize them according to their findings.

1. Larsen (1952) found that socially desirable behavior was overreported on the telephone, compared with face-to-face interviews.
2. Hochstim (1967) found very few differences by interviewing strategy, but the two differences that he notes suggest that people interviewed in person tended to give more socially desirable responses: Women were less likely to be candid about their drinking when interviewed in person than by mail or telephone and less likely to ac-

knowledge husband-wife discussion of women's medical problems.

3. Colombotos (1969) reported no statistically significant differences in the social desirability of responses given by physicians to interviews in person and by telephone; but, in one of his two studies, physicians gave more socially desirable answers to 8 of 12 items in personal, rather than telephone, interviews.

4. Rogers (1976:53) stated:

   > The quality of data obtained by telephone is comparable to that obtained by interviews in person. . . . If anything, the data suggest that those interviewed in person are somewhat more likely to give socially desirable answers than those interviewed by telephone. . . . Telephone data on reported education were more consistent, and on voting, more accurate, but face-to-face interviewing was more successful in obtaining income.

5. Wiseman (1972) found that the *lowest* number of socially *un*desirable answers was given on the telephone, compared with mail and personal interviews.

6. Henson, Roth, and Cannell (1974) reported that among Kansas City households, telephone interviewing yielded *less* frequent reporting of *un*desirable psychological states than did personal interviewing.

7. Locander, Sudman, and Bradburn (1976) found, in a Chicago sample, that there is less overreporting of socially *desirable* behavior in personal than in telephone interviews but less underreporting of socially *un*desirable behavior on the telephone than in person.

8. Johnson and White (1979) found that among a sample of Nebraskans aged 65 and older, telephone interviews were shorter; more respondents acknowledged ill health and health care seeking in personal interviews (but not significantly so); and telephone respondents tended to report lower income and less home ownership. There were no differences in reported satisfaction with some 20 different areas.

9. In a Texas study of respondents 55 and over (Stephens, 1979), there was some tendency for sensitive questions to be answered affirmatively more often in person rather than on the telephone, and some tendency for telephone respondents to be more reluctant to reveal personal information.

10. Herman (1977) reported more refusals to sensitive questions (how the respondent had voted in a union election) on the telephone than in person.

11. A study by Hoerner and Haas (1979) suggests that respondents whose attitudes have not crystallized are more likely to give socially desirable responses and to be less consistent in their responses on the telephone than in person.

12. Groves and Kahn (1979) report that telephone interviews tend to be shorter; that there is more missing data in telephone interviews; that respondents on the telephone tend to be more suspicious and less satisfied; that there is a greater frequency of multiple responses to open-ended questions in personal interviews; and that response rates are lower on the telephone than in person, with a much higher proportion of breakoffs.

13. Finally, the study reported in the paper by Jordan, Marcus, and Reeder tells us that the telephone sample had a lower response rate; more missing data for family income; more acquiescence, evasiveness, and extremeness response bias; and more responses to checklists (but not to open-ended questions). There were, however, no substantive differences in responses to attitude items, apparently because the scales were balanced for positive and negative responses.

If we take these findings together, I think the prudent conclusion would be that, until now, researchers have paid a price for using the telephone to interview people in terms of a whole series of components of response quality: overall response rate, item nonresponse, acknowledgment of sensitive behavior, response biases of various sorts, and respondent attitudes toward the interview. With the exception of the studies by Hochstim, Colombotos, Rogers, and, in part, Locander et al., such differences as exist favor the personal interview. But why is this the case? What do the findings mean?

The fact of the matter is that we really don't know. For instance, what about telephone interviewing is associated with a lower response rate? We can think of several possibilities.

1. One possibility that comes to mind is that, as in the study discussed in the paper by Jordan and his colleagues, interviewers had much more experience with personal interviews and indeed preferred them. The "method" effect is, therefore, likely to be in substantial part an interviewer expectation effect. Even if it is, how does it come about?

2. A second possibility is a variant of the "foot-in-the-door" explanation: Perhaps, having opened the door, it is harder to close it than it is to hang up the telephone, once having picked it up. Essentially, this explanation says that respondents are predisposed *not* to give an interview, and they find it easier to act on their predispositions on the telephone than in person. Thus, if motivation is high (or can be increased), telephone interviews will get as high a response rate as personal interviews; where motivation is low, personal interviews will have an advantage.
3. A third possibility is similar to the second in assuming that telephone interviews are more difficult to get than in-person interviews and that additional experience on the part of interviewers is required. Since the interviewers used in experimental telephone surveys are often newly hired for the study, they do worse on the telephone than in person.

When the issue is response quality, rather than overall response rate, we are confronted by a similarly bewildering array of possibilities. The authors of several of the publications cited above, for example, wonder whether the differences in response are, perhaps, a function of the variations in response rate—in other words, whether the people reached by telephone differ from those reached by personal interviews, so that the responses differ as a function of differences between the samples rather than as a function of differences between interviewing methods. Although very few of the comparisons have turned up demographic differences between samples interviewed on the telephone and in person, this is no guarantee that other, more subtle differences of attitude and response style do not exist. On the contrary, it is possible that the people who agree to a telephone interview are more acquiescent; if they were not, they would have refused!

Whither future research?

Summing up the discussion of their findings, Jordan and his colleagues point to several reasons for the effects that they observe. They note, for example, that interviewers often found routine pauses in the conversation, which can be tolerated in a face-to-face situation, intolerable on the telephone and that they missed visual cues for deciding when to probe or to pause for a more complete response. They speculate that

> the greater tendency to select extreme categories, acquiesce, and provide evasive answers . . . might be related to the more technical side of telephone interviewing—to the fact that voice transmission over the telephone is imperfect, which could result in greater confusion and misunderstanding; to the inability of the interviewer to use visual cues to discern when respondents need clarification or probing; . . . to the typically faster pace of telephone interviews; and so on. These considerations call attention to the need for further research on telephone interviewing.

The question is, what direction shall this research take? It strikes me as possible that advances here will come not from additional conventional comparisons between telephone and personal interview surveys but from controlled laboratory experiments. They are small; they are affordable; they permit the monitoring of interactions through videotaping and the like; they permit the sequential testing of cumulative hypotheses in a fairly short period of time.

In moving to the laboratory, we have a fairly substantial body of research awaiting synthesis and extension to the interviewing situation. Some of this research has been summarized by Williams (1977). I will give you just a few of the tantalizing findings that he reports, with the caution that some of these, too, are of the black box variety:

1. In cooperative problem-solving tasks with objective solutions, there are no differences between audio-only and face-to-face conditions in the accuracy with which solutions are achieved, but there are differences in the processes by which solutions are reached, for example, in the number of solutions discussed. On the average, face-to-face sessions last slightly longer.
2. Experiments on the effect of the medium of communication on the balance of cooperative and competitive interactions suggest that affect of all sorts, whether friendly or unfriendly, is amplified in face-to-face, compared with audio-only, conditions.
3. On the other hand, in experiments in which subjects meet two strangers via two different communication modes and are then asked to evaluate the conversation and the stranger, face-to-face conversations were preferred over telecommunicated ones and audio-visual conversations over audio-only. This series of experiments, and some others, suggest, according to Williams, that media richer in nonverbal cues lead to more favorable first impressions.
4. The hypothesis that there are verbal substitutes for visual cues was tested by Cook and

Lalljee with little success, and it remains an open question whether such substitutions in fact occur. However, even if they do not occur spontaneously, perhaps they can be deliberately created.

5. One hypothesis that has received some support is that in face-to-face communication we see others as real social beings, whereas over the more "distant" media, such as the telephone, we treat others as "more like semimechanical objects, which can be ignored, insulted, exploited, or hurt with relative impunity." Some evidence for this comes from the analysis of transcripts of conversations over various media; in a variety of studies, audio-only conversations were shown to be more depersonalized, argumentative, and narrow in focus, compared with face-to-face conversations.
6. There is some evidence that visual cues in a face-to-face situation serve as distractions, so that, e.g., more persuasion takes place in an audio-only situation, and lying is apparently easier to detect in such a situation.

These tantalizing findings provide only a beginning for the kind of research needed to dissect telephone interviewing as a black box. Further, I think the challenge of penetrating the black box of telephone interviewing offers unusual opportunities to shed new light on fundamental social-psychological processes such as the rules of acquaintanceship, the norm of reciprocity and the penalties for its violation, and other equally engaging topics.

Computer-based interviewing systems

The open discussion on Groves's paper was of two types: questions about computer-based interviewing systems and discussion of the significance and implications of such systems for data collection procedures.

In response to a query on how open questions are handled, Groves said that on most systems answers to open questions are typed right into the terminal. There seem to be two different ways for handling their storage and coding. In the Michigan system, the answers are stored in a separate file and tape-coded at a later date so that the processing of open questions does not delay dealing with the closed data. The two files are integrated when open coding is complete. Some other systems enter open responses into the same file as the closed data, using variable-length records.

When asked how the system in Michigan (at the Survey Research Center) compares with the UCLA system, Groves responded that the two systems evolved in different environments and for that reason have different design features. The SRC system in its current edition requires the machine to do file management work that the UCLA system did initially. (Groves noted that such systems are generally changing and this might not now be true.) Since the machine is doing a great deal of on-line processing of the sample base and interview data, response time on the SRC system is worse than at UCLA.

The next question concerned how computer-based interviewing systems affect the time needed to develop a questionnaire, administer a questionnaire, and process data. Groves responded that because this was the Survey Research Center's first study using a computer-based interviewing system, everything took longer, and that with more experience, the time should decrease. Cannell noted that, in particular, researchers had to learn to think about a questionnaire in a different way. The computer-based interviewing systems involve issues that researchers are not used to dealing with.

A key variable, according to Taylor, is the extent to which the researchers can build in detailed control over interviewer behavior. He thought that if the only goal is to produce a set of questions for interviewers to ask, questionnaire development could go as fast or faster than it does for traditional studies. However, as the goals come to include more complex control over interviewer activities, such as building in feedback, it takes longer to develop the protocol.

Groves added that although the field period for a study using a computer-based interviewing system does not differ from any other telephone study, the direct data entry has the potential to reduce the time needed for data reduction. He noted, however, an important tradeoff regarding data checks. Computer processing is slowed up if very many checks are done beyond ascertaining whether an entered response is a legal code, and this could interfere with the flow of the interview.

In response to a question on the kind of computer needed for a computer-based interviewing system, Groves said that SRC has moved to a minicomputer system supporting up to 14 terminals and having 192K of storage capacity. It is likely, in his opinion, that smaller and smaller computers specially geared to interviewing would be used, even to the point of having each interviewer using separate and independent microprocessors rather than hooking up to large central computer facilities.

These questions were supplemented by a discussion of several more general topics. One general focus of the discussion was the applicability of computer-based interviewing systems. Groves (in his paper) and Taylor (in his comments) had discussed the problems that had been encountered in making the systems operational and cost-effective. Freeman, in talking about the UCLA system, amplified on their

comments. He noted that the UCLA computer-based interviewing system was much more expensive to develop than they had anticipated and that there were many more problems than they had initially envisioned. He also noted that their goal had not necessarily been to set up a cost-beneficial system. They had been in an experimental mode to identify applications and problems with such systems. He said that a lot of their difficulties stemmed from simple (or not so simple) communications problems between programmers and survey researchers. Over time, they might well be worked out.

Freeman also said that it was possible to set up a system that would be simpler than the UCLA system and that would probably have fewer problems. He went on to note that even well-developed computer programs, such as SPSS and SAS, do not work perfectly. It may be inappropriate to judge any computer-based interviewing system during these early stages of development.

de la Puente also emphasized that the idea of computer-based interviewing should not be rejected too soon. Telephone interviewing generally is one of the most important interviewing methods currently being developed. Because there are numerous questions about when and how to use it, it is essential that we learn more about the underlying issues.

Horvitz amplified this point. Traditional surveys have been around for 30 or 40 years, and the bugs are not yet worked out. It therefore is not reasonable to judge computer-based interviewing systems at this early stage. He hypothesized that there may be some kinds of studies for which the computer-based system is much better than the alternatives, and there may be others for which it is not cost-effective. For example, in the National Medical Care Expenditure Survey, it was necessary to provide feedback to respondents' reports prior to the next interview. A direct data-entry system might well have solved problems associated with the process much better than they could have been solved in any other way.

Cannell noted that we are also still in the infancy of developing our understanding of how to carry out telephone interviews in general, with or without computer assistance.

Sudman responded that we do not know everything about interviewers and interviewing on the telephone, but there are some generalizations that can be made that apply equally to telephone and face-to-face interviews:

Intelligence is a virtue in all interviewers.

The good telephone interviewers may not be the same people as the good face-to-face interviewers.

There are large performance differences among interviewers, whether in person or on the telephone. We do not know very much about what causes those differences, although individuals' need for achievement may be a factor.

There is always a great deal of turnover among new interviewers. Typically, less than half of a newly trained staff will actually be productive. Attrition problems are not at all unique to interviewers on computer-based interviewing systems but are common to all interviewers.

At this point, Fuchsberg described what the National Center for Health Statistics is doing with telephone interviews as an example of the potential importance of this area of inquiry.

In a telephone survey on smoking that NCHS has just begun, they are using a staff of interviewers that they hired and trained themselves. Their experience, like that of others, is that interviewers learn quickly and are getting response rates in excess of 80 percent. This study is not computer assisted, although NCHS is very interested in the applications of computer-based interviewing for the future. The current survey will be compared with a similar personal interview survey as one aspect of the evaluation of telephone interviewing in the overall data collection efforts of NCHS.

Singer raised the question of how interviewers feel about using computer-based interviewing systems. Intelligence is always important to being a good interviewer, but computer-based interviewing limits interviewer discretion. Indeed, that is cited as one of its great virtues. In addition, interviewers are, or can be, closely monitored. How will that affect interviewer morale? Is evaluation of these issues part of the Groves-Cannell study?

Groves responded by saying that this particular study was not aimed at assessing these issues. He has heard that Chilton (a commercial firm that has used computer-based interviewing systems extensively in its surveys) decided to reduce machine processing and give some responsibilities back to the interviewers (e.g., dialing the telephone) because of feedback that the interviewers were dissatisfied.

Cannell then noted that these issues were very relevant to Miller's paper and that he and his colleagues had long been concerned about uncontrolled interviewer behavior. In particular, Marquis had pointed out that interviewers were reinforcing respondent behaviors in random

ways or worse. For example, an uncooperative respondent, or one who was not trying, was likely to get more reinforcement than one who was doing exactly what the researcher wanted.

As a result of observations such as these, procedures have been developed and tested to structure interviewer behavior so that they "train" respondents to behave in more consistent and constructive ways. These procedures include (1) building in standardized reinforcement for desirable respondent behavior; (2) minimizing all other reinforcement by having interviewers provide respondents with a standardized set of instructions stressing accuracy and good reporting, in order to standardize role expectations; and (3) having respondents sign a commitment form agreeing to try hard and to be accurate.

These efforts were all directed at increasing the similarity and consistent effectiveness of interviewer behavior and in reducing interviewer discretion. A computer-based interviewing system is only one means to achieve this.

Interviewing techniques in mass media research

Sudman began the general discussion on Miller's paper by pointing out that many studies have used "more is better" as the criterion for success. In studies using reinforcement, one needs to be concerned that the researcher is not artificially creating reports or verbal behaviors that are not "better."

Miller noted that this was a particular concern of his. A feature of his study that he considered important was that his hypotheses did not always involve reporting "more." For example, respondents in the experimental condition reported reading fewer books than did those in the control group and also reported less exposure to the newspaper editorial page. Since both behaviors are positively valued, one assumes a tendency to *over*report them. Further, the presumed overreporting tendency concerning books read is reduced by the experimental techniques even more markedly for well-educated respondents because these people are likely to place a very high value on book reading. The substantially lower mean score for "Hi-Ed" respondents in the experimental group is further support for the hypothesis that the experimental interviewing procedures produce more valid, and not simply more, reporting.

Marquis said that in the initial pilot work on developing reinforcement techniques in interviews, a social desirability scale was used. Respondents who were systematically reinforced for accurate reporting scored lower on the social desirability scale.

Cannell pointed out some of the difficulties in creating a standardized reinforcement interview. First, researchers have to work hard to creat an administerable questionnaire that reinforces the *process* of being a good respondent but not the *substance* of any particular kind of answer. Second, it is very hard for interviewers not to produce random reinforcement. Interviewers believe that they will alienate respondents if they do not give reinforcement. They are always surprised when they find that respondents are not bothered at all when interpersonal feedback is curtailed.

Fowler added that the discussion was focusing on reinforcement, but the treatment involved a good deal of standardizing expectations as well. The standardized interviewer instructions to respondents and the respondent commitment form were designed to make it clear—to both interviewer and respondent—what kind of interaction the interview was to be and what the respondent was supposed to do. Indeed, reinforcement too could be thought of as communicating expectations as much as making the respondent feel good. It should not be overlooked that one of the major problems the procedures were designed to address was that interviewers communicate very different expectations and standards to respondents.

Horvitz picked up the point that a critical outcome of the approaches being discussed should be to reduce interviewer variance. Cannell said that such a test was one of the important goals of the current SRC computer-based interviewing experiment. They found that one of the great virtues of the system for those interested in doing methodological research was the ease of generating randomized assignments to interviewers.

Dohrenwend questioned whether the procedures, particularly the goal of standardizing interviewer behavior, would apply to studies of mental health. She cited a study in which respondents were asked whether they had considered suicide. One-third of the interviews were taken by psychiatrists, the balance by professional survey interviewers. The rate at which respondents said that they had considered suicide was 25 percent when the interviewer was a psychiatrist and only 4 percent when a survey interviewer was used (Reissman, 1979). Dohrenwend felt that the experience of the psychiatrists was critical in this situation.

Marquis noted that it was likely that there were many other cues or signals being sent out by both groups of interviewers.

Cannell said that the SRC procedures were developed to address two critical underlying assumptions, based on research findings: First, respondents generally are not very motivated and are not geared to work very hard. Second, interviewers are likely to err in the direction of not asking for more than the minimum, sometimes even reinforcing minimal behavior. Although the procedures are directed to these key problems, this does not mean that there are not other problems that can affect the quality of data collected.

Verbrugge wondered if the results of the suicide study mentioned by Dohrenwend might have something to do with the fact that the psychiatrists were male. Gift noted, however, that sex of interviewer has seldom been associated with response error if other factors, such as level of experience, are controlled.

Fleishman asked again whether interviewers feel alienated from their work when their behavior is more controlled by procedures such as those used in computer-based interviewing systems.

Cannell responded that there are not systematic data on that topic from the SRC study. Anecdotally, he said that he has observed that some interviewers are intrigued by the gadget and seem to enjoy the whole process; others are frustrated and find interacting with the console to be tension producing. Some interviewers, he noted, clearly express a sense of relief at the reduced responsibility. In traditional interviewing, responsibility rests clearly on the interviewer for all of his or her behaviors and for the results obtained. If the researcher tells the interviewer what number to call, when to reinforce the respondent, and when to say nothing, the results of the interview become less the interviewer's responsibility. This provides less chance for a sense of success, perhaps, but also less chance for failure.

It is clear that the task of the personal interviewer in a face-to-face household survey is very different from that of the interviewer in a computer-based interviewing system. It is altogether reasonable that a person would not necessarily be equally suited to both. The critical question for the future is whether there are capable people who can and will do a good job of computer-based interviewing. Although there is still much work left to be done, the preliminary answer to date clearly seems to be "yes."

Response styles

The discussion of Jordan's paper started from Singer's basic conclusion that the studies comparing telephone and personal interview procedures generally produced a few small differences. Athough in each study these differences often appeared to her to be of minimal importance, when she looked over all the studies, they began to suggest, as the Jordan, Marcus, and Reeder study did, that there were some real differences and that the reasons for those differences were not at all well understood.

Bergner raised the question whether some of the differences could be studied by looking at interactive patterns using tape recordings of interviews. No one knew of any systematic attempt to compare the interactive patterns on the telephone and in person.

Warnecke noted that Jordan's findings of increased acquiescence on the telephone challenged the notion that increased social distance helps to minimize social desirability effects.

Actually, the data on such topics — such as whether mail procedures are better for sensitive information — are not at all consistent. Singer went on to make the same point with respect to telephone studies. When she looks across all the studies, she concludes that the weight of evidence is that telephone surveys do a bit less well than personal strategies for the collection of sensitive data. She notes the possible relevance of laboratory studies indicating that the telephone is a less "warm" form of communication.

Singer emphasized that her main concern is that the evidence is contradictory and that studies may not all be comparable. However, we do not know why different studies obtain different answers; we do not know what the relevant variables are.

Oksenberg drew attention to the importance of the differences in format between the telephone and personal interview forms for the items on which Jordan's key comparisons were made. On the telephone, response alternatives were read, while a show card was used when the interview was in person. Thus, the mode of presentation could be the key to the differences. Jordan agreed that the stimuli were different in this respect, as he had noted in the paper.

It was suggested that we need to understand much more about which kinds of items work best on the telephone and which work best in person, since they may be different. A related suggestion was that since the extremity of response was found to be greater on the telephone, perhaps due to the oral presentation of a 4-point scale, comparability could be obtained by asking the question in two parts: e.g., first asking agree-disagree, then asking whether or not the respondent felt strongly about the matter.

Walden then asked perhaps the critical question of this discussion. He stated that the paper being discussed, and the differences discussed, dealt mainly with attitude items and ordinal scales. In his work he has primarily been concerned with factual items—reporting of events and behaviors. He asked to what extent the group felt that such items produced different data by personal interview and by telephone.

Singer noted that there was some limited evidence that fewer health conditions were reported on the telephone. Cannell said that the studies were not clear on the point. However, his research has shown that reporting of such things as visits to doctors and health conditions were responsive to a wide range of factors in the way that an interview was carried out. It was almost certain that the issues being discussed in the interview affected the reporting of factual information on the telephone.

Fuchsberg said that he was convinced that the telephone procedures, particularly those using random-digit-dialing samples, had two critical problems that were different from personal interviews. First, interviewers had more trouble establishing legitimacy. Second, they had more trouble establishing rapport.

Cannell noted that it was easier to refuse on the telephone than in person. (However, this contradicts the fact that response rates seem comparable when there is an advance letter. In fact, one has the impression that response rates are about the same for random-digit-dialing studies as for national personal interview studies.)

Singer reemphasized that the number and size of the differences that she found in her review were not large. However, the persistence of some types of differences simply seemed to her to require some further thought and to demand questioning of the growing view that there were no differences between telephone and personal interview data.

The discussion on Jordan's paper closed with a comment from Patrick on the use of telephones in surveys in Britain. He stated that such surveys are not feasible in Britain. One reason is the comparatively low rate of telephone ownership. In addition, though, he said that a woman would not answer the telephone if she was home alone. The meaning of a telephone call can vary cross-culturally or within a culture. We do not know enough about what the telephone means to people, how that meaning varies, or how talking on the telephone enhances or constrains interaction. Like any other mode of interaction, there are, no doubt, people who prefer it and people who avoid it. The point is that we do not understand what the telephone does to interactive patterns and communication patterns. There appeared to be general agreement that simply treating the telephone interview as if it was the same as the personal interview was unrealistic. Yet, so far we do not know much about how to deal with it differently. Studies have shown that just doing a personal interview on the telephone produces comparable data some of the time, for some people, and for some kinds of items. However, it does not work all the time for all items—a reasonable enough conclusion. We now need to move ahead to understand the nature of the telephone interview so that we can better take advantage of this important research mode.

Recommendations

Three interrelated themes were dealt with in Session 2: (1) the uses and limitations of computer-assisted telephone interviewing; (2) the uses and limitations of telephone interviews more generally; and (3) the value of increased control of interviewer behavior in the interview, with telephone interviewing in general and computer-assisted telephone interviewing in particular providing a powerful mechanism for gaining that kind of control.

With respect to computer-assisted telephone interviewing, a number of technical issues need to be worked out. In particular, the most cost-effective mix between tasks needs to be developed. This mix should be assessed from the point of view of start-up and processing time, from the point of view of the effect of computer operations on the interviewer-respondent interaction, and from the point of view of the impact on interviewers' morale and satisfaction with their jobs. At this time, the potential of computer assistance cannot be adequately assessed on any of these criteria. Clearly, the whole area needs quite a bit more experience before it will be possible to make sound judgments about the best kinds of computer-assisted telephone interviewing systems, let alone make more detailed statements about the kinds of projects and applications that are or are not appropriate for such interviewing.

With respect to telephone interviewing, the conference raised some issues that have been discussed in previous conferences and also added some new perspectives. In general, the previous conferences had concluded that the weight of evidence was that if reasonably equivalent sampling schemes could be used, telephone surveys would produce data on most items that were as accurate as those data collected by personal interview. This conclusion was based on findings from several carefully

conducted studies. However, in most comparisons between personal and telephone interviews, there were a few differences. This conference took another look at those differences, and a renewed call was made for careful study of possible differences between telephone and personal interview data. In particular, the following issues emerged:

1. While some studies have found almost no differences between the aggregate figures based on telephone and personal interviewing procedures, others showed some tendency for sensitive data to be less well reported on the telephone. Although these differences are not very large, to understand why there are study-to-study variations further research is needed to produce better generalizations about when personal and telephone procedures are interchangeable and when they are not.
2. In particular, there is a need for better understanding of the interactive patterns on the telephone compared with those in personal interviews. It seems almost certain that the interaction is affected by the mode of communication. Understanding better how the interaction differs will open the way to designing more effective telephone procedures and training procedures for interviewers.
3. Research is needed to identify what differences, if any, exist between "good" telephone interviewers and "good" personal interviewers. Some interviewers have a definite preference for one mode of interviewing over the other and have styles that seem particularly effective in one kind of interviewing or another. There has as yet been virtually no research on these matters; such research would be helpful.
4. Response rates are of particular concern in telephone interviewing. In particular, there is a difficulty with the random-digit-dialing strategies, for which no advance letter is possible. Telephone response rates are not always lower than those for personal interviews. There are certain groups, such as those with good security systems or urban populations in general, who respond more often to telephone interviews than to personal interviews. Nonetheless, research on effective ways of presenting telephone interviews is needed in order to improve further the value of telephone interviewing strategies. High response rates are particularly important in telephone samples because 7 percent of the population is excluded since they have no telephones.

Finally, this conference included a paper on an application of experimental attempts to apply increased control of interviewer behavior to media research. In the preceding two conferences, there was discussion of the need to develop practical procedures for structuring interviewer behavior, making that behavior a positive force rather than a random or even negative force in the production of good respondent behavior. Miller's paper provided further evidence of the potential value of such efforts. Yet we still are very far from having a set of accepted practical guidelines for incorporating the results of experimental studies into standard interviewing procedures.

SESSION 3:
Use of informants and multiplicity estimates and the use of diaries and panels in health services research

Chair: Seymour Sudman, Department of Business Administration and Survey Research Laboratory, University of Illinois at Urbana-Champaign

Recorder: Richard B. Warnecke, Department of Sociology and Survey Research Laboratory, University of Illinois at Chicago Circle, and Illinois Cancer Council

Network sampling in health surveys

Monroe G. Sirken, National Center for Health Statistics

Introduction

Network sampling is being used in surveys of medical providers (Sirken, 1975) and in household surveys (Sirken, Graubard, and McDaniel, 1978) to estimate disease incidence, prevalence rates, and the cost of illness. In describing network sampling, it is convenient to pose three questions:

1. How are networks defined?
2. What is the unique feature of network sampling?
3. What is the estimation procedure?

In network sampling, the networks are defined by counting rules (Sirken, 1974). These rules provide the links between the individuals in the population and the enumeration units where they are eligible to be enumerated in the survey. By virtue of these rules, an individual is linked to a network containing the enumeration units that are eligible to report him, and conversely an enumeration unit is linked to as many individuals as it is eligible to report. For instance, the rules used in network surveys of medical providers are based on transactions between individuals and their medical providers; the rules used in household network surveys are based on relationships, such as kinship or friendship, between individuals in the population.

An essential difference between network sampling and traditional sampling is the network size. Traditional sampling is based on counting rules that uniquely link each individual to one and only one enumeration unit where he is eligible to be enumerated in the survey. The de jure and de facto residence rules are examples of counting rules used in household surveys based on traditional sampling. Rules that uniquely link each patient to one medical provider, such as his first or principal provider, are examples of rules for medical provider surveys that are based on traditional sampling. Network sampling, on the other hand, is based on counting rules that do not necessarily imply that an individual is uniquely linked to a single enumeration unit. Kinship and friendship rules are examples of counting rules in household surveys that are based on network sampling, and rules that link patients to several medical providers are admissible in medical provider surveys based on network sampling.

Since counting rules used in network sampling do not necessarily imply that an individual is uniquely linked to an enumeration unit, some individuals may not be linked to any units and others may be linked to several units. Both possibilities require careful attention to avoid bias.

One of the easiest and most direct ways of controlling counting rule bias when the rule fails to link some persons to enumeration units is to incorporate a traditional counting rule within the network rule. For instance, network sampling based on a rule that links individuals to the households of their siblings would miss individuals without siblings. The problem could be handled by modifying the counting rule so that individuals were linked to their de jure households as well as to the households of their siblings. The proposed modified rule would make individuals without siblings eligible to be enumerated at their own households.

Several unbiased estimators for network sampling have been proposed to handle the problem where several enumeration units are linked to the same individual (U.S. National Center for Health Statistics, 1965b). One of these estimators, the multiplicity estimator, is defined in the appendix. It is noteworthy that the estimator requires information about the network size of every individual who is counted in the sample enumeration units. The supplementary information is usually collected in the survey from the sample enumeration units where the individuals are enumerated. For instance, the

household reporting an individual in a survey based on a sibling rule would also report the number of other households eligible to report the individual because they were residences of the individual's siblings.

It is easy to confuse counting rules with respondent rules, since both rules are design features of household surveys that specify which individuals are eligible to report for one another. Respondent rules specify the conditions, if any, under which a proxy respondent is eligible to report the survey information for an individual selected in the sample. For instance, health interview surveys often adopt proxy respondent rules that permit individuals to serve as respondents for relatives in their households who are absent. On the other hand, counting rules specify the conditions, if any, under which an individual selected in the sample may serve as an informant for other individuals. In other words, an informant, as defined by the counting rule, is a sample individual who is eligible to report for other individuals; a proxy respondent, as defined by the respondent rule, is an individual who is eligible to respond for a sample individual.

Examples of counting rules

Table 1 presents a fictional population of 11 persons residing in six households. The relationships of individuals within households are listed in Column 3, and the relationships of the 11 individuals to the head of household A are listed in Column 4.

Table 2 lists five counting rule options for enumerating the 11 individuals shown in Table 1. Option I is the de jure residence counting rule used in traditional sampling. The remaining counting rules represent variations of kinship counting rules that might be used in network sampling. Table 3 lists the individuals in Table 1 who are enumerable at each household according to the five counting rule options, and Table 4 lists the households eligible to report each individual by these options. For instance, in compliance with Counting Rule III, B_1 and C_1 are eligible to be reported at household A; A_3 and C_1 are eligible to be reported at household B; A_3 and B_1 at household C; C_2 at household D; and D_2 at household E. Thus, A_3, B_1, and C_1 are eligible to be enumerated at two households, although not the same two households, and C_2 and D_2 are each eligible to be enumerated at one household. It is noteworthy that according to Rule III, six individuals, A_1, A_2, A_4, D_1, E_1, and F_1, are not eligible to be enumerated at any households. Similarly, 7 of the 11 individuals are not enumerable according to Rule II, and two individuals are not enumerable by Rule IV. On the other hand, all 11 individuals are eligible to be enumerated by Rules I and V.

Let us assume that a household sample survey is conducted of the population displayed in Table 1 and that the sample contains two households, namely, A and E. Let us now compare the information that would be reportable at these households if they were selected by traditional sampling based on Counting Rule I and by network sampling based on Counting Rule V.

In traditional sampling, individual A_1, A_2, A_3, and A_4 are enumerable at household A, and individual E_1 is enumerable at household E. In network sampling based on Rule V, B_1 and C_1 are enumerable in addition to A_1, A_2, A_3, and A_4 at household A, and D_2 is enumerable in addition to E_1 at household E. The multiplicity es-

Table 1
The fictional population

(1) Households in the frame	Individuals (2)	Relationships to household head (3)	Relationships to A_1 (4)
A	A_1	Head	Self
	A_2	Wife	Wife
	A_3	Son	Son
	A_4	Father	Father
B	B_1	Head	Son
C	C_1	Head	Son
	C_2	Wife	Dauthter-in-law
D	D_1	Head	Father of daughter-in-law
	D_2	Wife	Mother of daughter-in-law
E	E_1	Head	Grandfather of daughter-in-law
F	F_1	Head	Unrelated

Table 2
Kinship counting rules

Option	Rule
I	A household is eligible to report its residents.
II	A household is eligible to report nonresident siblings of its residents.
III	A household is eligible to report nonresidents siblings and children of its residents.
IV	A household is eligible to report nonresident siblings, parents, and children of its residents.
V	A household is eligible to report its residents and nonresident siblings, parents, and children of its residents.

Table 3
Individuals eligible to be reported by households in compliance with five counting rules

Household	Rule I	Rule II	Rule III	Rule IV	Rule V
A	A_1 A_2 A_3 A_4	B_1 C_1	B_1 C_1	B_1 C_1	A_1 A_2 A_3 A_4 B_1 C_1
B	B_1	A_3 C_1	A_3 C_1	A_1 A_2 A_3 C_1	A_1 A_2 A_3 B_1 C_1
C	C_1 C_2	A_3 B_1	A_3 B_1	A_1 A_2 A_3 B_1 D_1 D_2	A_1 A_2 A_3 B_1 C_1 C_2 D_1 D_2
D	D_1 D_2	—	C_2	C_2 E_1	C_2 D_1 D_2 E_1
E	E_1	—	D_2	D_2	D_2 E_1
F	F_1	—	—	—	F_1

Table 4
Households eligible to report individuals according to five counting rules

Individuals	Rule I	Rule II	Rule III	Rule IV	Rule V
A_1	A	—	—	B C	A B C
A_2	A	—	—	B C	A B C
A_3	A	B C	B C	B C	A B C
A_4	A	—	—	—	A
B_1	B	A C	A C	A C	A B C
C_1	C	A B	A B	A B	A B C
C_2	C	—	D	D E	C D
D_1	D	—	—	C	C D
D_2	D	—	E	C E	C D E
E_1	E	—	—	D	D E
F_1	F	—	—	—	F

timator also requires households A and E to report the number of other households eligible to report each of their respective reportable individuals. Thus, in the survey based on Counting Rule V, household A would report that two other households, namely, B and C, are eligible to report A_1, A_2, A_3, B_1, and C_1 and that no other households are eligible to report A_4; household E would report that two other households, namely, C and D, are eligible to report D_2 and that one other household, D, is eligible to report E_1. The estimator, which is defined in the appendix, indicates how the individuals enumerated at the sample households, A and E, would be weighted to obtain unbiased estimates.

Circumstances favoring network sampling

Network sampling offers design options that are not available in health surveys based on traditional sampling. Sometimes these options can substantially improve the design of surveys for which traditional sampling methods are inefficient. Like other sample design options, network sampling is not always advantageous, and it is more advantageous under some conditions than under others. Specific conditions that favor network sampling in health surveys are the following:

1. The health condition is relatively rare.
2. Some individuals are not covered by the sampling frame.
3. The health condition is a sensitive or threatening matter.
4. Individuals are not uniquely linked to enumeration units.

Conditions 1, 2, and 3 often arise in health interview surveys, and condition 4 often arises in medical provider surveys. Each of these four conditions is briefly discussed below.

Rare health conditions. Most serious health conditions are relatively rare. For instance, the prevalence rate for diabetes is less than 5 percent, and the cancer prevalence rate is only about 1 percent. On the average, more than 30 households would have to be enumerated in a survey based on traditional sampling in order to count one person with cancer. Using network sampling based on kinship counting rules increases the expected number of diabetics two or

threefold and often, but not necessarily always, substantially improves the reliability of the survey estimates.

Incomplete sampling frames. Since institutionalized individuals are not covered by sampling frames of household surveys, they are not counted in health interview surveys based on traditional sampling methods. For instance, about 22 percent of the deaths among those 65–84 years of age in North Carolina during 1975 occurred in institutions, and hence these individuals would be missed in the survey even if they were eligible to be enumerated at their former residence. However, nearly all institutionalized deaths in North Carolina are survived by close relatives and hence would be reportable by them in health surveys using network sampling based on kinship rules (Sirken and Royston, 1977).

Sensitive issues. Sensitive health conditions seem less likely to be reported in health surveys by the individuals with these conditions than by their close relatives or friends. For instance, estimates of heroin use that are based on reports of the users' close friends when network sampling is used are substantially larger than the heroin use estimates based on the self-reports of heroin users in conventional sampling (Sirken, 1979). The explanation may be that the anonymity of substance users is better assured when they are reported by their friends than when they report themselves. In contrast to health surveys based on traditional sampling using the de jure residence rule, health surveys based on network sampling using a close friend counting rule do not directly confront the heroin user and do not require that the close friends disclose the users' identities.

Multiple links. Diagnosis, treatments, and care for a seriously ill patient often involves many transactions with several medical providers. Thus, a patient is enumerable several times in a medical provider survey if all of his medical providers report him. Unfortunately, it is impossible to produce unbiased survey estimates by eliminating duplicates of the same patients who are reported by different sample medical providers. Network sampling presents a design strategy for coping with this problem.

Concluding remarks

Network sampling seeks to improve the design efficiency of health surveys by utilizing information that various enumeration units are able to report about the same individuals either by virtue of the social networks among these individuals and or by the transactions that they have with their medical providers. Thus, health interview surveys based on network sampling utilize information that close relatives and friends are able and willing to report about each other. Medical provider surveys based on network sampling utilize information that several medical providers are able and willing to report about the patients whom they have in common. Applications of network sampling to health surveys, under the specified survey conditions noted previously, have been encouraging. Network sampling often improves the quality of the survey estimates obtainable by traditional sampling methods. It frequently reduces the sampling errors and sometimes also reduces the nonsampling errors.

Appendix: The multiplicity estimator and its variance

Let $\Omega_I=\{I_1, \ldots, I_\alpha, \ldots, I_N\}$ represent a population of N persons with a specified disease. Counting rule r links the I_α ($\alpha = 1, \ldots, N$) individuals in Ω_I to the enumeration units (i.e., households or medical provider establishments) where they are enumerable in a survey. Let $\Omega_H=\{H_1, \ldots, H_i, \ldots, H_M\}$ represent a frame of M enumeration units. The links between the I_α in Ω_I and the H_i in Ω_H as defined by rule r are specified by the indicator variable,

$$ {}_r\delta_{\alpha i} = \begin{cases} 1 \text{ if } H_i \ (i=1, \ldots, N) \text{ is eligible to} \\ \quad \text{report } I_\alpha \ (\alpha=1, \ldots, N) \text{ by rule r,} \\ 0 \text{ otherwise.} \end{cases} $$

A random sample of m households is selected from Ω_H. The I_α ($\alpha=1, \ldots, N$) individuals eligible to be reported by the m households in compliance with rule r are enumerated in the survey. The multiplicity estimator, one of several network estimators that have been proposed (U.S. NCHS, 1965b) to estimate N, is

$$ \hat{N}_r = \frac{M}{m} \sum_{i=1}^{M} a_i \, ({}_r\lambda_i), \qquad [1] $$

where the Bernoulli variable

$$ \alpha_i = \begin{cases} 1 \text{ if } H_i \ (i=1, \ldots, M) \text{ is sampled,} \\ 0 \text{ otherwise,} \end{cases} $$

and the variate

$$ {}_r\lambda_i = \sum_{\alpha=1}^{N} {}_r\delta_{\alpha i} W_{\alpha i} $$

is the weighted sum of the I_α ($\alpha=1, \ldots, N$) that are eligible to be reported by H_i ($i=1, \ldots, M$).

Ignoring the effect of nonsampling errors, the multiplicity estimator is unbiased if ${}_rs_\alpha > 0$ ($\alpha=1, \ldots, N$) and if the counting rule weights, $W_{\alpha i}$ ($\alpha=1, \ldots, N$) ($i=1, \ldots, M$), satisfy the following conditions:

$$\sum_{i=1}^{M} {}_r\delta_{\alpha i}\, W_{\alpha i} = 1 (\alpha=1, \ldots, N).$$

Several kinds of counting rule weights satisfying these conditions have been proposed (Sirken and Royston, 1976). For instance, weights depending only on the number of households eligible to report the I_α are

$$W_{\alpha i} = \frac{1}{{}_rs_\alpha} \quad (\alpha=1, \ldots, N)\ (i=1, \ldots, M), \qquad [2]$$

where

$${}_rs_\alpha = \sum_{i=1}^{M} {}_r\delta_{\alpha i}$$

is the number of H_i ($i=1, \ldots, M$) that are eligible to report I_α ($\alpha=1, \ldots, N$) by rule r. Since the multiplicity estimator requires these weights only for the I_α that are enumerated in the survey, the information needed to determine them is collected from the sample households.

Unless ${}_rs_\alpha > 0$ ($\alpha=1, \ldots, N$), $\hat{N}_r$ is a biased estimate of N. Thus,

$$\text{Bias}\ (\hat{N}_r) = N - T, \qquad [3]$$

where T is the number of the I_α ($\alpha=1, \ldots, N$) for which ${}_rs_\alpha = 0$.

Based on simple random sampling of households with replacement, the sampling variance of $\hat{N}_r$ is

$$V(\hat{N}_r) = \frac{M}{m} V(\lambda_r), \qquad [4]$$

where

$$V(\lambda_r) = \frac{1}{M} \sum_{i=1}^{M} ({}_r\lambda_i - \bar{\lambda}_r)^2 \text{ and } \bar{\lambda}_r = \frac{1}{M} \sum_{i=1}^{M} {}_r\lambda_i = \frac{N}{M}$$

Assuming no more than one person is eligible to be counted at an enumeration unit, that is, assuming

$$\sum_{i=1}^{M} {}_r\delta_{\alpha i} \leq 1\ (i=1, \ldots, M),$$

it follows that

$$V(\lambda_r) = P(K_r - P) = P(1-P) - P(1-K_r), \qquad [5]$$

where

$$K_r = \frac{1}{N} \sum_{\alpha=1}^{N} \frac{1}{{}_rs_\alpha}$$

is the inverse of the harmonic mean of the s_α ($\alpha=1, \ldots, N$). The second term on the right side of [5] represents the reduction in variance due to using network sampling instead of traditional sampling.

Discussion: Network sampling in health surveys

George S. Rothbart, Center for Policy Research

My comments on the Sirken paper come from the perspective of a practically-oriented user of network sampling. In a national study of Vietnam Era males, my colleagues and I are using the network method (I am adopting Sirken's terminology here, but the term "multiplicity sampling" is also employed elsewhere) to draw a sample of veterans of military service during the period of the Vietnam War; we used a modified version of network sampling to obtain persons in particularly rare sample cells in an earlier version of this study. The present sample is a fairly complex one, since it is stratified by race and ethnic status and by service in Vietnam. I should declare the obvious, that I am enthusiastic about network sampling. Thus, I will make numerous approving comments. Nevertheless, as discussed later, there are some problems and limits in its application that should not be ignored.

There is one difference between our use and the examples given in Sirken's paper. He refers only to the use of network sampling to enumerate a population, i.e., to determine how many persons exist with a particular characteristic. We are using network sampling not only to discover the existence of sample persons but also actually to locate and interview them. This option is highly relevant for survey research. It creates some special problems that are worth mentioning, even though this usage is not the subject of the paper under discussion.

Network sampling is an important technical innovation. It not only solves specific sampling problems but also makes possible surveys of rare populations that would otherwise be impossible. Its development comes in a period when innovative methods for the sampling of rare populations are quite important for at least two reasons. First, we are being increasingly asked to conduct studies of persons with special characteristics—such as unemployed workers or people with particular categories of illness—in order to provide data for policy development. As research becomes more pointed, more informed by what has gone before, and more oriented toward practical implications, the general survey is often unsatisfactory to our needs. Second, field-drawn samples are becoming more essential. One *can* study some rare populations using institutional lists, even though the lists are often seriously biased. But even where the lists are good ones, their availability is increasingly problematic. The right of citizens to privacy and the requirement of informed consent can be and often are interpreted so as to deny the use of population lists for social surveys. Network sampling, by offering what is often a quite radical increase in yield, not only makes a study independent of lists but often produces a sample of superior quality.

The yield improvement of network sampling varies in some very important ways by the characteristics of subpopulations. From a practical standpoint, the most important variation may be by ethnic and racial status. We have obtained a higher than average yield among minority groups. The reasons for this are not yet completely evident, but they obviously include the size of families and their complexity as a result of multiple unions. The yield improvement also undoubtedly results from the superior access that network sampling provides to persons without personal telephones. We are presently conducting a study of the variation in telephone accessibility of persons who came into our sample via the network versus the non-network routes, and we will eventually be able to provide some direct information to explain yield variability. Whatever the reasons, the yield of blacks and Hispanics compared with whites has doubled. Obviously, network sampling is a tool of great promise in studies of minorities.

Network sampling improves the yield of any subgroup that is less likely to live in ordinary households with telephones. Persons who are rarely at home, who have no stable address, who are in institutions, or who by their life organiza-

tion are peripheral are individuals who are less likely to appear in a household sample (a "traditional survey" in Sirken's usage) than in a network sample. If the sample is a complex one like ours, calling for strata of blacks and Hispanics and having other characteristics correlated with the life-style characteristics cited above, the overall yield improvement will be commensurately great.

Our reduction of sampling expense was highly significant, the telephone screening cost being close to 50 percent less than for an ordinary household sample, estimated on the basis of the actual within-household yield in the survey. In surveys where the interview is short and conducted by telephone (not the case in ours), network sampling can radically alter the practical limits of a research design.

One problem with network sampling is response bias. Individuals may fail to mention eligible persons who are in their network or they may provide false eligibles—respectively called undercoverage and overcoverage bias. According to our data, however, these biases should be set alongside the biases of a household sample, which also undercovers persons, as we have noted. Since the undercovered persons tend to be of specific types, the bias is a significant one.

The major thrust of my discussion has been a pecuniary one (network sampling saves money); the minor theme is that of sampling quality (network sampling is a tool that can provide superior representation of some segments of the population). I will now delve into the problematic side of the major theme. The problem is a simple one. If money is going to be saved, it is important and even essential to know in advance how much the savings will be and under what conditions they will occur so that the resultant design will be the most efficient one. If estimated savings are small, one may wish to reject network sampling, because the method is complicated and therefore messy from a practical point of view and it may even result in a reduction in efficiency.

In addition to the cost of information to determine counting rule weight, we have found that field costs to contact household respondents also increase with network sampling. Additional questions to elicit potential network responses must be asked. They cannot be properly asked if they are simply "squeezed" into a single screening question. Likewise, a special explanation must be given of the need for questions about the person's network, and any resistance to such questions must be overcome. Typically, only a small amount of time is necessary to add this material, but it must be asked of all the thousands of persons who do not produce eligible respondents. If the characteristic being enumerated is in any way complicated, asking and answering may take considerable time, possibly negating the advantages of the network method.

The major factor in estimating cost is the yield of persons per household. An estimate for this requires knowledge of average network size. That figure, however, is extremely difficult to ascertain in advance of the survey. The average network size may vary by the characteristic being enumerated *in conjunction with* the counting rule that is employed. For example, if it is cancer patients who are being sought, they are likely to be older than average and thus to have fewer parents and siblings and more children than average. If they are veterans of the Vietnam War, they are young and therefore have more parents and fewer grown children than average. Furthermore, if the sample definition is area based, that fact reduces network size in a complicated way: persons in the network but outside the area are ineligible. Estimating yield, in other words, is much more difficult in network sampling than in traditional sampling. In the latter, screening costs can usually be estimated from available survey or census data.

The solution to estimating multiplicity yield is, in my opinion, to conduct a pilot study in the design stage. In our case, less than 400 phone calls were necessary to arrive at an adequate estimate. It was on the basis of that estimate that we decided which counting rule to employ, a decision that could not be left to guesswork.

Finally, I turn to the subject of locating eligible respondents who were enumerated in the screening stage. In the design phase, we had some anxiety regarding the costs and problems of location. Would the household "nominator" be able to supply sufficient information for location? Would a great deal of door-to-door field work be required to track down the eligibles? If a substantial number of eligibles could not be located, a serious question of bias would surface.

In contrast to those anxieties, our experience with location has been positive. First, *most* persons nominated as eligible turned out to be relatively easy to locate by telephone. Second, the very large cost savings of network sampling made it easy to spend considerable time and effort in locating the few persons who presented difficulties. Third, persons who were difficult to locate were often those who would have been lost in a household sample. Thus, they presented their own justification for the location difficulties.

What is essential in the design of a network sample involving location is that the screening

interview be carefully designed to elicit as much information as possible that will aid in the location task. This can include multiple telephone numbers and addresses. Needless to say, callbacks to the nominator (the original informant) are sometimes very useful. It is our experience that locating a potential respondent can almost always be accomplished if one is fortified with a name and reference points to a network whose members are relatively more accessible. This is true even when the eligible respondent does not have a stable address.

Almost all persons have some network links to ordinary households. By providing probabilistic estimation methods for network samples generated from such households, Sirken has made it possible to do field surveys of highly unusual populations without lists and within the comfortable tradition of scientific sampling.

Methodological analyses of Detroit health diaries

Lois M. Verbrugge, Department of Biostatistics and Survey Research Center, University of Michigan

Charlene E. Depner, Survey Research Center, University of Michigan

Introduction

For several decades, the main source of data about individual health and health behavior has been face-to-face, retrospective interviews. Concerned about rising field costs, health researchers are now considering and using other field instruments such as self-administered questionnaires and telephone interviews. Self-administered forms can be retrospective or prospective. If prospective, they are called "health diaries."

Health diaries have been used for three purposes: (1) in methodological studies to compare reporting levels for retrospective and prospective instruments, (2) as memory aids to improve recall of health events in a retrospective interview, and (3) as a primary data source. In a recent review, Verbrugge (1978, 1980) describes health diary studies conducted between 1938 and 1978. Of the 19 studies, 6 were for methodological purposes, 6 for memory aids, and 7 for primary data.

Compared with retrospective instruments, health diaries have two principal advantages for content: (1) higher reporting levels for numerous health indicators and (2) information about "minor" health events as well as "major" ones.

In health diary studies, morbidity and health actions are reported on the day that they occur. This reduces two common problems of retrospective data: memory lapse (forgetting an event entirely) and telescoping (remembering an event but forgetting its correct date). As a result, diaries produce higher counts and rates than retrospective interviews for most health indicators (Allen et al., 1954; Kosa, Alpert, and Haggerty, 1967; Sudman and Lannom, 1979; Sudman, Wilson, and Ferber, 1976; U.S. National Center for Health Statistics, 1972; U.S. Public Health Service, 1962; Wilcox, 1963). Diaries are especially able to elicit reports of acute and chronic episodes that do not cause disability or require medical care, recent acute conditions and diffuse symptoms, and disability days taken for acute conditions. Compared with interviews, diaries produce similar counts of chronic conditions. They are not clearly better or worse than interviews for counts of health actions.

Because recall error is minimal in health diaries, information about "minor" symptoms and health actions can be collected. Diaries can therefore provide data on *all* symptoms experienced during the diary period, *all* disability days taken for symptoms, and *all* curative and preventive health actions. This permits a very comprehensive view of individual health and provides population health indicators not often found in retrospective data.

Experimental studies have evaluated health diaries as a survey procedure. Several types of field instrument have been tested (e.g., self-administered form, telephone interview, face-to-face interview).[1] The following outcomes can be compared: (1) response rates, (2) data quality (item completion), (3) validity,[2] (4) trends in reporting levels,[3] and (5) survey costs. Experimental studies using health diaries are Allen et al. (1954), Sudman and Lannom (1979), Sudman et al. (1976), U.S. NCHS (1972), U.S. PHS (1962), and Wilcox (1963). Results from these studies will be reviewed later in this paper.

Valuable information about procedural aspects of health diaries can also be obtained from nonexperimental studies that use a single strategy for all respondents. Analysis focuses on between-individual rather than between-instrument comparisons. Outcomes can be related to respondent characteristics and also to staff actions toward respondents. Such analyses help researchers understand which population groups respond best and worst to the instrument and which staff activities are most and least effective.

Interested readers are referred to Verbrugge (1978, 1980) for detailed reviews of the following topics: levels of reporting in health diaries

compared with retrospective interviews; recall error; validity of health reports in diaries; value of diary data for individual-level analysis and for studies of health dynamics; respondent cooperation; conditioning effects; data quality; survey costs; and complexity of data collection and processing. Characteristics of previous health diary studies are also described: study place and dates; study purpose; study population; sample size; who recorded information in diary; for whom data were reported; duration of diary; purpose of diary; diary contents; interviews conducted during the study; whether diary was picked up or mailed in; whether respondents were compensated; diary format; and whether daily entries were required.

Analyses for a single-strategy study

This paper discusses methodological analyses for a health diary study conducted in Detroit in late 1978. The Health In Detroit Study used a single strategy for all respondents: it included a six-week health diary. The study provides numerous measures of respondent characteristics, staff actions toward respondents, and respondent performance. There is ample variation in these measures: (1) The population sampled (white adults in the Detroit metropolitan area) varies greatly in sociodemographic and health characteristics. (2) The number and types of staff activities for each respondent (editing, mail contact, telephone contact) varied. (3) The respondent task (daily entries in a health record for six weeks by the sampled person) was demanding, and respondents varied in how well they fulfilled it.

The Detroit data are a fine resource for studying differentials among population groups in response rates and diary-keeping behavior, differentials in the volume and types of staff effort given to population groups, and the relationship of staff inputs to respondent performance. Three aspects of "respondent performance" are response rates, data quality, and trends in reporting levels over the diary period. The Detroit data also provide information on costs for a strategy that varies staff inputs toward respondents.

Analysis of the Detroit diaries can help answer these often-raised questions about health diaries:

1. *Response rates.* Can high response rates be achieved in health diary studies for general population samples? Which population groups are lost most quickly from the study?
2. *Data quality.* Can high-quality data with little missing or unclear information be achieved for diaries that require daily entry?
3. *Conditioning effects (trends in reporting levels).* Do respondents become sensitized to health during the diary period, so that they perceive their symptoms differently than before or change their typical health behavior? Do they tire of keeping the daily health record and provide less careful and complete reports as the diary period lengthens?
4. *Survey costs.* How costly in terms of money expenditures, staff time, and staff activities are procedures designed to achieve high response rates and minimize missing data?

This paper presents a research agenda. Some preliminary results available at the time of the Biennial Conference are reported. First we describe the Health In Detroit Study. Then we discuss analysis plans for response rates, data quality, conditioning effects, and survey costs.[4] For each topic, we review relevant literature, summarize empirical results from other health diary studies, state research questions, and outline specific analysis plans. The research agenda is then compared with recent work by Sudman and his colleagues. The methodological analyses will be performed in 1979–82.

The Health In Detroit Study

Purpose of study. The main purpose of the Health In Detroit Study is a substantive one—to explain sex differentials in health and health behavior. Health indicators from surveys tend to show higher rates of illness, restricted activity, and health services use among females than among males. (Exceptions are higher male rates for some chronic conditions, injuries, and major disabilities from chronic conditions.) The study is designed to test hypotheses about how social and social-psychological factors explain the differentials. The theoretical model is discussed in Verbrugge (1979).

To achieve the substantive aims, the data need the following characteristics: (1) detailed reports about *all symptoms and health actions* in a time period; (2) *self-reports,* not proxy reports; and (3) *a sociomedical perspective* of health rather than a medical one. A health diary promised to fulfill these three criteria:

1. The usual sources of individual health data are health services records and retrospective interviews. Health services records are available only for people who seek medical or dental help; they exclude symptomatic people who do not seek help. Retrospective surveys include a broader sample of the population, but they still limit the symptoms and health

actions queried. Questions focus on "major" health events such as illnesses that cause bed disability or require medical care. The two standard sources of health data therefore capture only the "tip of the iceberg" of individual health. A prospective instrument such as a health diary promised to give data on all symptoms and health actions with minimal recall error.

2. Diaries are a good vehicle for self-reported health data from adults.
3. From a sociomedical perspective, health is "what individuals experience" rather than "what medical professionals diagnose." Health diaries fit the sociomedical perspective well, providing an especially rich view of health as perceived by individuals.

Field instruments. The Health In Detroit Study has a sequence of field instruments for each respondent: (1) a face-to-face Initial Interview, (2) 42 consecutive days of self-administered Daily Health Records, and (3) a Termination Interview conducted by telephone.

The Initial Interview contains items on health attitudes and beliefs, health habits, access to and use of health services, health insurance, lifestyle behaviors, attitudes about social roles, time constraints, general health status, known chronic conditions and limitations from them, recent symptoms and restricted activity from them, stress and anxiety, and sociodemographic information. The interview provides the main predictors for testing hypotheses about sex differentials in health and health actions. It is, however, also designed to stand on its own as a health survey and has numerous questions about current health status, recent illnesses and injuries, and recent use of health services.

The Daily Health Record (DHR) contains items about general health status on a given day, physical symptoms and discomforts, conditions underlying the symptoms, seriousness of symptoms, restricted activities as a result of symptoms, curative health actions (medical and dental care, medications, and treatments), consultation with lay people about symptoms, preventive health actions, mood, and special events that day. The Daily Health Records provide the dependent variables for analysis of sex differentials.

The Termination Interview contains items about general health status, changes in health habits during the diary period, important life events during the period, and reactions to the diary task. The main purpose of the Termination Interview is methodological—to study conditioning effects and respondents' performance of the diary task from their own perspective.

Sample. The study population is noninstitutionalized, civilian white adults (18 years or older) who reside in the Detroit metropolitan area. This area comprises Wayne, Oakland, and Macomb Counties and includes the city of Detroit. A multistage, stratified probability sample of households was designed. In each eligible household, one respondent was selected.

Using information from prior health diary studies and from recent Detroit surveys, response rates were projected for all phases of the study. The sample was chosen to yield about 600 cases with complete data (Initial Interview + 42 Daily Health Records + Termination Interview). The sample was designed and drawn by the Detroit Area Study (Department of Sociology, The University of Michigan) with consultation from the Sampling and Field Sections of the Institute for Social Research. Altogether, 1,184 addresses were selected. These yielded 714 Initial Interviews and 535 complete cases.

Several comments should be made about the sample restriction to white adults: The Health In Detroit Study design differed substantially from previous health diary studies, and there were many uncertainties about response rates. First, the study insisted on self-reports. Consequently it would include a much higher proportion of *male* respondents than previous studies. (Most prior studies asked one person to keep the diary for all household members; that person was usually a female.) Second, the Health In Detroit task was more demanding than in many studies, requiring *daily* entry for six weeks. Third, a *general* population was to be used rather than a special group, such as health plan participants. Fourth, the population was *metropolitan* and thus resided in areas where willingness to participate in surveys may be declining (Hawkins, 1977; Marquis, 1978b). Fifth, no estimates of health diary response rates for *nonwhite* populations were available. Faced with so many uncertainties, we opted to limit the study to the white population. We felt that this would fulfill the substantive goals and also reveal important problems of diary response for males, a daily-entry task, and general population samples in metropolitan areas. The experience would be useful in designing later studies including nonwhites.

Data collection.

Respondent activities. At the end of the Initial Interview, the interviewer took out a Daily Health Record and filled it out with the respondent for "today." Respondents were then asked to complete a Daily Health Record (DHR) each day for the following six weeks. If they

agreed, they were given further instructions and a telephone number to call if they had problems.

Respondents (R) began to keep DHRs the day after the Initial Interview. On Day 2 of the diary period, the field interviewer telephoned R to help with any questions or problems. Diaries were bound into week-long booklets, and respondents were asked to mail in each booklet upon completion. The interviewer left booklets for Weeks 1 and 2 with R. Booklets for Weeks 3–6 were sent later.

At the end of Weeks 1, 2, and 6, a telephone interviewer called to remind R to mail in the completed booklet for that week. Reminder postcards were sent for Weeks 3, 4, and 5.[5] R also received telephone calls if booklets were not returned on schedule, if received booklets had ambiguous or missing responses that needed clarification, or for other problems. If R quit during the diary period, a refusal conversion letter was sent, followed by a telephone call to persuade R to continue in the study. When R finished six weeks of DHRs, a Termination Interview was conducted by telephone. Normally this occurred at the same time as the Week 6 reminder call. When possible, a Termination Interview was also conducted for respondents who completed one or more booklets but then quit the study.

All respondents who completed the Initial Interview received a summary of interview results. Respondents who completed all phases of the study were compensated $10.

Staff activities. Data were collected and coded from September 1978 through January 1979. During this period, the following staffs worked simultaneously: field interviewers, control room staff, telephone interviewers, research staff, clerks, and coders. Interviewers, control room staff, and coders were regular employees of the Field and Coding Sections of the Survey Research Center. Figure 1 diagrams the integrated activities of these staffs.

Research staff activities were directed at achieving high continuation rates and high-quality data from diary keepers. The progress of each respondent was monitored daily by scanning a computer printout that showed the documents received or overdue for each person. Research staff sent letters and arranged telephone calls for people who agreed to keep DHRs but did not start and for people who started keeping them but quit. By telephone and mail, research staff assisted respondents who had special difficulties fulfilling the study task.

A critical activity for quality control was editing the incoming data. Each day research staff edited portions of incoming DHRs to spot incomplete, ambiguous, or missing information. The rest of the DHR was edited by specially trained coders. Editors recorded problems on a special form, called a Problem Sheet. After the editing of a booklet was completed, research staff made a final decision about problems: to code N.A., to attribute responses from previous DHRs or the Initial Interview, or to recontact R. If the decision was to recontact R, research staff wrote specific instructions to telephone interviewers on the Problem Sheet and transmitted these directions and the DHR booklet to the telephone staff.

Data for methodological analyses. The main documents for methodological analyses are the Coversheet, Non-Interview Form, Daily Health Record, Termination Interview, and Respondent Folder. The Daily Health Record and Termination Interview have already been described. Coversheets were available for all sampled addresses. Non-Interview Forms were filled out when no interview occurred at a sampled address. A Respondent Folder was kept for each R who agreed to keep Daily Health Records. The folder contained records of all mail and telephone contacts after the Initial Interview, and it had all Problem Sheets generated for R's booklets.

The research agenda

Methodological analyses of Detroit health diaries will help identify factors that influence respondent performance and population groups that need ample staff attention to keep them in a health diary study. Here we discuss analyses of response rates, data quality, conditioning effects, and survey costs.

Response rates. No standard method for calculating survey response rates is now used (American Statistical Association, 1974; Bailar and Lanphier, 1978). Response rates that are reported must therefore be reviewed and compared with caution. In general, a response rate should be a ratio of the number of completed interviews to the number of eligible respondents (Cannell and Monteiro, 1978). Response rates become artificially inflated if some nonresponse categories are excluded from the denominator of eligibles (Kviz, 1977).[6] Also, response rates for samples of special groups (such as health plan participants or insurance policyholders) are not readily comparable to those for general population samples. Special groups typically

Figure 1. Flow chart for field instruments and staff activities, Health In Detroit Study.

exclude people who might be more difficult to interview.

Most health diary studies have an initial interview followed by a diary period of several weeks or months. Four types of response rates can be evaluated in such a design:

1. *Initial interview response rate.* This is identical to the response rate calculated for typical one-time interview surveys.
2. *Diary agreement rate.* This is based on interviewed respondents who agree to maintain the health diary; they constitute the study panel.
3. *Segment completion rate.* This is measured at regular intervals of the diary period. It indicates retention of diary keepers in the study (or its converse, panel attrition).
4. *Study completion rate.* This refers to panel members who remain in the study at the end of the diary period. It can include only the people who submit diaries for the entire period or also people who participate for the entire period but submit partial data (having skipped some days or weeks). If the study has a final interview, another completion rate can be computed based on panel members who complete the interview as well as the diaries.

Using uniform definitions, response rates for health diary studies have been computed and are shown in Table 1.

1. Interview response rates in diary studies are similar to typical interview surveys. The frequently cited decline in survey response rates does not appear in Table 1.[7]
2. Diary agreement rates in prior health diary studies are very high, usually above 85 percent.
3. The segment completion rates show that attrition during the diary period tends to happen early, being concentrated in the first segment.
4. Total attrition is quite small, and diary studies have shown high study completion rates (88–100 percent). A distinction must be made between diaries that are mailed in by respondents and those picked up by interviewers. Mail-in and pick-up procedures have similar diary agreement rates, but the mail-in procedure has lower segment completion and study completion rates. Compensation improves completion rates for the mail-in procedure (Sudman and Lannom, 1979).

Response rates may vary across population groups. Selectivity in response can lead to biased estimates of parameters (Daniel, 1975; Goudy, 1976, 1977; Mandell, 1974). Information on selectivity helps researchers to design future studies by pointing out "troublesome" groups that need special attention.

There is little information about selectivity at any stage of health diary studies. Muller, Waybur, and Weinerman (1952) found that young people and those with high acute illness rates tend to drop out of the diary panel more than others. However, Sudman and Lannom (1979) found that households with *few* illness episodes in the past year tend to drop out more than households with numerous episodes,[8] and that low-education households tend to quit more than high-education ones.[9]

In summary, the largest loss of eligible respondents for a health diary study is at the very beginning, in refusals to have a face-to-face interview. This is a problem that all interview surveys face. Among diary starters, people who quit the study usually do so early in the diary period. Surprisingly, diary agreement and study completion rates do not seem related to the length of time respondents are asked to keep the diary. Long diary periods are needed for stable estimates of some health indicators, and they provide rich data for individual-level analysis. Since long periods do not appear to scare off respondents, Sudman and his colleagues now recommend that respondents be asked to keep diaries for three months or longer.

Much remains unknown about response rates in health diary studies: What reasons do people give for refusing to keep a health diary? Which population groups are least likely to agree to keep diaries and to maintain them throughout the diary period? Do reluctant diary keepers come from the same population groups as those who refuse interview surveys? What reasons do people give for dropping out of the study? What fraction of them can be persuaded to return? Are returnees selected for certain sociodemographic or health characteristics? How successful are mail and telephone attempts to bring dropouts back into the study? What do respondents say about the diary task at the end of the study—did other people help them fill out the record? Did they tire of keeping the diary? What did they do if they skipped a day?

The Detroit study can provide some answers. The following analyses are planned:

1. *Selectivity of interviewed respondents.* Examine reasons for noninterview among sampled addresses. Compare characteristics of interviewed and noninterviewed households. Compare interviewed households with census data for the same population.
2. *Selectivity of respondents who agree to keep Daily Health Records.* Describe reasons for refusal.

Table 1
Response rates in health diary studies

Category	Muller et al. (1952) Berkeley 1949	Mooney (U.S. PHS, 1962) San Jose 1952	Roghmann and Haggerty (1972) Rochester, N.Y. 1969	Wilcox (1963) Tecumseh, Miss. 1962	Laurent et al. (U.S. NCHS, 1972) Detroit 1968	Sudman et al. (1976) Marshfield Wis. 1973–74	Sudman et al. (1976) Chicago 1973–74	Sudman and Lannom (1979) Illinois 1976
	Source, place, and date							
Eligible	815	1,632	628	50	121	321	337	1,077[h]
Interviewed	564[b, c]	1,474	543[b]	—	—	288	251	873
Interview response rate[a]	69%	90%	86%	—	—	90%	74%	81%
Agreed to keep diary	564	1,430	534	48	108	282	240	765[i]
Diary agreement rate[a]	69%[d]	97%	86%[d]	96%	89%[d]	98%	96%	88% (90,86)[j]
Segment length	Month	Month	Month	Month	Week	Month	Month	Month
Number of segments	5	1	1	1	1	3	3	3
Finished segment 1	NA[e]	1,400	512	48	101	277	220	639
Segment 1 completion rate[a]		98%	94%	100%	94%	98%	92%	84% (97,71)
Finished segment 2	NA	—[g]	—	—	—	269	216	604
Segment 2 completion rate		—	—	—	—	95%	90%	79% (95,64)
Finished segment 3	NA	—	—	—	—	267	211	576
Segment 3 completion rate		—	—	—	—	95%	88%	75% (94,58)
Completed study	504	1,400	512	48	101	267	211	576
Study completion rate[a]	89%[f]	98%	94%	100%	94%	95%	88%	75% (94,58)

[a] Interview response rate = Number interviewed/Number eligible for interview. Diary agreement rate = Number who agreed to keep diary/Number interviewed. Segment completion rate = Number returning diary for segment n/Number who agreed to keep diary. Study completion rate = Number who kept diary for entire diary period/Number who agreed to keep diary.

[b] All people who agreed to do the interview also agreed to keep the diary.

[c] Most loss of eligibles was due to change in employment (which meant leaving the study population) or other unavailability. Only 63 (8% of 815) persons refused to participate in the study.

[d] Calculated as Number who agreed to keep diary/Number eligible for study.

[e] NA = Not applicable. Study had no initial interview, or no segments.

[f] According to Muller et al. (1952), all attrition was due to change in employment (which meant leaving the study population). No one dropped out because of unwillingness to continue. The study completion rate includes people who missed middle segment(s) of the diary period but stayed in the study.

[g] Mooney also had some respondents keep a diary for approximately 4 months with monthly return. Response rates cannot be computed from published data. Attrition was 2–3% per month (U.S. PHS, 1962:47).

[h] The 873 respondents were chosen from 5,214 people who completed an initial screener. The overall noninterview rate for the screener was 19%. This is used here to estimate the number of eligibles for the diary strategy.

[i] This is the number of screened respondents who had an initial interview before the Month 1 diary. It is not clear if all of them agreed to keep the diary.

[j] For half of the sample, diaries were picked up by interviewers at the end of each month. For the other half, respondents mailed in diaries monthly. Response rates for pick-up and mail-in procedures are shown in parentheses under the overall rate.

Compare agreers and refusers on sociodemographic and health characteristics.

3. *Selectivity of diary keepers who complete the study.* Describe reasons people gave for quitting the study (dropouts). Compare dropouts to people who keep DHRs for the entire diary period. Compare dropouts who return to the study with those who do not return. Also compare people who do not mail DHRs on schedule (delinquents) with those who mail DHRs on time.
4. *Staff efforts to retain diary keepers.* Compare the volume and types of staff contacts for three groups: dropouts who return to the study, dropouts who do not return, and continuous diary keepers. (Measures of staff contact must adjust for the variable lengths of time respondents are in the study.)
5. *Prediction of study completion.* Estimate a model that predicts study completion, using respondent characteristics and staff contacts as predictors.
6. *The diary task.* Report diary keepers' reactions to the diary task as reported in the Termination Interview. Examine sociodemographic and health correlates of these items.

Some preliminary results on response rates are available for the Detroit study. Table 2 shows response rates by sex for all stages of the study. (1) The interview response rate is rather low. Eligible addresses had an unusually large number of people incapable of participating because of mental, severe physical, and language problems. Reasons for refusal by eligible households will be examined in detail. (2) The diary agreement rate is comparable to that of previous health diary studies. This general similarity masks an important result: Most prior studies have restrictions on the study population, which favor high response. Only the San Jose, Illinois State, and 1978 Detroit studies have general population samples with few restrictions (cf. Table 1). All achieve high diary agreement rates. This indicates that health diaries can be successfully used in general populations. (3) The Health In Detroit Study has the finest record to date of segment completion and study completion for a mail-in procedure. This is especially notable because the respondent task was sizable (daily entry and self-response for six weeks). For a mail-in procedure that did not require daily entry or self-response, Sudman and Lannom (1979) report a 71 percent segment completion rate at the end of one month, and 64 percent after two months. The Health In Detroit Study has an 84 percent segment completion rate at the end of one month, and 81 percent at the end of six weeks.[10]

Data quality (missing or unclear responses). We use the term "data quality" to mean the frequency of missing or unclear information in an instrument. Indicators of data quality are not standardized in survey research. Some useful indicators are the frequency of skipped items, ambiguous responses, inappropriate responses, nonspecific responses, and nonsubstantive "don't know" responses. If survey forms are edited before they are coded, the frequency of editor entries can also be used as an indicator of data quality.

Researchers seldom evaluate a survey for its overall levels of missing and unclear information.[11] This is unfortunate, since information on the frequency and correlates of data quality can help researchers design questionnaires (both interviewer-administered and self-administered).

Most health diary studies do not report on data quality at all; the information from the few

Table 2
Response rates for the Health In Detroit Study

Category	Total	Male	Female
Eligible	1,041	480[b]	564[b]
Interviewed	714	302	412
Interview response rate[a]	69%	63%	73%
Agreed to keep diary	652	275	377
Diary agreement rate[a]	91%	91%	92%
Segment length	Week		
Number of segments	6		
Finished segment 1	588	241	347
Segment 1 completion rate[a]	90%	88%	92%
Finished segment 2	568	234	334
Segment 2 completion rate	87%	85%	89%
Finished segment 3	549	225	324
Segment 3 completion rate	84%	82%	86%
Finished segment 4	545	225	320
Segment 4 completion rate	84%	82%	85%
Finished segment 5	534	222	312
Segment 5 completion rate	82%	81%	83%
Finished segment 6	528	220	308
Segment 6 completion rate	81%	80%	82%
Completed study	528	220	308
Study completion rate[a]	81%	80%	82%

[a] See Table 1, Note a for formulas.

[b] Sex of potential respondent was not known for 157 addresses. Assuming a male/female ratio of 45/55, estimates of eligibles by sex are made and interview response rates are computed based on them.

[c] For men on the sample, diaries picked up half, respondents mailed in diaries monthly. Response rates for pick-up and mail-in procedures are shown in parentheses under the overall rate.

[d] NA = Not applicable. Study had no initial interview, or no segments.

[e] According to Muller et al. (1952), all attrition was due to change in employment (which meant leaving the study population). No one dropped out because of unwillingness to continue. The study completion rate includes people who missed middle segment(s) of the diary period but stayed in the study.

that do provide such results is very limited (Mechanic and Newton, 1965; Muller et al., 1952; Roghmann and Haggerty, 1972; Sudman and Lannom, 1979; U.S. NCHS, 1972; U.S. PHS, 1962; Wilcox, 1963). Their results are compiled and reported in Verbrugge (1978, 1980). In general, studies that give respondents strong incentives to keep diaries tend to have higher quality data than studies that offer weak incentives.

There are many possible reasons for poor-quality data. Respondents may be careless about making entries, lack motivation about the study, or feel embarrassed about certain items.[12] The diary may have difficult wording for some items, unclear skip patterns, or confusing chart format. Researchers who want to use health diaries in a general population need to know (1) if some population groups have more difficulty than others in filling out the form correctly and (2) what diary format and question wording are best for all population groups (or for specific groups). Even if researchers find differentials in data quality for sociodemographic groups or for diary items, the underlying reason for those differentials may not be obvious.

To date, there is limited evidence on these issues. Muller et al. (1952:Table 4) found that blue-collar workers tend to keep incomplete records (to skip whole segments of the diary period) more than do white-collar workers. Sudman and Lannom (1979) report that interviewers had to obtain additional information from low-education households more often than from high-education ones. (Whether this was to clarify responses or to fill in missing data is not noted.) No differences in data quality appear for households with high illness incidence in the past year versus those with low incidence.

Sudman and his colleagues (1976) tested two diary formats. A ledger had separate pages for different topics (e.g., work absence, doctor visit). A journal asked for all details about a health event on the same page. (It also had a page for details about health expenditures.) Neither form required daily entry of information. When they compared reporting levels for the ledger and the journal, Sudman et al. (1976) found no important differences.

One other result has been reported: Sudman and Lannom (1979) compared the frequency of editor entries for various health indicators in the diaries. There was little difference, with one exception. Hospital visits often needed additional questioning because the beginning and ending dates were missing. The reason for this is not obvious.

In summary, little is known about data quality in health diary studies that can help researchers in designing a diary instrument. These questions remain: Which population groups produce data with fewest skipped days, skipped items, unclear answers, and editor-altered items? Which groups need the most staff inputs aimed at clearing up errors in records and at improving response quality for the rest of the diary period? How successful are staff interventions to improve data quality? What items produce the least missing and unclear information, and is their success related to features of the diary format?

The following analyses of data quality are planned for the Health In Detroit Study:

1. *Description of data quality and staff actions to handle data-quality problems.* Indicators of data quality come from the Daily Health Records and Problem Sheets. Indicators of staff actions are from Problem Sheets and from Respondent Folder records of mail and telephone contacts with R.
2. *Correlates of data quality and staff actions.* Examine the relationship of data quality to respondents' sociodemographic and health characteristics. Also examine the relationship of staff actions to respondent characteristics. (Which respondents need most attention for poor-quality DHRs?)
3. *Effects of staff interventions on data quality.* For respondents who were contacted about problems in edited DHRs, examine levels of data quality before and after those calls. Using the results, estimate levels of data quality assuming no staff intervention and then re-examine differentials among population groups.
4. *Diary format.* Assess levels of missing and unclear entries for each DHR item. Characterize items by their placement, difficulty, boldness of print, and other features. See if these are related to data quality and make recommendations about diary format.

Conditioning effects (trends in reporting levels). Panel studies provide repeated measures of respondents' attitudes and behaviors. Participating in a panel may actually cause those attitudes and behaviors to change, or it may influence the way that people report their stable attitudes and behaviors. Such changes are called conditioning effects. They can have a sizable impact on estimates of rates and other parameters, causing trends to appear when the estimates are plotted over time (Bailar, 1975).

Two important conditioning effects are sensitization and fatigue. Respondents may become more sensitive to the attitudes and behaviors queried, so that their feelings and actions actually change. They may tire of participating in

the study and begin to provide less complete or thoughtful responses. Generally, fatigue is thought to reduce reporting levels over time. Sensitization can boost or diminish reporting levels over time depending on the item's salience, social desirability, etc.

Trends in panel data do not necessarily reflect conditioning effects. Other factors that can produce trends are seasonal variations, holidays, major political or community events, panel losses (selective attrition), sampling variation, changes in field procedures or interviewers, recall error that affects some periods more than others, and changes in coding schemes or data processing. These factors should be considered before trends are attributed to conditioning effects.

Sensitization and fatigue are quite likely to occur in health diary studies. First, while keeping the diary, respondents may become more aware of their health and more interested in it. As a result, they may perceive symptoms more readily than before, and they may change typical health behaviors. Second, as the diary period lengthens, respondents may tire of keeping the records. Consciously or not, they may become less thorough in reporting health events. It is generally thought that sensitization will boost reports of symptoms, disability, and health actions. The boost may be temporary (e.g., at the beginning of the diary period) or persistent. Fatigue acts to reduce reports over time.

What is the evidence for conditioning effects in health diaries? Several researchers have found that reporting levels for health problems and health actions tend to drop over time (Kosa et al., 1967; Sudman and Lannom, 1979; Sudman et al., 1976; U.S. PHS, 1962).[13] Sudman and Lannom's data are most comprehensive: Drops in rates appear for illness, disability days, bed days, visits to health professionals, and purchase of nonprescription medical supplies. No drops appear for hospital visits, purchase of prescription medical supplies, and payments to health care providers. In general, the drops in health diaries appear regardless of the season in which the studies were conducted. Conditioning effects appear to be the most plausible reason for the trends. Sometimes researchers attribute drops to sensitization at the beginning of the diary period, sometimes to fatigue as the diary period lengthens.

Let us summarize this approach: Rates are computed for segments of the diary period and then plotted. If trends are found, they are attributed either to sensitization or to fatigue. The approach uses aggregated data and strong inference. Actually, it is impossible to know if the interpretations are correct. Usually both sensitization and fatigue are plausible explanations for a particular trend.

A more formal approach would be to state plausible models of sensitization and fatigue well in advance of data analysis, then to test those models on the data. The perspective and statistical techniques of cohort analysis could help identify seasonal and conditioning effects. (In the language of cohort analysis, the first are "period" effects; the second are "age" effects.) This formal approach would produce more cautious but also more convincing conclusions than previous research.

Another very different approach that is possible relies on individual-level analysis and specific indicators of sensitization and fatigue. Respondents can be asked directly about changes in their health perceptions and behavior and about their reactions to the diary task. Fatigue indicators can also be obtained from office records about editing problems, delinquent return of diaries, etc. All of these variables can be associated with patterns in the diary reports over time.

In summary, although trends in health reporting have been found in diaries, little is known about the reasons for these trends. Are they due to initial sensitization to health, fatigue in keeping health records, or some combination of the two? What are plausible models of sensitization and fatigue, and how well do observed trends fit them? How can the concepts of sensitization and fatigue be operationalized to obtain scores for each respondent? Precisely what aspects of health are sensitized (e.g., increased awareness of symptoms, increased propensity to take medicines)? Which population groups are most likely to be sensitized or fatigued? How do the conditioning effects influence individual reports over time?

Extensive analysis of conditioning effects is planned for the Detroit data using both aggregate-level and individual-level approaches:

1. *Aggregate analysis.* Develop models of sensitization and fatigue. Test these models on rates for health indicators in the DHRs.
2. *Individual-level analysis.* Develop indicators of sensitization and fatigue from the Termination Interview, Respondent Folder, and other office records.[14] Examine sociodemographic correlates of those indicators. Dichotomize respondents as "sensitized" or "nonsensitized." Compute rates for diary segments separately for these two groups, and compare their levels and patterns over time. Do the same analysis for fatigued and nonfatigued respondents.

Survey costs. There is no standard scheme for computing and reporting costs in survey research. When costs are reported, the components included vary greatly among studies. This is partly due to different accounting conventions used by research agencies, but it is also partly due to the absence of a standard scheme. Such a scheme should encompass costs of study design and sampling, data collection, and initial data processing. In lieu of that, researchers must be careful to state clearly which components are included and excluded in their calculations.

In any case, knowing a survey's money costs is of limited use for other investigators even if they want to use the same design. Inflation quickly makes the figures outdated; and many factors contribute to a particular study's costs that may go unreported (e.g., accessibility of respondents, strictness of quality control procedures).

There are other ways to measure survey costs that may be more useful than money costs. Information on the distribution of costs among components and the ranking of costs among data collection strategies are probably more durable over time. In addition, other kinds of "costs" can be examined, such as staff time and staff activities directed toward the respondents. All of these help researchers estimate the inputs necessary to achieve their study objectives.

Brief mentions of money costs for health diary studies appear in Allen et al. (1954), Peart (1952), and Roghmann and Haggerty (1972). Detailed information on costs appears in Sudman et al. (1976) and Sudman and Lannom (1979). Sudman and his colleagues compare money costs for three strategies of data collection. Costs include interviewer time and travel, mail and telephone charges, and respondent compensation. Average costs are lowest for a strategy with repeated telephone interviews. A strategy with mail-in diaries is slightly more expensive. Repeated face-to-face interviews are much more expensive. If diaries are picked up or if diary keepers receive compensation, the costs of a diary procedure climb quickly. When both features are used, diary costs exceed those of all-face-to-face interviews.

In general, procedures to improve response rates and data quality in a diary study will increase all types of survey costs whether counted in terms of money, time, or activities. Researchers must weigh the value of the scientific product against those expenses. To do this intelligently, they need more information from studies that preceded their own. For example, what is the relative distribution of money (and time) costs across components of a health diary study? If a study has numerous procedures for retaining respondents and improving their reporting, which procedures must be used most? Which population groups require the most attention during the study? How skewed is the distribution of staff activities—do a few respondents get inordinate attention?

Using the Detroit data, we can analyze survey costs from the perspectives of money expenditures, staff time, and staff activities. Money and time data are aggregated and thus are available for the entire project but not for each respondent. Data on staff activities are available for each respondent.

Cost analyses will inform researchers about the work load involved in a health diary study and the population groups that generate the highest costs. The cost-effectiveness of staff activities can also be studied by asking how they help retain respondents and improve their reporting. (These analyses were outlined in the sections on Response Rates and Data Quality.)

The following analyses are planned:

1. *Money expenditures.* Report total money costs for study design and sampling, data collection, and data processing. Report costs per respondent. Show the distribution of costs among components.
2. *Time costs.* Report the distribution of hours worked by various staffs. Report time costs per respondent for each staff group.
3. *Activity costs.* Using Respondent Folder materials, describe staff activities for respondents, especially the number and purpose of mail and telephone contacts. Examine sociodemographic and health correlates of staff activities. (Who required the most staff attention?)

Conclusion

There is increasing interest in health diaries as a procedure for collecting health data. The advantages of diaries for *content* are well known: They can provide data on minor as well as major health events, and recall error is minimized for virtually all items. Many investigators, however, are worried about *procedural* aspects of diaries: response rates, quality of diary entries, conditioning effects, and survey costs. To date, research on these topics suggests that high response rates and high-quality data can be obtained if staff inputs are ample and well organized. Obviously this means that high-quality diary studies will cost plenty of money, time, and staff effort. There is limited evidence about conditioning effects in health diaries. Some trends in reporting have been shown, but

whether they are due to sensitization or fatigue (or both) is unknown.

The most extensive and careful analyses of procedural issues to date are by Sudman et al. (1976) and Sudman and Lannom (1979). The Health In Detroit Study offers a fine opportunity to study the issues further. The main goals of methodological analyses for the Detroit study are to learn about (1) the differentials in respondent performance, (2) the staff inputs given to different population groups, (3) the relationship of staff interventions to respondent performance, and (4) the costs of obtaining high continuation among diary keepers. In the appendix, we compare the study design and methodological analyses of the Health In Detroit Study with the studies by Sudman and his colleagues.

Results from the Health In Detroit Study will be reported in ways to help other researchers design health diary studies. For example, the Detroit study will identify which population groups are most difficult to recruit and keep in a study, which mail and telephone contacts improve the quality of diary reports, and which staff activities are least successful in maintaining respondents or converting dropouts.

Appendix: Design and methodological analyses for four studies (Chicago, Marshfield, Illinois, Detroit)

In 1973–1974 and 1976, Sudman and his colleagues conducted studies using health diaries in Chicago, Marshfield (Wisconsin), and Illinois. How do the design and analysis plans of the Health In Detroit Study differ from them?

The Chicago, Marshfield, and Illinois studies used experimental designs to compare several strategies of data collection. Three field instruments were tested: a health diary, repeated face-to-face interviews, and repeated telephone interviews. The principal purpose of the studies was methodological—to compare response rates, reporting levels, costs, data quality, and validity across strategies.

The Health In Detroit Study has a nonexperimental design; one strategy is used for all respondents. The main field instrument for health data is a health diary. The study has both substantive and methodological purposes: The health diary is a primary data source for testing hypotheses about health and health behavior. The study also provides for methodological analyses of response rates, reporting levels, survey costs, and data quality.

Table A1 shows the treatment, control, and dependent variables for the four studies. For the Chicago, Marshfield, and Illinois studies, diary segments were one month long. The time interval between interviews was also one month. For the Detroit study, diary segments were one week.

Analysis of the Chicago and Marshfield studies is presented in Sudman, Wilson, and Ferber (1976). Treatment effects on the dependent variables are discussed. Interaction effects among treatments are also examined. Analysis of the Illinois study is presented in Sudman and Lannom (1979). Treatment effects are again examined, as are control effects and treatment-control interactions. Based on results from these studies, researchers can choose among several strategies for health surveys knowing their advantages and disadvantages with respect to response rates, reporting levels, and money costs. There are more limited results for data quality and validity.

For the Detroit study, analyses consider "treatment" and "control" effects on the dependent variables. (Here "treatment" refers to mail and telephone contacts, which vary among respondents; "control" refers to sociodemographic and health characteristics of respondents.) Relationships between "treatment" and "control" variables will also be examined, as will their net effects on dependent variables.

The Chicago, Marshfield, and Illinois studies provide fine information about health diaries. The Detroit study increments it in several ways:

1. *Respondent selectivity.* We can do extensive analysis of respondent selectivity at all stages of the study. The Coversheet and Initial Interview provide ample information about sampled households and respondents. The other three studies obtained limited information about sample units.

2. *Dropouts who return to the study.* The Detroit study can examine diary keepers who dropped out of the study but were persuaded to return. Reporting levels, data quality, and staff actions for these returnees can be compared with respondents who continued without interruption in the study. The Chicago, Marshfield, and Illinois studies did not discuss dropouts, procedures to recruit them again, and subsequent performance of returnees.

3. *Mail and telephone contacts.* For the Detroit study, there is great variation in staff actions applied to respondents. We can evaluate their worth in improving continuation and data quality. In particular, the importance of telephone contacts for encouraging respondents to mail in diaries can be examined. The other

Table A1
Designs of the Chicago, Marshfield, Illinois, and Detroit studies

Study	*Treatment variable[a] (method)*	*Control variable (sample)*	*Dependent variable*
1973–74 Chicago and Marshfield	Data collection strategy: F+DP+DP+DP F+F+F+F Diary format: [b,c] Ledger Journal Compensation[d] None $10 (Chicago) Health report (Marsh.)	(None)	Response rates (cooperation) Reporting levels and trends in reporting Validity (external record check) Costs
1976 Illinois	Data collection strategy:[e] F+DP+DP+DP F+DM+DM+DM F+F+F+F T+T+T+T Compensation:[d] None $15	Education[f] Illness experience in past year[f]	Response rates Reporting levels and trends in reporting Data quality[b] Costs
1978 Detroit	F+DM+DM+DM+DM+DM+DM+T[g] plus nonscheduled mail and telephone contacts during diary period[h]	(No experimental controls)	Response rates Reporting levels and trends in reporting Data quality Costs

[a] F=face-to-face interview, T=telephone interview, D=diary, P=pick-up, M=mail-in.

[b] For diary strategies only.

[c] The ledger has separate pages for different topics (e.g., work absence, doctor visit). The journal asks for all details about a health event on the same page. (It also has a page for details about health expenditures.)

[d] For all strategies.

[e] All of these strategies occurred after an initial screening interview by telephone. The screener was a short interview on household composition, illness experience in the past year for all household members, and education of female head or spouse of male head.

[f] Education=education of female head or spouse of male head. Illness experience in the past year is an index based on reports of illness episodes, restricted activity, and hospital stays for all household members.

[g] Each respondent received four scheduled telephone calls: one by the field interviewer two days after the Initial Interview, and three reminder calls at the end of Diary Weeks 1, 2, and 6.

[h] Written guidelines were used for deciding whether or not to make these contacts.

three studies had less variation in contacts with respondents.

4. *Data quality.* Detailed records of editing problems were kept for the Detroit study, permitting close analysis of data quality—its level and also how staff interventions affected it. Data quality is a minor topic in the other three studies.
5. *Conditioning effects.* The Detroit data have specific measures of sensitization and fatigue for each individual. These measures can be related to patterns of reporting over the diary period. This individual-level analysis of conditioning effects is an innovation in health diary studies. Sudman and his colleagues have studied trends in rates and suggested their causes. Aggregate analysis will also be done for the Detroit study, using a formal modeling approach.
6. *Survey costs.* For the Detroit study, survey costs can be reported in terms of money expenditures, staff time, and staff activities. Costs for a wide range of data collection and data processing components can be included. For the Chicago, Marshfield, and Illinois studies, cost analysis focuses on money expenditures for data collection.
7. *Health indicators.* The Detroit data have some health and health behavior variables that are not available in the other three studies. The

Table A2
Health indicators from the Chicago, Marshfield, Illinois, and Detroit studies

Health indicator	*Chicago and Marshfield*	*Illinois*	*Detroit*
Illness:[a]			
*Illness days	x	x	x
*Symptom days	—	—	x
*Type of illness[b]	—	x	x
Treatment of illness[c]	—	x	x
Seriousness of symptoms	—	—	x
General health status (daily)	—	—	x
Disability:			
*Disability days[d]	x	x	x
*Type of illness on disability day	—	x	x
*Bed days[e]	—	x	
Treatment of illness on disability day[c]	—	x	x
Visits to health professionals:			
*Doctor contacts[f]	x	—	x
*Total visits to health professionals	—	x	x
Type of health professional visited	—	x	—
Reason for visit (type of visit)[g]	—	x	x
Reason for visit (name of condition)	—	—	x
*Made appointment for visit[h]	—	—	x
*Telephoned doctor or dentist for advice[h]	—	—	x
Hospital care:			
*Outpatient visits	x	x	x[i]
*Overnight stays	x	x	x
Reason for overnight stay (name of condition)	—	—	x
Lay consultation:			
*Discussion of symptom with relatives or friends	—	—	x
Type of person R talked with	—	—	x
Use of medicine or treatments:			
*Medicine/treatment days	—	—	x
Name of medicine/treatment[j]	—	x	x
*General type of medicine/treatment (prescription, over-the-counter)	—	x	x
*Reason for taking it (preventive, curative, other)	—	—	x
Reason for taking it (name of condition)	—	—	x
Medical supplies and payments:			
*Number of items purchased	x	x	—
Type of item purchased	—	x	—
Whether covered by insurance or not	—	x	—
Price	—	x	—
Payments to health care providers:			
*Number of payments	x	x	—
Whether covered by insurance or not	—	x	—
Amount paid	—	x	—

[a] Item wordings for the Illinois study are in Sudman and Lannom (1979). The Chicago and Marshfield items are very similar. Item wordings for the Detroit study are in the Daily Health Record.

[b] To code illnesses and injuries, the Detroit study used the Reason For Visit Classification developed for the National Ambulatory Medical Care Survey, with some modifications to fit the Detroit data. The Illinois study used the Short Index of Diseases, Injuries, and Impairments developed for the Health Interview Survey, with some modifications.

[c] Categories used for Illinois: prescription, other medicine, other treatment, none. More detail is available and coded for Detroit.

[d] All studies ask about missing work/school and cutting down usual activities. The Detroit study also asks about cutting down other planned activites such as sports or club meetings.

[e] Bed days are one type of disability day.

[f] Categories for Chicago and Marshfield: telephone, acute care visit, routine visit.

[g] Categories for Illinois: acute illness, checkup, chronic illness, consultation, therapy, tooth cleaning. More detail is available and coded for Detroit.

[h] This can be added into counts of visits for an overall count of contacts with health professionals.

[i] The Detroit data include outpatient hospital visits with doctor/dentist visits; i.e., they are not separable.

[j] The Illinois study uses 14 categories; Detroit uses 95.

NOTE: x means the indicator is used in the study. — means the indicator is not used (and usually that the necessary data were not collected). * indicates that counts of events are the dependent variable. For other items, distributions or averages are computed instead.

variables provide more detail about the illness experience and its social consequences. The Detroit data have fewer indicators for purchases of medical supplies and payments to health care providers than do the other studies. Table A2 compares the indicators used in analyses in the four studies.

8. *Self-response.* An especially important feature of the Detroit data is self-response. Females have long predominated as reporters in health diary studies. For the Chicago, Marshfield, and Illinois studies, data are proxy-reported for the majority of males. The Detroit data provide an opportunity to assess response rates, reporting levels, and data quality for males in a health diary study and to compare their performance with that of females.

Footnotes

[1] "Type of instrument" is a treatment variable in experimental studies. The studies often test other treatments too (e.g., compensation). A specific combination of treatments is called a strategy. Each respondent is assigned to a strategy.

[2] Validity is operationalized in various ways. Some studies compare interview and diary reports with medical records. Some compare interview and diary reports with each other and assume that the instrument that yields higher rates is more valid.

[3] Trends may reflect "conditioning effects" in a panel. These are reactions to being in a panel study that influence people's answers over time.

[4] Some analysis of validity is planned that compares diary reports with reports in an initial interview. Because the purpose is to evaluate interview items, not diary items, the plans are not discussed here.

[5] The original study design planned for a telephone call to R at the end of each week. For administrative reasons, this was reduced to three calls shortly after data collection began. Thus, some respondents did receive 4–6 reminder calls, but most received 3.

[6] For example, Sudman and Lannom (1979) report response rates for a study in Illinois that used health diaries. Households were initially contacted by telephone, and a screener interview was conducted during the call. Overall, 19 percent of eligible housing units did not yield a screener interview. In the response rates, denominators are based on all households where a screener interview occurred, not on all households with working telephones.

[7] Some studies show a decline, but not others (Marquis, 1978b). As a general principle, high interview response rates can be achieved if numerous attempts are made to contact eligible respondents (Kish, 1965; Sprately, 1973; Sudman, 1967). Urban areas may require the most effort, since they have concentrations of people difficult to contact and interview (e.g., older people, foreign-language speakers, restricted access dwellings) (Market Research Society, 1976).

[8] The results are probably not contradictory. In the Muller et al. study, respondents were university employees. Most dropouts were people who changed jobs. These tended to be young people—at ages with relatively high acute condition rates. In the Sudman and Lannom study, most dropouts were respondents unwilling to continue.

[9] Low education is defined as "less than 12 years for the female head or spouse of male head."

[10] In San Jose, respondents were asked to mail them, but interviewers picked up any that were not received by mail (U.S. PHS, 1962).

[11] They do show concern for the quality of specific variables to be used in substantive analyses. Researchers often make adjustments for missing or unclear entries for these variables. Their procedures are based on extensive literature about how item nonresponse can produce biased estimates and about strategies to correct such bias. (See Afifi and Elashoff, 1966; Elashoff and Elashoff, 1970; Hertel, 1976; Hutcheson and Prather, 1977; Miller, 1970.)

[12] Sociodemographic groups may vary in how well they recall health events; but this in unlikely to affect health diary reports, since people record information on the day a health event occurs.

[13] Panel studies of consumer expenditures also find drops over time (Neter and Waksberg, 1964a; Turner, 1961; U.S. Bureau of the Census, 1968). Bailar (1975) discusses sensitization effects in the Current Population Survey.

[14] The Termination Interview is the principal source. Respondents were asked direct questions about sensitization and fatigue (e.g., if they noticed health problems more during the diary period than before, if they handled their health problems differently, if they tired of filling out the DHRs). The Termination Interview also repeats some health items from the Initial Interview, which were expected to remain stable over the diary period. Changes may be taken as evidence of sensitization.

Evaluation of health diary data in the Health Insurance Study

Kent H. Marquis, The Rand Corporation

Introduction

I would like to discuss some of my own recent research with health diaries. This research uses the diary not only as a primary source of health information but also as a drive for a more extensive collection of detailed data from medical care providers. In the first part of this discussion, my remarks will illustrate that the use of health services can be severely underreported on diaries under some conditions. In the second part, my remarks are more optimistic: They indicate that the dual-purpose diary methodology can achieve its objectives under other circumstances.

The data that we will look at are from the Health Insurance Study, a large-scale, social experiment that will estimate the demand for health services as a function of price and the effect of different kinds of health insurance on health status. In this study, the participant reports details about his health and his use of health services in a diary; he also assumes responsibility for providing the study with technical medical information from the health service provider.

Participant sample

The analyses reported here are for Health Insurance Study enrollees residing in the Dayton, Ohio, site during 1975. They include 1,140 persons enrolled in experimental plans and 631 persons enrolled in the control group. Eligible participants are a systematic sample of civilian site residents under 65 years of age. Dayton, Ohio, was the first site to be studied, and 1975 was the first year of the experiment.

The utilization reporting system

The documents and methods of obtaining information about the use of health services in the first year of the Health Insurance Study are summarized in Figure 1. Two principal documents are used to obtain the data:

1. The MER (Medical Expense Report), which will be referred to as the "claim." Each time a Study participant uses a medical service, he asks the provider of the service (a doctor, dentist, hospital, etc.) to fill out the medical expense report and send it to the Health Insurance Study.
2. The diary (called the health report). The health report is filled out either weekly or biweekly by the participant, or someone in his family, and mailed to the Health Insurance Study. Information furnished in the diary concerns the participant's use of health services and the days of restricted activity because of health.

Supplementary reporting forms are also used. These consist of both a calendar and some small forms that can be torn off each claim. The supplementary forms make it possible for each participant to record the use of health services and days of restricted activity as soon as they occur. This information is then copied onto the health report when it reaches the family by mail.

Incentives. As indicated in Figure 1, a family is paid either $2 (weekly) or $4 (biweekly) for sending in a completed diary to the Health Insurance Study. Approximately half of the first-year-sample participants filled out weekly diaries, for which they were paid $2 for each diary sent in. The remainder of the sample filled out diaries on a biweekly basis and were paid $4 for each completed diary mailed in. The reporting of control group participants is shown in Tables 1, 2, and 3. This group received no additional payment for providing the study with data on claims. Figure 3 and Table 4 present data from experimental groups. These participants had health insurance administered by the Health Insurance Study and could receive reimbursement for some or all of the charges

listed on the claim forms. Thus, the experimental group had some incentive to provide the study with claims for each health service used.

Follow-up procedures. All data forms received were edited for completeness, and item nonresponse was followed up. After a period of approximately six weeks, an attempt was made to match every health service (e.g., doctor visit) reported in the diary with its corresponding visit reported on a claim (MER). An effort was then made to secure any missing claims. Usually this was a request from the Health Insurance Study for the participant to deliver a blank claim form to the provider and for the participant to ask the provider to fill out the form with the details of the visit in question. It was possible to receive claims for visits that were not mentioned on the health report. No attempt was made, however, to obtain additional health report information for these visits.

Control group results

Underreporting of health services. During the first year of the study, the control group in Dayton appeared to underreport the use of health services substantially. Table 1 shows (1) the expected annual rates of health service use per person under 65 years of age based on the 1975 NCHS Health Interview Survey (HIS) data, (2) the annual per-person rates obtained from diary data, and (3) the annual per-person rates obtained from claims data. Aside from the fact that the National Center for Health Statistics is a major sponsor of this conference, I have several other reasons for basing my expected Dayton use rates on their national estimates: The NCHS estimates are of presumably high quality, the national estimates do not differ importantly from synthetic estimates for the Dayton area, and they do not differ importantly from annual per-person use rates observed in our own baseline survey of the Dayton population. Table 1 shows that the control group reported less than 75 percent of the expected number of annual doctor visits in the diaries and less than 50 percent of the expected number of visits on the claim forms. A similar pattern of underreporting is observed for dental visits. Also, only about 70 percent of the expected number of telephone consultations were reported in the diaries, which is about the same underreporting rate as that observed for outpatient M.D. office visits.

The reasons for the underreporting are not known, but it is possible to hypothesize what may have gone wrong. My guess is that we placed too much uncompensated burden on the

Figure 1
Utilization reporting system: Dayton first-year documents, incentives, and follow-up procedures used to measure consumption in the first year of the experiment

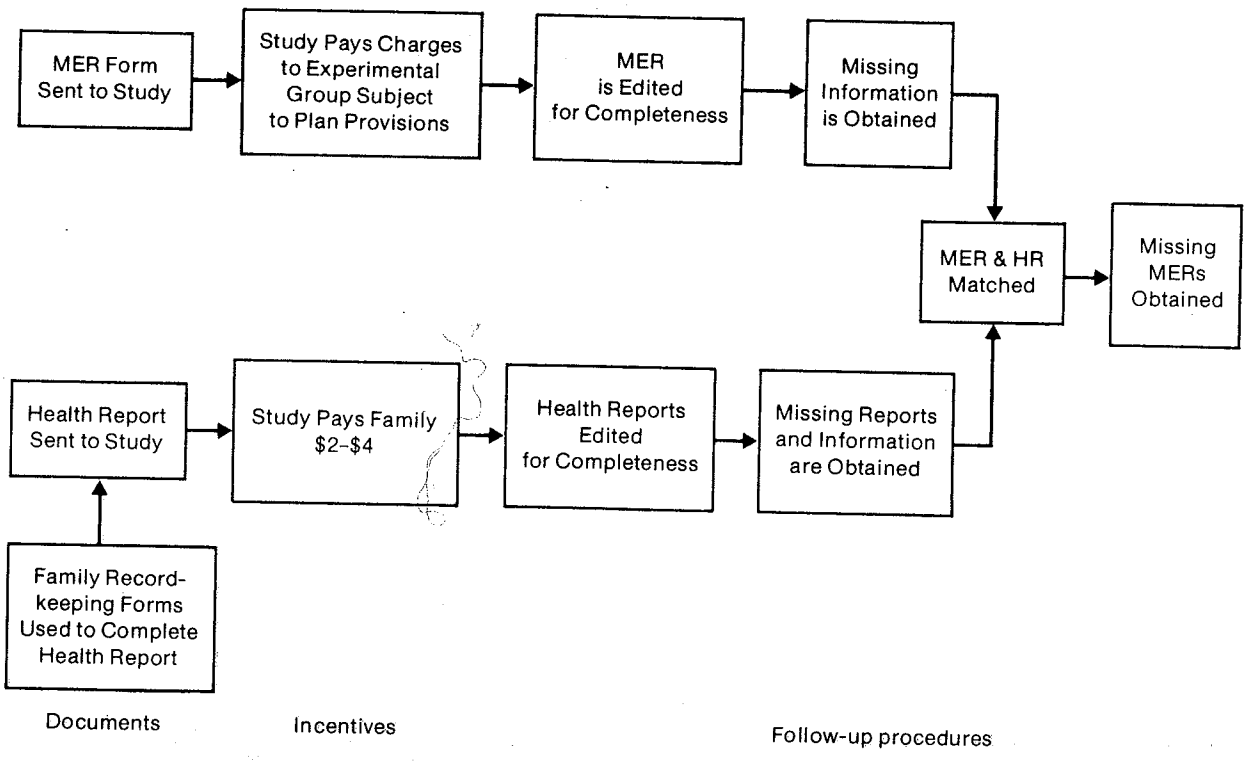


Table 1
Mean annual health service visit rates per person: Dayton control group (N = 631 persons)

Type of visit	NCHS, Health Interview Survey, 1975	Health Insurance Study	
		Diary	Claims
Outpatient (M.D. and non-M.D.)	4.2 (M.D. only)	3.0 (0.25)	1.9 (0.16)
Dental	1.7	0.7 (0.05)	0.5 (0.05)
Telephone	0.7	0.5 (0.06)	Not obtained

NOTES: Means for experimental group (not shown) are in expected range as defined by NCHS estimates and information about price. Standard errors of the control group means, uncorrected for design effects, are shown in parentheses. Telephone consultations are not reported on claims. Both Health Insurance Study and NCHS estimates are for persons under 65 years of age.

Table 2
Mean annual health service visit rates, by type of service and data collection treatment (Control group, N = 631 persons)

Type of service	Data collection treatment		t
	Weekly ($2)	Biweekly ($4)	
	Diary		
Outpatient physician	2.97	3.04	−0.13
Dental	0.66	0.63	0.26
Telephone	0.61	0.38	1.83
	Claims		
Outpatient physician	1.88	1.92	−0.13
Dental	0.55	0.42	1.38

NOTE: t-statistics are based on variances that are unadjusted for design effects. Such adjustments generally increase the size of the estimated variance, hence reducing the absolute value of t.

control group participant. The participant was required to persuade the provider to fill out a claim for each health service used. Neither the participant nor the provider received any additional reimbursement for filling out the claim. If the participant had reported the visit on the health report, the study would contact him after six weeks and insist that he furnish the medical claim information for each visit that he had reported. The participant was motivated on the one hand to remain in the study because it was paying him over $100 per year for filling out diaries. He was not motivated to fulfill the claim-filling requirements if his health service provider voiced objections to the extra paperwork and/or threatened to charge for the extra burden. Participants may have decided on a compromise strategy of not reporting on health reports health services used if they anticipated that a health service provider would object to filling out a claim.

There may be some lessons to be learned from this experience. Just as survey sampling textbooks insist that we must have 100 percent response rates, experimental design textbooks insist that we include control groups in our experiments. But people do not behave like the agricultural plots on which experimental design theory is based. We can be worse off with a control group in an experiment than without it.

Effect of reporting period length. With the possible exception of telephone visits, the difference between the weekly and biweekly data collections had no important or statistically significant effects on the mean reported health service visit rates in the control group. No differences were observed either for diary data or for claim data, as shown in Table 2. Undoubtedly, whatever motivating or memory facilitation effects weekly data collection might have over biweekly collection, they were too subtle to overcome the major underreporting biases observed in the control group.

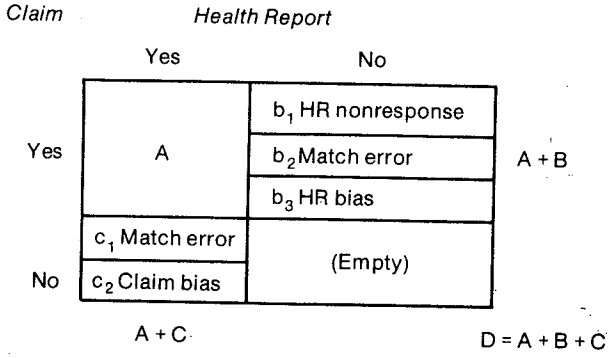
Quality of restricted activity day data. The diary asked two questions about restricted activity due to poor health: (1) the number of days in which any time was taken off from work because of health problems and (2) the number of times that usual daily activities were curtailed for more than half a day because of health. Similar concepts are measured in the NCHS Health Interview Survey, although full days of activity restriction must be experienced before they are counted. Table 3 suggests that the Dayton control group reports work loss at the rates that might be expected based on HIS data. However, the control group appears to report much less restricted activity of other kinds than is observed in the HIS across the nation. Reasons for this last discrepancy are unclear, especially because subsequent efforts by the Health Insurance Study to improve techniques of measuring other restricted activity have not increased the observed rates. I suggest that this is a fruitful area for methodological research.

Table 3
Mean restricted activity days: Dayton control group and national NCHS-HIS

Restricted activity	*Control group*	*NCHS-HIS (1975)*
Work loss	5.7	5.2
	(1.2)	(0.16)
All restricted activity	7.5	15.5
	(2.6)	(0.16)

NOTE: Standard error of mean in parentheses not adjusted for design effects or effects of imputation for diary nonresponse. Annual rates for persons under 65 years of age are shown.

Figure 2
Cross classification of claim and health report utilizations



A = Number matched.
B = Number of unmatched claims.
b_1 = (Number claims for utilizations during period of HR nonresponse)
b_2 = [(Number HR utilizations with "MERleft" checked yes but not matched to claims)/Number HR utilizations with "MERleft" checked yes] × total health report utilizations.
$b_3 = B - (b_1 + b_2)$.
C = Number of unmatched health report utilizations.
$c_1 = b_2$.
$c_2 = C - c_1$.

Proportions of errors in diary/claims data

We will now consider data from the experimental group in the first year of the Health Insurance Study in order to evaluate the characteristics of the outpatient physician visit data that were obtained. Recall that the objectives of the data collection system were to obtain diary reports of all outpatient physician services used by participants and then to obtain claim information for each reported visit. The present data cannot be used to evaluate how well we met the first objective (complete enumeration of all "true" visits). They can be used to evaluate the second objective (obtaining claims for visits mentioned in the diaries) and to gain some insights into the kinds of visits that take place but are not mentioned in the diaries. This evaluation relies on comparing the information in the diaries with that on the claims, matched according to a specific set of match rules.

Match rules. A service is deemed reported on both documents (matched) if both documents agree about:

1. The name of the person receiving the service;
2. The type of service used (doctor visit, inpatient hospital, dentist visit);
3. The date the service was rendered (plus or minus ten days); and
4. The name of the provider of the service (if available).

If a health service is reported on both documents, the service is deemed "matched" and appears in Category A of Figure 2. Unmatched claims appear in Category B; unmatched health report services are in Category C.

Error proportion estimators. The analysis objective is to obtain estimates of the proportion of known service uses that are not reported on either one or the other of the documents. The desired estimates (see Figure 2) are proportions of claim bias (c_2), health report form nonresponse bias (b_1), and residual health report bias (b_3). The estimation problem is one of separating random match errors from reporting biases. Theory-based procedures for estimating random cross-classification error proportions must assume independence of the two data sources, an assumption that clearly does not hold in the present experiment. Instead, I use another procedure that also rests on strong assumptions but appears more appropriate for these data. It uses information provided by the participant concerning whether he left a claim document with the provider when he obtained the services of the provider. If the participant said he left the claim, I assume that this is true; I also assume that the provider filled the claim out and returned it to the Health Insurance Study. Thus, there is a claim in the data pool that should match the information on the health report for this particular service. The ratio of the number of such unmatched health report services to all health report services for which the participant initiated a claim, multiplied by the total number of health report utilizations, is an estimate of the random match error rate for both the health report and the claim information (b_2, c_1).

The proportion of unmatched claims due to health report nonresponse (b_1) is obtained by counting the number of services reported only on claims if the date of service is within a period of health report form nonresponse. The proportions of residual health report bias and claim bias are obtained, then, by differencing, as indicated in Figure 2.

Outpatient physician visit error. The results of this kind of estimation for outpatient physician visits in the first year of the Health Insurance Study (experimental participants) indicates that over two-thirds of the detected services could be matched and that most of the failures to match were due either to health report underreporting (7 percent from nonresponse bias, 9 percent from residual bias) or to random match errors (14 percent), as shown in Figure 3. The claim omission bias appears to be very low (2 percent). This indicates that we failed to obtain claim information for about 2 percent of the physician visits that we learned about from the Utilization Reporting System. So the dual-diary system appears to have achieved its second objective, namely, to get claim data for all visits detected on the diaries.

This analysis does not attempt to estimate the number of outpatient physician visits that occurred but were reported on neither of the documents. An accurate estimate cannot be made from these data, but the Health Insurance Study will be collecting other kinds of information to obtain the estimate.

Characteristics of visits not reported on diaries

Table 4 presents the results of a very preliminary analysis of the characteristics of outpatient physician visits that were detected only on

Figure 3
Error proportions: Outpatient physician visits (Experimental group)

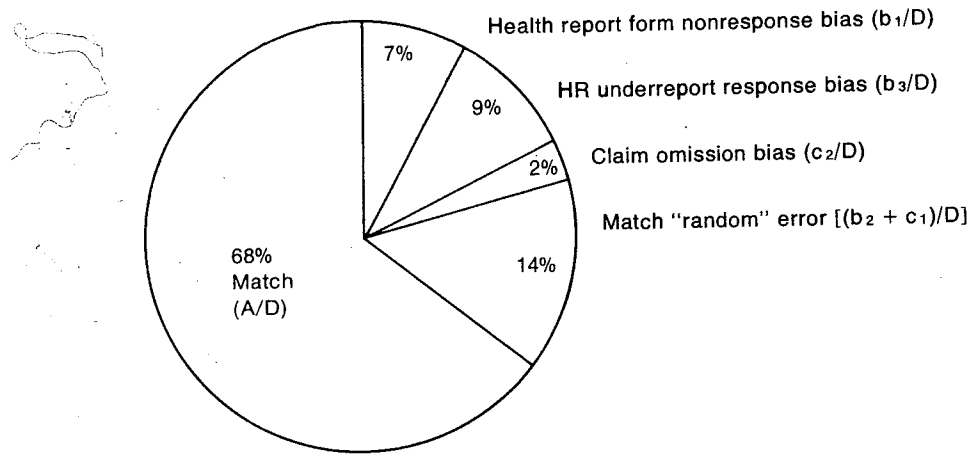


Table 4
Regression analysis of physician visits not reported on diaries (Experimental group)

Explanatory variable	*Regression coefficient*	*Standard error*	*Standardized coefficient*
Place of service:			
Independent laboratory	.43	.03	.15
Emergency room	.31	.02	.16
Other—not in M.D. office	.24	.02	.14
M.D. office		(omitted)	
Type of service (multiple types per visit possible):			
Radiology	.10	.02	.06
Pathology	.01	.01	.01
Psychiatric	.01	.02	.01
Total cost of visit	$-.53\times10^{-3}$	$.20\times10^{-3}$	−.03
Length of time in study	$-.95\times10^{-4}$	$.4\times10^{-3}$	−.02
Diary form nonresponse	.80	.02	.49
Constant term	.16		

NOTES: The unit of analysis is the "visit." The dependent variable equals 0 if the visit is reported on both the diary and the claim or 1 if the visit is reported on the claim only. A positive regression coefficient indicates that the variable is a characteristic of visits not reported on the diary. 6,210 visits of a possible 6,236 are used in this exploratory analysis. An ordinary least squares regression procedure was used. Pathology service and independent lab place of service are positively correlated so that independent effects are difficult to determine. Constant term includes classification error and unexplained bias effects.

claims (i.e., they were not reported in the diaries). The coefficients reported are ordinary least squares regression coefficients. The dichotomous dependent variable has the value of zero if the doctor visit is reported both in the diary and on the claim; it has a value of one if the visit is reported on the claim only.

Diary-form nonresponse is the most important reason for diary underreporting of physician visits. The other characteristics of visits underreported in the diary are that they take place somewhere out of the ordinary or that the service rendered is ancillary (e.g., radiology). These two sets of predictors are highly correlated; the ancillary services are often rendered outside of the physician's office.[1] These results point to some possible definition problems in that respondents appear not to share the research worker's definition of what constitutes a separate outpatient service. A patient may go to the doctor's office and consider this one visit. However, if the visit results in independent radiology or laboratory services, the researcher considers that additional services have been performed. The patient may consider one trip to the emergency room as a single visit. The researcher considers that several services have been rendered if, for example, separate bills are received for physician services in the emergency room and for the basic emergency room service.

Note that visits involving "psychiatric" services are not underreported at higher rates than other visits.

The negative coefficient on the total cost of the visit suggests that less expensive visits are more likely than more expensive ones to be unreported in the diaries.

The length-of-time-in-study coefficient suggests that some learning may take place: The probability of a diary underreport decreases over time, presumably as participants become more familiar with their reporting roles.

Conclusions

On the basis of these preliminary analyses, the diary/claim system appears to work fairly well in obtaining complete information about the use of health services if incentives are enough to compensate for the participant's reporting burden. If incentives are insufficient, which may have happened for the control group, the quality of both diary and claim data may suffer. The diary system may miss some kinds of visits owing to form nonresponse, forgetting, and definitional problems. Finally, there is some question about the validity of the data reported for restricted activity days; estimates from the diary are substantially lower than estimates from the national Health Interview Survey. Neither NCHS nor the Health Insurance Study has validated their restricted activity day self-reports. Perhaps, this is an area that needs further measurement evaluation research.

Footnote

[1] If the prediction equation does not include independent lab as a separate dummy variable, the pathology type of service coefficient is statistically significant.

On the use of memory aids in the Los Angeles Health Survey*

Alfred C. Marcus, School of Public Health
University of California at Los Angeles

Verbrugge's call for more research on the use of the memory aid technique is especially timely from our point of view because we are currently engaged in such research on the Los Angeles Health Survey (LAHS). We have described the methodology of the Los Angeles Health Survey in more detail in our paper on telephone interviewing (Session 2). I would now like to share some unpublished data that we have on the use of the memory aid technique in our survey.

These data come from a one-year panel survey conducted as part of the LAHS (1976–1977). The sampling design was a three-stage random probability sample of Los Angeles County that was drawn from the master sampling frame at the Institute for Social Science Research (N = 1,210). At the completion of the first panel interview, which was conducted face-to-face, respondents were randomly assigned to a memory aid or a no memory aid condition. The memory aid was designed to record information pertinent to questions asked in subsequent waves of the panel (e.g., accidents, illnesses, injuries, bed days, doctor visits, etc.). For the next 12 months, respondents were interviewed by telephone at approximately six- to eight-week intervals. The final interview was then conducted face-to-face. All data presented here come from the first and last interviews only.

Comparisons of the memory aid/no memory aid conditions produced the following findings:

1. *Demographic comparisons.* At the start of the survey, we had 1,210 respondents, 606 (50.1 percent) of whom were assigned to get memory aids, while the remaining 604 (49.9 percent) received none. As expected from the randomization process, no significant differences were found between the two groups in their sociodemographic characteristics at the start of the survey. Generally, the same comparability was found at the last interview; hence, no evidence of differential attrition existed.
2. *Attrition.* For the sample as a whole, we had 327 noncompletions at the last interview in the panel survey, for a noncompletion rate *from all sources* of 27 percent. When broken down by memory aid group, we found nearly identical rates at the last interview. For the memory aid condition, there were 170 noncompleted interviews (28 percent), compared with 157 noncompleted interviews (26 percent) in the no memory aid condition. These differences are not statistically significant. Further analysis is currently in progress to determine whether differences exist within particular categories of attrition (e.g., refusals).
3. *Reporting of health information.* As shown in Table 1, we consistently found greater reporting of illnesses and disability in the memory aid condition. There was also more frequent reporting of doctor-recommended medications and treatments. However, there were no differences with respect to self-reported compliance with medical advice, doctor visits for recent health problems, and the number of current preventive health behaviors engaged in by the respondent. We should note that these data are generally consistent with previous research (see, for example, Verbrugge, 1980). That is, most studies have found higher reporting levels for illness and disability within the memory aid condition, but little difference has been found in the reporting of health actions (i.e., use of physician services, compliance with medical advice, practice of preventive health behavior).
4. *Respondent perceptions of the memory aid technique.* During the last interview of the panel, respondents in the memory aid condition were asked several questions about their ex-

*This paper is an expanded version of remarks made during the open discussion on the Verbrugge and Depner paper and the Marquis presentation.

periences with the memory aids. Those who indicated that they had used their memory aids at least once during the one-year study period (N = 252, or 60 percent) were asked the following:

(1) Did the memory aid help you answer the questions we asked during the past 12 months?

Of those who used the memory aid at least once, 62 percent answered "yes" (156/252).

(2) Did the information you included in your memory aid in any way influence your feelings or behavior regarding your own health?

Of those who used the memory aid at least once, 34 percent answered "yes" (85/252).

Responses to the latter question suggest that the memory aids may have been reactive for a substantial proportion of our respondents. This finding was examined further to determine whether sociodemographic subgroups differed in their reactiveness. For most of the demographic variables that we examined (e.g., age, sex, education, total family income, and marital status), no significant differences were found in reactiveness for the memory aid technique. However, we did find that racial/ethnic minorities were more likely to report being influenced (see Table 2), as were respondents employed in occupations *other*

Table 1
Comparison of memory aid and no memory aid condition in reporting of health data

Items from final panel interview	*Scoring*	*Memory aid (N = 436)*	*No memory aid (N = 447)*	$p \leq$
Was R sick in past two months from an acute illness?	% Yes	41.5%	34.7%	.02
Did R have to cut down on normal activities because of sickness (past two months)?	% Yes	26.8	22.5	.06
Were there other times when R just wasn't feeling as well as usual (past two months)?	% Yes	46.4	41.3	.06
Is doctor currently prescribing or advising certain medications or treatments?	% Yes	43.6	37.0	.02
Has R seen doctor about not feeling well in past two months? ...	% Yes	16.5	17.1	n.s.
Has R done anything in past two months to maintain or improve health that was not doctor recommended?	% Yes	47.7	51.0	n.s.
Has R complied with all doctor recommendations?	% Yes	77.9	74.0	n.s.

Table 2
Racial/ethnic differences in reported reactiveness of memory aid

		Race/Ethnicity		
		White/Anglo	*Hispanic*	*Black*
Did the memory aid influence your feelings or behavior regarding your own health?[a]	Yes	47 (26.7%)	16 (51.6%)	17 (58.6%)
	No	129 (73.3%)	15 (48.4%)	12 (41.4%)
	Total	176 (100.0%)	31 (100.0%)	29 (100.0%)

[a] Question was originally asked of those respondents who reported using the memory aid at least once (N = 252).

$p < .001$.

than professional/technical (see Table 3). Respondents were also asked to describe their reactions to the memory aid in more detail. Using an open-ended question format, we found the most frequent reactions included becoming more aware of health (29 percent) and becoming more careful about health (10 percent). Both of these responses showed the same occupation and racial/ethnic differences described above.

Please note that these findings are preliminary and based on a relatively small number of cases. We have shared these data to illustrate the need for more systematic research on the memory aid technique. We are hopeful that the LAHS, together with Verbrugge's Detroit survey, will extend our understanding of the use of memory aids in health surveys.

Table 3
Differences among occupational groups in reported reactiveness of memory aid

		Occupational Group			
		Professional/ technical	*Managerial/ sales*	*Clerical*	*Crafts/ operatives*
Did the memory aid influence your feelings or behavior regarding your own health[a]	Yes	10 (19.6%)	13 (40.6%)	26 (40.6%)	21 (37.5%)
	No	41 (80.4%)	19 (59.4%)	38 (59.4%)	35 (62.5%)
	Total	51 (100.0%)	32 (100.0%)	64 (100.0%)	56 (100.0%)

[a] Question was originally asked of those respondents who reported using the memory aid at least once (N = 252).

$p < .08$.

Medical Economics Survey-Methods Study: Cost-effectiveness of alternative survey strategies

Richard Yaffe, Health Services Research and Development Center, Johns Hopkins Medical Institutions

Sam Shapiro, Health Services Research and Development Center, Johns Hopkins Medical Institutions

Introduction

Data needs relating to policy and planning issues involving national health insurance, cost containment, and administration of federally supported health care programs, among others, have led to the launching of two large-scale national surveys during the last three years. The National Medical Care Expenditure Survey, sponsored by NCHS and NCHSR, is nearing completion, while NCHS and HCFA are about to begin the National Medical Care Utilization and Expenditure Survey, which is expected to become a continuing survey on a biennial basis. The emphasis in both of these panel surveys is the collection of current data on health care expenditures and the extent of third-party coverage and payments. This paper will present findings from a methodological study of health survey procedures that have important implications for the planning of national health expenditure surveys and for the interpretation of data from such surveys.

At state and local levels, new interest in health surveys has been created by the requirements of the planning legislation PL 93–641. Many Health Systems Agencies and State Health Planning and Development Agencies are feeling handicapped in their planning activities by the lack of small area estimates on health status, health care utilization, health insurance coverage, and health expenditures. Although the current funding level for HSAs does not, in general, provide the resources to carry out household surveys, several areas have found ways to do so through pooling resources of state and local agencies, an example being the HSA in Virginia. Although local areas are not likely to utilize survey methods as complicated as those evaluated in this study, the results raise concerns about reporting accuracy that should be important to all users of health survey data.

The purpose of the study was to test the feasibility and effectiveness of several different survey strategies that were designed to obtain accurate information concerning medical care utilization, expenditures, and the extent of third-party payments. Results will be presented concerning the feasibility of successfully enrolling and maintaining household respondents in a panel survey with repeated callbacks and the relative effectiveness of telephone versus in-person interviews and monthly versus bimonthly callbacks. The need for obtaining information from provider and/or third-party-payer records to supplement household reporting will also be addressed. Results on reporting accuracy will be presented separately for different types of health care services and for various population subgroups.

The Medical Economics Survey-Methods Study (MES-MS) was carried out under contract with the National Center for Health Statistics during 1975–76 (Health Services Research and Development Center, 1977; Shapiro et al., 1976; Yaffe et al., 1978). In earlier attempts to gather health expenditure data through the Health Interview Survey (HIS), NCHS found several problems (U.S. National Center for Health Statistics, 1975). For example, the household respondent is frequently unaware of the total charges and total expenditures for many health services, since often at least part of the costs are met by insurance or public financing programs with the dollar figures involved unknown to the respondent. Furthermore, the HIS is limited to a two-week recall period for certain types of utilization, which permits the estimation of average annual expenditures but not the distribution of expenditure levels.

To overcome these problems, NCHS concluded that a panel sample survey procedure would be required in which information was obtained repeatedly from the same households over a relatively long period of time. This design would allow the collection of data for an entire year of experience for a sample of households, while maintaining a reasonable re-

call period. In addition, charges that were unknown at the time of an interview because a bill had not yet been received could be asked about at a follow-up interview. It was also clear, however, that a panel survey by itself would still not, in many cases, obtain data on charges that were billed directly to third-party payers and that supplementary sources would be needed.

In designing a panel survey, several options had to be considered: in particular, the periodicity of interviews and the extent to which telephones could be used in place of in-person interviews. In addition, the feasibility of gaining and keeping the participation of a panel over an extended period of time had to be tested, and the extent to which data from record sources were required needed to be assessed.

Study design, household survey, and record check procedures

The survey was designed to obtain, for all members of the sampled households, information on medical utilization and expenditures over a seven-month period (one month retrospectively at the initial visit and six months prospectively through repetitive contacts). The structure and content of the interview departed from the HIS primarily by probing for greater detail concerning utilization of medical, hospital, and dental care; prescriptions filled; identification of sources from which care was received and regular source of care; insurance coverage; health care charges; and sources of payment (insurance and out-of-pocket).

The household sample originally included 1,035 households distributed between two geographic areas: (1) 624 sample households in Baltimore City and two adjacent counties, with a population of 1.5 million and (2) 411 sample households in Washington County, Maryland, with about 100,000 residents, two-thirds of whom live in rural areas and the other third in a small central city, Hagerstown. One-half of the total sample was drawn as an area probability sample; the other half was drawn from health insurance record sources, including HMOs in the area, and consisted of families with past experience predictive of relatively high utilization during the study period.

Household respondents were requested at the initial interview to keep a ledger-type diary, which was meant to serve as a record-keeping system and memory aid for the respondent rather than as a primary data collection instrument. The recall period for the follow-up interviews was always the time period since the previous interview. Prior to each follow-up interview, and at the end of the study period, a cumulative Summary Report was mailed to each household. Respondents were asked to review the report and to make any necessary additions or corrections, including entries about bills received for services obtained during an earlier period.

Through a randomization procedure, the eligible households were divided into four experimental groups, which were assigned different follow-up techniques, as shown in Table 1, for collecting information over the six-month period after the initial interview. At issue was whether periodicity (monthly versus bimonthly interviews) and type of contact (in-person versus telephone) are significant factors in the cost-effectiveness of conducting this type of panel survey.

Participation in the survey was measured in terms of the extent to which households agreed to participate in the initial interview (initial response rate) and the extent to which households completed the entire schedule of interviews (completion rate), i.e., the proportion of households completing initial interviews and all succeeding interviews. As shown in Table 2, the initial response rate was 77.5 percent overall—72.1 percent in the metropolitan Baltimore area and 85.6 percent in largely rural Washington County. The completion rate was 67.0 percent overall — Baltimore area, 60.1 percent; Washington County, 77.4 percent. The results for the monthly alternating telephone-and-in-person approach (Procedure 3 in Table 1) were signifi-

Table 1
Experimental study design

Survey strategies	*Initial interview*	*1*	*2*	*3*	*4*	*5*	*6*
1. Monthly telephone	IP[a]	Tel[b]	Tel	Tel	Tel	Tel	IP
2. Bimonthly telephone	IP	—	Tel	—	Tel	—	IP
3. Monthly telephone and in person	IP	Tel	IP	Tel	IP	Tel	IP
4. Bimonthly in person	IP	—	IP	—	IP	—	IP

[a] IP = In-person interview.

[b] Tel = Telephone interview.

cantly different from the other three. For this procedure, the combined effect of the initial response rate and attrition resulted in only 59 percent of eligibles completing the study. The corresponding figure for all three of the other procedures was close to 70 percent. By far the largest difference between Procedure 3 and the others was found in the Baltimore area.

At the completion of the household survey, each reported medical care provider and third-party payer was contacted for information on utilization and expenditures. The objective was to obtain information required to estimate the extent of reporting error in household data under each of the alternative survey strategies and to determine the importance of different types of record data in such surveys.

Providers were asked to furnish independent reports of utilization and expenditures for the time period involved for individuals in the survey who reported receiving one or more services from them. If the household failed to report the name of a provider whom they utilized during the study, a provider record check could not uncover such omissions. Several factors in the study were designed to minimize the effect of this situation. First, respondents were asked to identify the "regular doctor and dentist" for each household member at the initial interview. All "regular" providers were contacted in the record check whether or not any actual utilization was reported by the household. The problem was further reduced by the fact that some part of the unreported utilization could be discovered in the record check of third-party claims data.

A summary of completion rates achieved in the provider record check is shown in Table 3. The completion rates were generally higher in the Baltimore area, with the greatest difference being for doctor forms (81.8 percent in the Baltimore area versus 42.6 percent in Washington County). The low response rate for doctors in Washington County compared with the Baltimore area can be attributed to the combined effect of a much greater average response burden per physician in Washington County (see Table 4) and a greater reluctance among

Table 2
Completion of interviews among eligibles

Household procedure	Eligible	Completed	All scheduled interviews completed	Percent initial completed	Percent all completed
			Washington County		
1. Monthly telephone	100	86	80	86.0	80.0
2. Bimonthly telephone	101	85	78	84.2	77.2
3. Monthly telephone and in person	101	88	75	87.1	74.2
4. Bimonthly in person	109	93	85	85.3	78.0
All	411	352	318	85.6	77.4
			Baltimore area		
1. Monthly telephone	157	113	96	72.0	61.1
2. Bimonthly telephone	158	118	103	74.7	65.2
3. Monthly telephone and in person	159	105	79	66.0	49.7
4. Bimonthly in person	150	114	97	76.0	64.7
All	624	450	375	72.1	60.1
			All areas		
1. Monthly telephone	257	199	176	77.4	68.5
2. Bimonthly telephone	259	203	181	78.4	70.0
3. Monthly telephone and in person	260	193	154	74.2	59.2
4. Bimonthly in person	259	207	182	79.9	70.3
All	1,035	802	693	77.5	67.0

Table 3
Provider record check completion rates, by type of provider and geographic area

Provider type	Baltimore area		Washington County		Total–both areas	
	Total number	Percent complete	Total number	Percent complete	Total number	Percent complete
Doctor	500	83.2	147	63.9	647	78.8
Forms	1,055	81.8	1,269	42.6	2,324	60.4
Dentist	231	92.1	69	76.8	300	88.7
Forms	676	93.9	778	63.4	2,454	77.6
Pharmacies	176	97.7	39	87.2	215	95.8
Forms	550	95.1	540	96.3	2,090	95.7
Hospital/clinics	75	78.7	27	81.5	102	79.4
Forms	553	89.9	273	86.1	806	88.6
Other	52	90.4	21	90.5	73	90.4
Forms	80	91.2	80	86.3	160	88.8

Table 4
Average response burden (Forms/provider)

Provider type	Baltimore area	Washington County
Doctor	2.11	8.63
Dentist	2.93	11.28
Pharmacies	3.13	13.85
Hospital/clinics	7.11	10.11
Other	1.54	3.81

the Washington County doctors to participate in research studies of this nature. The great majority (78 percent) of all third-party payer (TPP) forms were completed by four insurers/payers. The remaining 22 percent of the TPP forms were sent to 50 different insurers/payers, and 55 percent of these forms were completed. The overall completion rate for all TPP forms was 90 percent.

Data matching

After all household and record information was collected, data matching procedures were undertaken in order to determine the "Best Data" set, that is, the criterion to be used to measure accuracy and completeness of reporting in the household survey. For several reasons it was not possible to use the record data as the sole criterion. Record information was not available in many cases because the record sources refused to participate in the study or because in a small proportion of households respondents refused to give permission to have their records checked (10.1 percent in Washington County and 19.2 percent in the Baltimore area). Also, record sources themselves often contained errors and omissions, so that it could not be assumed that an unverified visit did not take place.

The approach taken in the matching process was to use all data available from the household, provider, and third-party payer to determine the Best Data set using specified rules and guiding principles. The matching was carried out through a computerized process, which was followed by a manual review of all decisions. The process was complex and time consuming, and only a brief account will be given here.

In general, when information was available from one or more record sources and these sources disagreed with the household data, preference was usually given to the record sources, especially in the case of likely underreporting on the part of the household respondent. When the household reported utilization that was not confirmed by a cooperating record source, the general rule was to include the household-reported utilization in the Best Data set. Due to the nature of the methods used in the household survey, i.e., relatively short recall periods and the Summary Report feature, the assumption was made that unconfirmed utilization was more likely to be a problem of incomplete provider records than of overreporting on the part of the household respondent.

In the computerized matching process, all data on utilization, charges, and payments were sorted in a hierarchical fashion according to household, person, service, provider, and date. For each unique Patient-Provider Pair (PPP), the matching program then attempted to match utilization reported by the household, provider, and third-party payer. All of the computer matching decisions were printed in highly formatted outputs and were reviewed manually. The majority of situations that had to be dealt with manually fell into two categories: un-

matched utilization and differing levels of aggregation of the data reported in the various data sources.

Measures of reporting accuracy

Accuracy of reporting was examined for utilization, charges, third-party payments, and out-of-pocket expenditures in specific categories of utilization such as physician outpatient and inpatient care, hospital utilization, dental care, and prescription drugs. Two different measures of reporting accuracy were used in comparing household-reported data with the Best Data set. First, the Percent Best—All Data was calculated by comparing household (HH) data and Best Data considering all data in both the HH and Best Data sets, even if there was no verifying record source; i.e.,

$$\text{Percent Best—All Data} = \frac{\text{All Household Data}}{\text{All Best Data}} \times 100.$$

Unverified HH data may arise because the respondent did not give permission to check records, because the record sources did not respond, or because the record source did not report a particular utilization reported by the household, which was nevertheless incorporated in the Best Data set. The second comparison uses the Percent Best—Verified Data, calculated by eliminating all unverified HH data from both the HH and Best Data sets; i.e.,

$$\text{Percent Best—Verified Data} = \frac{\text{Verified Household Data}}{\text{Verified Best Data}} \times 100.$$

There are potential biases inherent in both measures. The Percent Best—All Data measure tends to overestimate the accuracy of reporting, since in comparing HH and Best, it is implicitly assumed that all unverified data are 100 percent accurate. For all strategies combined, 58.9 percent and 68.9 percent of all charges data in the Best Data set were verified in the Baltimore area and Washington County, respectively. In considering only the verified data, the implicit assumption is that the unverified data were reported with the same degree of accuracy as the verified data.

Since both measures of accuracy of reporting are subject to certain biases, the absolute value of the percentage must be interpreted with caution. In this paper, in making comparisons among household survey strategies and among types of medical services, the Percent Best—Verified Data is used. In making comparisons among population subgroups, the Percent Best—All Data is used.

Summary of study results

As mentioned previously, household samples were drawn from two distinctly different geographic areas: one a large metropolitan area and the other a largely rural area with a small central city. In general, household reporting in Washington County was considerably better than in Baltimore; therefore, results for the two areas are shown separately.

The accuracy of reporting various categories of utilization by households is shown in Table 5. In the Baltimore area, the services reported with the greatest accuracy were emergency

Table 5
Accuracy of utilization reporting in the household survey, by type of service and area

Type of service	Baltimore area		Washington County	
	Utilization rate[a] (Best Data)	Percent Best—Verified Data[b]	Utilization rate[a] (Best Data)	Percent Best—Verified Data[b]
Physician outpatient:				
Total	2.73	65.8	2.38	79.3
Office	1.68	72.9	1.99	83.5
Clinic	.923	53.5	.238	39.1
Emergency room	.130	94.2	.155	96.2
Other medical providers	.552	56.9	.364	79.4
Dentist	.940	81.5	.929	86.3
Prescribed drugs	4.61	60.5	4.76	75.0
Hospital:				
Discharges	.094	93.9	.079	96.6
Days	.655	91.5	.555	92.4

[a] Average per-person utilization rate for the seven-month study period as determined from the Best Data set.

[b] Percent Best—Verified Data = $\frac{\text{Verified Household Data}}{\text{Verified Best Data}} \times 100.$

room visits (94.2 percent) and inpatient hospital utilization (93.9 percent). Physician outpatient visits were reported with considerably less accuracy (72.9 percent for office visits and 53.5 percent for clinic visits), as were visits to "other medical providers" (56.9 percent) and prescription drugs (60.5 percent). Dental utilization was reported with greater accuracy than physician office visits (81.5 percent).

In Washington County, the pattern of differences in reporting accuracy among service types was similar to that in the Baltimore area. However, reporting accuracy was better in Washington County for every service type with the exception of clinic visits. As was the case in the Baltimore area, the services reported with the greatest accuracy included hospital discharges (96.6 percent), hospital days (92.4 percent), and emergency room visits (96.2 percent). Physician outpatient visits to private offices were reported with less accuracy (83.5 percent), as were visits to other medical providers (79.4 percent), dentists (86.3 percent), and prescriptions (75.0 percent). Visits to clinics were reported least accurately (39.1 percent).

The accuracy of reporting of charges by household is shown in Table 6. Charges for all services were reported at 75.7 percent in the Baltimore area and 93.0 percent in Washington County. In Baltimore, services for which charges were reported with greatest accuracy included dental charges (89.5 percent), hospital charges (86.9 percent), and physician inpatient charges (78.0 percent). Outpatient services, including physician charges, prescriptions, and other medical providers were reported with less accuracy.

In Washington County, charges for hospital inpatient care and physician inpatient services were slightly overreported; this apparent overreporting is probably due to imputation procedures that were used to estimate missing charges in the household data. The accuracy of charges reporting closely parallels the accuracy of utilization reporting noted earlier for the other service types.

Table 7 summarizes the two-way comparison of accuracy of utilization and charges reporting based on Percent Best—Verified Data for the telephone versus in-person and the monthly versus bimonthly strategies for the Baltimore area. In-person strategies achieved a higher degree of accuracy of utilization reporting for all service types but one, other medical providers, for which the difference was slight. In comparing the monthly and bimonthly strategies, neither strategy was clearly superior. For charges reporting, the in-person strategy achieved a higher degree of accuracy for every type of service except physician office visits, for which the difference is very slight. No such clear-cut difference is apparent in comparing the monthly and bimonthly strategies. Analysis of variance indicated a significant effect for telephone versus in-person for both utilization ($p = .008$) and charges ($p = .030$) reporting, while no significant differences were found for the monthly versus bimonthly strategies.

In comparison to the Baltimore area, where the in-person strategy was superior for utiliza-

Table 6
Accuracy of charges reporting in the household survey, by type of service and area

Type of service	Baltimore area		Washington County	
	Charges[a] (Best Data)	Percent Best—Verified Data[b]	Charges[a] (Best Data)	Percent Best—Verified Data[b]
Physician outpatient:				
Total	60.26	55.5	38.24	76.7
Office	30.98	68.2	27.15	78.3
Clinic	19.93	38.0	2.89	30.6
Emergency room	9.25	64.4	8.10	90.4
Other medical providers	14.96	61.1	4.64	83.0
Dentist	28.83	89.5	27.25	104.3
Prescribed drugs	14.19	64.7	16.27	77.3
Physician inpatient	23.33	78.0	24.06	95.9
Hospital	106.34	86.9	59.63	100.6
All services	247.92	75.7	169.98	93.0

[a] Average per-person charges for the seven-month study period as determined from the Best Data set.

[b] Percent Best—Verified Data = $\frac{\text{Verified Household Data}}{\text{Verified Best Data}} \times 100$.

tion reporting, there is no consistent difference among the survey procedures in Washington County (Table 8). For charges reporting in Washington County, the in-person and bimonthly strategies did slightly better based on a comparison of charges for all services. However, this difference is not consistent across service types. Analysis of variance showed no significant differences among strategies.

Table 9 displays the percentage distribution of persons by level of total medical charges for the study period, as determined by the HH and Best Data sets for the Baltimore area and Washington County samples. In comparing HH and Best in Baltimore, the percentage in each group decreases from HH to Best for the $0–$99 categories, is the same for the $100–$149 category, and increases for all higher categories. In Washington County, the percentage in each group decreases from HH to Best for the categories below $100 and increases for those categories above $100. This shift to higher charge categories in going from HH to Best was apparent for all survey strategy groups.

Considerable differences were found in reporting accuracy among population subgroups in the Baltimore area, while few differences were found in Washington County. For exam-

Table 7
Two-way comparisons of Percent Best—Verified Data measures among household strategies for utilization and charges, by type of services (Baltimore area)

Type of service	Telephone vs. in-person		Monthly vs. bimonthly	
	Utilization	Charges	Utilization	Charges
Physician outpatient:				
Total	(−5.6)	(−6.1)	+6.8	(−5.0)
Office	(−1.6)	2.1	(−3.8)	(−5.2)
Clinic	(−10.0)	(−9.4)	+17.7	(−3.7)
Emergency room	(−11.4)	(−19.6)	(−1.3)	(−28.4)
Other medical providers	0.90	(−8.1)	(−6.9)	13.3
Dentist	(−2.1)	(−13.2)	2.0	(−4.4)
Prescribed drugs	(−9.5)	(−8.5)	3.7	15.1
Physician inpatient		(−17.9)		(−3.8)
Hospital:		(−8.1)		17.3[a]
Discharges	(−1.1)		6.5	
Days	(−10.0)		15.5	
Total charge		(−5.8)		4.4

[a] This large discrepancy was due to the presence of two outliers for the bimonthly strategies.

NOTE: Figures shown are the signed differences between the Percent Best—Verified Data measures for strategies indicated. Negative values (shown in parentheses) indicate better accuracy of reporting for the second procedure listed in each comparison (in-person or bimonthly).

Table 8
Two-way compensation of Percent Best—Verified Data measures among household strategies for utilization and charges, by type of service (Washington County)

Type of service	Telephone vs. in-person		Monthly vs. bimonthly	
	Utilization	Charges	Utilization	Charges
Physician outpatient:				
Total	(−1.9)	(−3.0)	(−3.4)	(−1.7)
Office	(−3.4)	(−6.2)	(−4.3)	(−4.6)
Clinic	8.2	6.2	(−1.2)	5.4
Emergency room	(−7.0)	3.4	(−1.8)	(−3.6)
Other medical providers	11.4	19.6	(−2.7)	4.8
Dentist	0.9	10.7	(−1.0)	(−7.3)
Prescribed drugs	7.5	9.6	2.3	2.2
Physician inpatient		(−6.9)		(−8.3)
Hospital:				
Discharges	(−0.4)	1.3	(−7.3)	1.2
Days	(−4.3)		(−8.7)	
Total charge		(−3.0)		(−5.6)

NOTE: Figures shown are the signed differences between the Percent Best—Verified Data measures for strategies indicated. Negative values (shown in parentheses) indicate better accuracy of reporting for the second procedure listed in each comparison (in-person or bimonthly).

ple, as shown in Table 10, the reporting acuracy for charges was considerably poorer for the lowest income group (below $5,000). The accuracy of reporting charges was greater for those households where the head of household had completed high school than for those households where the head of household did not complete high school (90 versus 77 percent). Reporting accuracy was considerably better for whites than for nonwhites (89 versus 74 percent).

Table 11 shows analogous data concerning the reporting accuracy for physician outpatient visits. As with total charges, those with the lowest incomes, those with less than a high school education, and nonwhites reported less accurately than others. These findings were consistent for all types of medical services.

Tables 12 and 13 show the accuracy of household reporting for total charges and for physician outpatient visits by population subgroups and survey strategy. For those subgroups with the poorest reporting (i.e., those with lower incomes, with less education, and nonwhites), the in-person and monthly strategies did improve somewhat the accuracy of reporting. However, even with these more intensive interview methods, the reporting accuracy for these

Table 9
Percentage distributions of persons by level of total medical charges for the study period: household data compared with Best Data by area

Level of total medical charges (per person)	Baltimore area		Washington County	
	HH	Best	HH	Best
Total persons	1,259	1,259	1,041	1,041
$0	24.5	22.5	29.9	28.8
$1–$49	29.1	26.6	33.7	32.9
$50–$99	16.3	15.6	18.9	18.1
$100–$149	9.1	9.1	8.6	9.1
$150–$199	4.4	5.7	4.3	5.0
$200–$299	4.8	6.6	4.3	4.8
$300–$499	3.3	3.7	2.8	3.7
$500–$999	3.4	4.2	3.8	4.0
$1,000–$1,999	3.1	3.3	2.1	2.2
$2,000+	2.0	2.6	1.4	1.3
	100%	100%	100%	100%

Table 10
Household reported changes data shown as a percent of Best Data, by household income, education of head of household, and color of household respondent (Baltimore area)

Category	Number of persons	Per-person charges (Best Data)	Percent Best—All Data
Household income:			
$0–$4,999	198	$291	69
$5,000–$9,999	246	235	86
$10,000–$14,999	225	274	91
$15,000–$19,999	164	214	94
$20,000–$24,999	142	328	93
$25,000+	185	182	89
Unknown	99	201	81
Education level of head of household:			
0–8	187	$286	76
9–11	290	206	78
12	318	297	91
Some college	437	236	90
Color of household respondent:			
White	861	$284	89
Nonwhite	386	171	74

Table 11
Household reported utilization of physician outpatient services shown as a percent of Best Data, by household income, education of head of household, and color of household respondent (Baltimore area)

Category	Number of persons	Visit per person (Best Data)	Percent Best—All Data
Household income:			
$0–$4,999	198	3.46	66
$5,000–$9,999	246	2.10	73
$10,000–$14,999	225	2.79	84
$15,000–$19,999	164	2.47	92
$20,000–$24,999	142	3.24	87
$25,000+	185	2.70	82
Unknown	99	2.46	80
Education level of head of household:			
0–8	187	2.80	71
9–11	290	2.67	72
12	318	2.71	78
Some college	437	2.84	88
Color of household respondent:			
White	861	2.89	84
Nonwhite	386	2.31	65

population subgroups was still below the average levels for the study population as a whole.

Discussion and conclusions

Based on the findings of this study, it appears that there were substantial deficiencies in the household reporting, even with the intensive methods employed in the survey. The accuracy of reporting was found to differ greatly among the various types of medical services as well as among various population subgroups. Charges for all services reported in the household survey were only 76 percent of the Best Data in the Baltimore area, while they were 93 percent in Washington County. As found in earlier studies (U.S. NCHS, 1977), service types that might be expected to have a greater impact on the re-

Table 12
Accuracy of household reporting of charges, by population subgroups and survey strategy (Baltimore area)
(Percent Best—All Data)

Subgroup	All strategies	Telephone	In-person	Monthly	Bimonthly
Household income:					
$0–$4,999	69	66	77	77	65
$5,000–$9,999	86	82	90	78	93
$10,000–$14,999	91	88	93	91	91
$15,000–$19,999	94	90	97	94	94
$20,000–$24,999	93	99	69	89	99
$25,000+	89	88	90	95	82
Unknown	81	83	76	76	88
Education level of head of household:					
0–8	76	74	84	87	68
9–11	78	76	80	76	79
12	91	89	93	90	92
Some college	90	92	88	89	93
Color of household respondent:					
White	89	88	90	89	89
Nonwhite	74	72	78	75	73
Total study population	86	84	88	86	86

Table 13
Accuracy of household reporting of physician outpatient visits, by population subgroups and survey strategy (Baltimore area)
(Percent Best—All Data)

Subgroup	All strategies	Telephone	In-person	Monthly	Bimonthy
Household income:					
$0–$4,999	66	62	71	71	62
$5,000–$9,999	73	74	72	69	77
$10,000–$14,999	84	76	90	84	83
$15,000–$19,999	92	97	87	92	91
$20,000–$24,999	87	93	81	84	93
$25,000+	82	86	78	89	77
Unknown	80	73	95	83	77
Education level of head of household:					
0–8	71	67	77	78	65
9–11	72	71	73	77	67
12	78	74	80	73	81
Some college	88	90	86	88	88
Color of household respondent:					
White	84	84	84	84	84
Nonwhite	65	61	70	71	61
Total study population	79	78	81	81	78

spondent, such as inpatient utilization and emergency room visits, were reported with greater accuracy than were physician outpatient visits or prescriptions. There was considerable variation in reporting accuracy by population variables in the Baltimore area, while little such variation was found in Washington County. Reporting accuracy in Baltimore was poorer, in general, for those with lower incomes, those with lower educational levels, nonwhites, those insured by Medicaid, older respondents, and those living in one-person households.

The importance of record check data for use in improving estimates derived from household survey information is supported by the findings of this study. Record information is important for correcting the general underreporting of utilization and expenditures found in the household reports. Perhaps equally important, record data is required to correct for the differential accuracy of household reporting found among service types and among population subgroups and geographic areas. Use of uncorrected household data could have consequences for policy development or planning.

The study showed no consistent differences in reporting accuracy among any of the survey strategies in Washington County for either utilization or charges reporting. In the Baltimore area, however, the in-person strategies did consistently better than the telephone procedures for both utilization and charges reporting. Earlier record check studies based on interview procedures similar to the Health Interview Survey have clearly demonstrated that the accuracy of recall of medical events decreases with time (U.S. NCHS, 1965a, 1965c, 1965d, 1967). In these studies, although the recall of hospitalization was found to remain quite good for several months, a sharp decrease in recall of events of lesser significance, such as doctors' visits, was found to occur within the first few weeks after the event. Thus, the lack of a clear-cut difference between the monthly and bimonthly groups in the current study is somewhat surprising. It appears that in the pilot MES, the use of a panel approach, calendar-diary, and cumulative Summary Report significantly reduced expected differences between the one- and two-month recall periods.

There are several factors that may account for the superiority of the in-person strategies in the Baltimore area, especially among those population subgroups with the poorest reporting. In an in-person interview, the interviewer can review written documents such as bills, receipts, appointment reminders, checkbooks, prescription labels, and the study diary; such a review cannot be done in a telephone interview. In addition, the interviewer can simultaneously interview all members of the household who are present during an in-person interview, whereas by telephone it is possible to interview only one respondent at a time. In Washington County, reporting accuracy was at a relatively high level regardless of strategy.

It is important to note that the monthly strategies suffered appreciably higher attrition over time, indicating a greater unwillingness of households to participate in a monthly survey. Thus, considering the substantial cost savings (approximately 30 percent of data collection and processing costs in this study) of a bimonthly compared with a monthly interval, the bimonthly interval appears to offer advantages.

The consistent superiority of the in-person compared to the telephone procedures in Baltimore, particularly for certain population subgroups such as those with lower incomes, those with lower educational levels, and nonwhites, indicates that allocation of additional resources for data collection and processing in certain areas or for specific population subgroups might be justified by the improved accuracy of the data obtained. Cost projections for a year-long survey, based on costs incurred in the methods study, indicate only about a 10 percent difference in costs for data collection and processing between in-person and telephone strategies.

It is important to note that the results presented here reflect a single study carried out in only two geographic areas. The results were used in planning for the National Medical Care Expenditure Survey (NMCES), a year-long panel survey during 1977 of 13,000 households that was jointly sponsored by the National Center for Health Statistics and the National Center for Health Services Research and carried out under contract with Research Triangle Institute, National Opinion Research Center, and Abt Associates. The NMCES used a recall period of 2–3 months with extensive use of telephone follow-ups. The NMCES did include a record check for a portion of the sample households, for all physician and hospital utilization. This national study represents an excellent opportunity to examine some of the findings reported here to determine if the variation of household reporting accuracy with sociodemographic characteristics and type of medical service is similar on a national basis to that found in two areas of Maryland.

In summary, the findings presented here, particularly if confirmed by NMCES, should have important implications for those planning future national and local household surveys of

health care utilization and expenditures. Record check surveys are expensive and difficult to carry out. However, some record data may be important to correct the inaccuracies in household reporting and the distortion in comparative findings that may result. Dependent on the objectives of a given survey and the characteristics of the population, it may be possible to obtain record data for selected types of utilization or only for certain subgroups of the population.

When a record check is not possible, previous studies such as this one and the NMCES may provide a basis for making judgments concerning the range within which errors may be present in the data. Particularly when sociodemographic comparisons are important, consideration of the findings of previous record check studies will be important in identifying differences in reporting accuracy among population subgroups that should be considered in interpreting survey results.

Discussion: Medical Economics Survey-Methods Study

Clyde Pope, Health Services Research Center, Kaiser Foundation Hospitals

The matter of cost-effectiveness of alternative survey strategies is an important issue. It is unfortunate that it has not received more systematic attention concomitant with the development of survey research. It is not surprising, however, that it has not. On the one hand, there has been little support from funding agencies for formal experimentation to determine either the costs or the effectiveness of alternative strategies. On the other hand, investigators have tended to be more interested in substantive than in methodological issues and in postcollection methodology of an analytical nature rather than in data collection methodology relating to cost-effectiveness. Methods of data collection have thus been relegated more to art than to science. perhaps the Yaffe and Shapiro study and these Biennial Conferences are indicators that the situation is changing, at least in health survey research.

The Yaffe and Shapiro study had a specific purpose: to determine the cost-effectiveness of different survey strategies for application to a planned national study. The task was to obtain from a panel of respondents accurate information on medical care utilization, expenditures, and third-party payments. The strategies involved monthly versus bimonthly interviews over the course of several months and in-person versus telephone contacts. Effectiveness was measured by completion of the series of interviews and by completeness and accuracy of utilization and expenditures data. The latter were independently verified through a process involving providers and third-party payers. In this context, then, cost and effectiveness are fairly clearly delimited.

However, as Yaffe et al. (1978) have shown, the cost of obtaining survey data is only one component of the cost of survey research; other components account for most of the cost. Hence, the differences among the costs of alternatives are greatly diminished when total study costs are considered. This does not minimize the importance of determining the most cost-effective alternative for obtaining data. It only points out the need to determine the cost-effectiveness of alternatives applicable to other components of the research process (e.g., capturing the data to computer).

Of course, costs are affected by the design of the study, including the type and size of the sample and its geographic dispersion, as well as rules about who are acceptable respondents and the number and type of follow-ups to be made. The effects differ depending on the type of contact: in-person, telephone, or mail. There is literature on cost differentials of different strategies (e.g., Dillman, 1978; Sudman, 1967). However, because cost factors are subject to dramatic change over sometimes very short periods of time, there is a need for continual evaluation of costs for alternative strategies. Continuously updated information about the costs of alternative strategies needs to be widely disseminated among survey researchers. Costs need not be stated solely in monetary terms; they could usefully be defined in terms of inputs required, such as type and amount of staff needed. Every federally funded survey should probably be required to provide standardized cost information to an agency designated to compile and distribute such data to the survey community. This could become increasingly valuable as more and more emphasis is placed on health services planning, which often causes planning bodies, such as Health Systems Agencies, to become involved in health surveys.

Another important cost factor is the process by which data are captured for computer analysis. More experimentation in this area is needed, especially as relevant software continues to evolve. This information also should be systematically gathered and communicated to the survey community.

Certainly, cost should not be separated from effectiveness, which, among other considerations, includes the completeness and quality

of the data provided. A major problem in evaluating the quality of much survey data is that even when there are objective data theoretically capable of verification, it is usually not considered practical to do the verification. Commonsense judgments about the quality of the data and an implicit belief that the probable magnitude of the error is tolerable for the specific purposes of the survey often suffice. And for most subjective data and measures of concepts, quality can, at best, be determined by methodological strategies ranging from studies of factors influencing item responses to formal studies of the reliability and validity of instruments. Such activities are not generally incorporated into substantively oriented surveys. Neither budgets nor the narrower interests of investigators provide for methodologically oriented analyses as part of a substantive project.

It is relevant to point out here that health maintenance organizations provide an excellent setting in which to conduct certain methodological studies relating to health surveys. One advantage is access to medical and administrative data for comparison with survey data. Another advantage is the greater ease with which a population can be followed over time, as is necessary for various methodological purposes such as evaluating instruments designed for prediction.

In thinking about surveys and the question of effectiveness, I am taken back to the discussion on "Total Survey Design" (Horvitz and Lessler, 1978) and the paper on "Standard Measures of Standard Variables" (Aday and Andersen, 1978) at the Second Biennial Conference and to the recommendations growing out of them. Among the problems identified were the lack of a common reporting vocabulary for many aspects of surveys and the lack of agreement on concepts used in health research.

Recommendations were made for supporting research that would provide for such things as (1) measuring the components of error in surveys, including the reliability and validity of instruments, sample biases, interviewer effects, and the like; (2) developing a unified total survey error model; (3) investigating alternative procedures or techniques for adjusting data for measurement errors and measurement bias; and (4) determining the feasibility of a computer retrieval system containing the cumulated measures of the components of survey errors.

Besides research into these and other methodologically oriented issues, the need was recognized for a centralized repository of such information and for a system of dissemination of that information to health survey researchers. It was suggested that one of the national health data collection agencies take on the latter responsibilities. This not only would permit a better determination of the gaps in our knowledge regarding survey methodology and cost-effectiveness of alternative strategies but would also help prevent that kind of information from becoming buried in the literature or lost because of its being scattered throughout a diversified literature.

Open discussion: Session 3

Network sampling

During the open discussion on Sirken's paper, several questions were raised regarding the specifics of the implementation of the multiplicity estimation and network sampling proposed by Sirken for studies of rare populations. In the description of his study of Vietnam veterans, in which the multiplicity estimator was used to identify eligible veterans, Rothbart noted that the probability of locating a veteran was a function of the size of the network. Network size was in turn a function of the counting rule used to determine the network scope. In general, however, multiplicity estimation with nonfamily networks must also take into account the likelihood that the respondent will know an eligible respondent, that is, how many persons in the network are eligible given certain requirements such as age and also how many eligibles are likely to be living at the time the study is done. This would be an issue, for example, in a study of patients with a disease such as cancer. In the veterans study, location was critical because the design called for locating only those veterans who lived within a well-defined area.

A further query about how a particular veteran's kinship structure was determined and the total eligibility from the network estimated led to a discussion about the principles of network sampling and the various weight factors that may be used to estimate eligibility. Sirken summarized these as

1. Sampling weight, which is the same as that used in traditional sampling;
2. Counting rule weight, which is the weight reported in terms of the number of eligible informants within the network.

Since different kinds of information are needed depending on the survey needs and the content of the study, different counting rules are required for different studies. A basic dilemma in defining the network is whether to use a relatively widely defined network that will yield a small sampling error but a higher risk of response bias due to poor reporting or to use a more narrowly defined network that may yield a better response but also a higher sampling error. Among the number of different estimators that can be used, the multiplicity estimator was most fully discussed. Its main advantage is that it does not require matching to establish individual respondents. Other estimators allow matching to eliminate case duplication. Still other estimators do not require weighting of the whole sample but only of the subsample. Emphasis was placed on consideration of the cost benefits of each type of estimator.

One attraction of multiplicity estimation that was noted was its focus on both sampling and nonsampling error. This observation prompted a further question about nonresponse effects: What information is required to adjust for nonresponse? Sirken responded that nonresponse is handled primarily by getting data from other sources, i.e., from two respondents or perhaps from other documents.

Another issue raised was the problem of asking sensitive questions. It is assumed that a person with a sensitive characteristic may be reluctant to report these about himself but may be willing to report them about a best friend or someone else. Sirken described a study by the National Institute on Drug Abuse in which they had asked about members of an individual's network who use heroin. This approach was very successful and produced estimates 50 to 100 percent higher than those obtained by traditional forms of inquiry. These estimates were also closer to those obtained from independent sources other than network reporting. The sampling errors of these estimates appeared to be about one-half as large as those occurring in other sources that were used for validation purposes. Using the network approach in this particular study *appeared* to correct for both sampling and nonsampling error.

For locating missing respondents, network sampling may be useful in two respects. First, individuals who are missed because of institutionalization may be described by members of their network. Second, nonresponse about attributes, i.e., underreporting, may be corrected because people are more likely to talk about others' experiences than they are about their own.

Axelrod questioned the implications of this technique for invasion of privacy and wondered whether there might be any special requirements for informed consent in studies using this technique.

Sirken described a survey being planned at NCHS in which network sampling will be used to locate cancer cases that are estimated to occur in about 1 percent of the general population. The procedure being developed explicitly takes into account the possibility that the patient may not wish to be identified as having cancer.

To accomplish this, all identified cases will be mixed with a probability sample drawn from the general population. Thus, neither the interviewer assigned a particular case nor the case will be aware at the time of the interview that the case has been identified by a person in his/her network as a cancer patient. When the interviewer begins a series of questions designed to determine whether the respondent is a cancer patient, patients may opt not to disclose their illness to the interviewer, and they will be treated accordingly. Such patients will be eliminated from the study, although their prior identification will permit seeding into the sample a higher potential yield of eligible respondents and also allow estimation of the extent of response bias due to refusal to participate by failing to identify oneself as a patient. Obviously this approach increases the cost of the survey, since eligible respondents are lost and two forms of screening are required, the first using the network to identify the respondents and the second using the interviewer to identify the patients in the sample. However, Sirken noted that the cost was necessary to protect the privacy of the patient.

In response to a general query about other uses of the multiplicity estimation technique, Groves described a University of Michigan study in which the technique did not work successfully. In that study, the objective was to identify a Chicano population. A sample of known Chicanos was selected, and they were interviewed and asked to identify and provide addresses for others in their families including parents, siblings, and children 18 and over living outside the household. The individuals in the sampled networks simply refused to disclose the requested address information.

In a second study described by Groves, network sampling is being used to identify black families with three generations surviving. In this study, the primary objective is enumerating the families rather than locating all members of a family network. Addresses are being requested only for one parent or one child in order to validate the presence of three generations. This design is expected to be more successful than the Latino study, in part because the data requested are less extensive and in part because the location of relatives is considered to be less sensitive for blacks than for Chicanos, for whom immigration status may be an issue.

Health in Detroit Study methodology

Two major issues were raised by the Verbrugge paper: the methods of respondent training and the reliability in interpretation by respondents and by the investigator of certain concepts that may be ambiguous. These issues were raised in response to a point made by Verbrugge that considerable effort had been made to train the respondent to differentiate between symptoms of a condition and the condition itself. This question related in large measure to the methods used to train the respondents in diary procedures.

Standardized procedures were used by Verbrugge and her staff at Michigan to describe the study and its procedures. The instruction in diary procedures was organized around recording health events that had occurred within the two weeks prior to the interview in a sample diary. In addition, initial entries were made in the diary for the date of the interview. Both these sample diaries were then left with the respondent as a model. Respondents were strongly encouraged to call the Survey Research Center if they had any specific questions about the diary procedures, such as how to record specific events and where in the diary format to record them. Questions about what to record were treated neutrally by the project staff with the respondent being encouraged to report any relevant events.

Numerous calls were received in the early phases of the study from the 700 respondents. Records indicated that as many as 15 to 20 calls per week were received, mostly concerning the method of coding symptoms of mental illness. In fact, the problem of coding symptoms and differentiating them from conditions seemed to have been a major issue in the Verbrugge project. Respondent training attempted to focus on the distinction between the two concepts and to

clarify the issue for the respondents in terms of getting them to think about symptoms *of* a particular condition. Moreover, the problems in making this distinction were not confined to respondents. Verbrugge reported that coders also experienced considerable difficulty with this distinction. It is clear that if this type of data is to be obtained by nonmedical personnel (either respondent or coder), detailed and highly specific instructions will have to be developed. Even then, the distinction in particular cases may not be clear-cut.

Health Insurance Study data evaluation

With respect to the Marquis presentation, questions seemed to focus on incentives and on estimating nonresponse by both the experimental and the control cases in that study.

In response to a question about incentives, Marquis noted that the only incentive used was payment to either the respondent or the carrier who provided the claims data.

In general, the results of the study indicated that there was substantial underestimation of health care service utilization when claims data were used, and much speculation occurred on why this might be the case. One likely source for this underestimation was the delay in claims submission. The real difficulty with this study, however, was in estimating the percentage of respondents who had used health care services but had neither filed claims nor completed diaries. In the control situation, this number is very hard to estimate because these were only claims data, although it was thought to be between one-third and one-half of the respondents. In the experimental condition, efforts were being made to estimate this percentage using provider data as a substitute for diaries. However, as noted by both Marquis and Sirken, this method of estimation is only reliable if one can make the assumption that the two sources of data are, in fact, independent. At the present time, it appears that the investigators at Rand are not satisfied that this assumption had been demonstrated, although efforts are still being made to determine whether this might be the case.

In the discussion of the Marquis results, it was pointed out that the finding that diaries made persons more aware of their health might be a basis for studies of the use of the diary as a means of public health education and that it will be interesting to compare results of several diary studies on these dimensions.

In summarizing this discussion, Sudman noted that in studies dealing with health, it is possible to get people to keep diaries for prolonged periods of time. Experience at the University of Illinois, for example, shows that people will keep diaries for periods of up to three months, and there is no indication that three months is a maximum limit. The main difficulty seems to be obtaining initial cooperation. Once respondents are recruited into a diary study, there seems to be very little attrition. This latter observation was also made by Verbrugge. Finally, Verbrugge, Marquis, and Sudman have all noted the beneficial effect of compensation on cooperation.

Medical Economics Survey-Methods Study

Open discussion of the Yaffe and Shapiro paper focused primarily on the possible causes of the noticeable response variation found in the data obtained from two areas around Baltimore. The study described by Yaffe was a pretest for the National Medical Care Expenditure Survey conducted by NORC and RTI and was conducted in Baltimore and Washington County, a rural area in western Maryland. There were major differences noted in the characteristics of the populations in the two regions that appear to have affected response. The main difference seemed to be the percentage of rural respondents in Washington County and the greater percentage of black respondents in Baltimore. Although the socioeconomic status in Washington County was somewhat higher, the ranges on this variable overlapped substantially in the two samples. However, as Yaffe noted, there was considerable difference in the sociological characteristics of the two populations.

The differences observed were also subsequently found in the National Medical Care Expenditure Survey and should provide further clarification of the interaction between socioeconomic status and participation in such studies.

Besides socioeconomic status, another major difference was the relative experience of the interviewing teams in the two regions. All interviewers used in the study were employed and trained by Westat, and standard training procedures were used. However, the staff in Washington County was more experienced and had worked on other studies for Johns Hopkins. The staff in Baltimore was less experienced and had been recruited specifically for this study.

A question was raised about the various conditions used in this experiment and whether there were consistent differences between telephone and personal interview responses. Yaffe indicated that no significant differences were found that could be attributed to the sociodemographic differences in the two samples. It was

noted that interviewer experience seemed to eliminate whatever variation might exist, since no differences between personal and telephone interviews were noted in Washington County.

Where more than one type of data collection was being used, the order in which methods were introduced seemed to be a factor. The Yaffe study found the least cooperation in situations where initial telephone contacts were followed by personal interviews. Respondents wanted to know why a personal interview was necessary. In the reverse situation and when the interval was longer, results seemed somewhat better.

Pope described a study carried out at the Kaiser Plan in San Francisco in which an initial mailed questionnaire was sent to respondents (selected from the files at Kaiser) and then telephone contacts were made as follow-ups. This was a study of the incidence of influenza, and weekly telephone contacts were made during the flu season. Pope described very high respondent commitment to the study and indicated that frequently respondents would call the interviewer and arrange for alternative interview dates if they were not going to be at home at the prearranged time.

A further question, based on the Yaffe experience, was raised about the effect of socioeconomic status on cooperation rates. It was suggested that with such groups, obtaining data directly from providers might be better.

In response to this comment, Sudman noted that one of the major points to be made from the papers presented by both Yaffe and Verbrugge was that diaries seem to work particularly well with the less educated. It would appear that "if they can be found, they will participate."

Horvitz summarized much of this discussion by making several general points based on the studies on health care expenditures being done by Research Triangle Institute. First, he noted that the alternative strategies of obtaining provider data and respondent data seemed to be complementary, and the relative cost-effectiveness of using either strategy or some combination of them had yet to be established. In the National Medical Care Expenditure Survey, preliminary work concerned whether providers should be contacted and a check made on behavior reported in the diaries. The data indicate that reports from the aged and nonaged poor may require some validation. Even among these groups, however, the extent of need for verification depends on what is being studied and the level of accuracy required. These factors need to be carefully considered, and the issue seems far from settled.

Horvitz suggested that the next step may be to differentiate between these alternative methods of data collection. The marginal costs of validation in terms of improved data quality could then be calculated. Decent cost data appear to be coming from studies like that described by Verbrugge. Moreover, in light of Marquis' findings it remains a moot point whether the provider data will be better than the respondent data. Although the dual-record approach may be useful, the issue of independence of the two sources is still unsettled, and the validity of the assumption of independence is questionable.

Recommendations

The following recommendations emerged from the discussion in this session:

1. Multiplicity estimation and network sampling as techniques need further investigation regarding their applicability to health survey research.
2. Multiplicity estimation appears to be particularly useful for locating rare respondents, and its applicability in this regard should be further evaluated.
3. Information about the accuracy of reporting is another particularly critical issue, and further research into this technique must be carried out.
4. Another issue related to the accuracy of information obtained by these techniques is the counting rule used to define the network for sampling purposes. The tradeoffs between widely extending the network to reduce sampling error versus tightening the network to reduce response effects need further study.
5. Confidentiality and the implications of network sampling and multiplicity estimation for invasion of privacy are critical issues that will have to be addressed if this approach is widely used.
6. Diaries have been shown to be an effective procedure for obtaining health information about less obvious health events. Because health is such an important topic to respondents, cooperation rates with health diaries have been high, with little loss of respondents over time. We recommend that diary procedures be utilized in continuing studies of health behavior and expenditures such as the Health Interview Survey (HIS) and the National Medical Care Utilization and Expenditure Survey (NMCUES).
7. Further research on diary format and whether diaries should be used as a primary data collection procedure or as a memory aid is necessary.

8. With the increased use of diaries, more needs to be known about the conditioning effects of diary reporting. As a corollary, whether the diary is a viable means of public health education in the field of health behavior should also be explored and experiments using it in this way should be developed.
9. Research is needed on the cost of improving diary cooperation versus the impact of noncooperation on total survey error.
10. The issue of the quality of health data obtained from telephone versus face-to-face interviews remains unresolved. More research is needed on the relative costs, benefits, and alternative mixtures of these two procedures.
11. A careful investigation is needed to determine whether studies that combine telephone and personal interviewing procedures should use separate cadres of telephone and personal interviewers and under what conditions a single instrument is adaptable to both techniques.
12. We recommend further study of the interaction of effects due to procedure (telephone, face-to-face, diary), population characteristics (age, socioeconomic status, rural/urban residence), and design (cross section versus longitudinal). It seems likely that there will be optimal combinations of these three critical characteristics that can be recommended for future work.

SESSION 4: Methodological issues in developing standardized measurement of long-term behavior

Chair: Jack Elinson, School of Public Health, Columbia University

Recorder: Naomi D. Rothwell, Bureau of the Census

What brief psychiatric screening scales measure*

Bruce P. Dohrenwend, Department of Psychiatry, Columbia University

Barbara Snell Dohrenwend, School of Public Health, Columbia University

Lois Oksenberg, National Center for Health Services Research

Diana Cook, Department of Psychiatry, Columbia University

Patrick E. Shrout, School of Public Health, Columbia University

In 1957 Macmillan published the Health Opinion Survey, a 24-item psychiatric symptom scale, and in 1962 Langner published a similar 22-item psychiatric symptom scale. Since 1964 at least 14 epidemiologic studies have used one or the other of these instruments, or a 20-item version of the Health Opinion Survey (Leighton et al., 1963:208–14), in community surveys. In reporting their results, investigators have described these measures in terms of such diverse constructs as "mental health," "mental illness," "psychiatric disorder," "emotional adjustment," "emotional disturbance," "symptoms of stress," and "psychophysiological symptoms" (Seiler, 1973:257). Given this state of conceptual anarchy, we must ask what these instruments measure.

Implicit in the labels used to describe them are two hypotheses. One is that these scales measure psychopathology—mental illness or psychiatric disorder. This hypothesis implies that persons identified as having the characteristic measured are in need of treatment. It is, moreover, consistent with Macmillan's and Langner's intentions that their scales be used to screen for psychopathology in the community. The alternative hypothesis is that these scales measure a response to circumstances or events—symptoms of stress, emotional disturbance—which does not necessarily or even usually imply pathology and need for treatment. The study that we will report started with the first hypothesis and ended with the second.

The purpose of this study was to build a better instrument to use in screening for psychopathology in the community. To clarify this purpose, let us review the procedures used in constructing the Health Opinion Survey and Langner's 22-item scale.

These instruments were developed in the course of two epidemiologic studies: the Stirling County Study (Leighton et al., 1963), which involved a sample of about 1,000 residents in a rural Canadian county, and the Midtown Study (Srole et al., 1962), which surveyed a sample of about 1,600 residents of a section of Manhattan. Consistent with such psychiatric orientations as those of Adolph Meyer (Lief, 1948) and Karl Menninger and his colleagues (Menninger, 1963), investigators in both studies treated all disorders as lying on a continuum. Responses to fairly extensive questionnaires administered in structured interviews were used to identify instances of disability. The questionnaires relied heavily on fixed-alternative symptom questions from sources such as the Neuropsychiatric Screening Adjunct developed by the Army Research Branch in World War II, the Cornell Medical Index, and the MMPI, as well as on similar questions created by the study staffs. The interviewers in these studies were not psychiatrists or clinical psychologists, but in both studies the procedure for identifying disabled individuals involved judgments by study psychiatrists based in large part on their review of the written records of questionnaire responses. Thus, psychiatrists in the Midtown study judged the extent to which each respondent appeared to be "impaired" (Srole et al., 1962:399). In the Stirling County study, psychiatrists made similar ratings and also judged the likelihood that the respondent in question would actually be a

* This research has been supported by Research Grant MH 10328 and Research Scientist Award KO5 MH 14663 from the National Institute of Mental Health, U.S. Public Health Service, and by the Foundations' Fund for Research in Psychiatry.

We would like to thank several of our colleagues for their help: DeWitt Crandell, Gladys Egri, and Frederick Mendelsohn for providing a priori clinical groupings of the symptoms; Sheppard Kellam for pointing out the plausibility of what turned out to be the rejected hypothesis that we were measuring neurosis; H.B.M. Murphy for sending several papers expressing his own skepticism about the relation of the screening measures to clinical psychological disorders; and Donald F. Klein for suggesting the relevance of Jerome K. Frank's concept of demoralization for our findings. We are also grateful to Joseph L. Fleiss for statistical advice, and to Alexander R. Askenasy and Thomas J. Yager for helpful comments.

"case" if given a full diagnostic evaluation (Leighton et al., 1963:121). Although psychiatrists in both studies made judgments about the presence or absence of various types of symptomatology or symptom "patterns," the "caseness" or "impairment" ratings were the primary measures.

In order to construct brief screening scales, comparisons were made between responses to symptom questions given by psychiatric patients and nonpatients (Langner, 1962; Macmillan, 1957). Comparisons for the Midtown study were between a group of psychiatric patients diagnosed as neurotic or psychotic or in remission and a group of individuals judged "well" by a study psychiatrist on the basis of a half-hour interview. Langner's brief screening scale consisted of the 22 items that discriminated most sharply between the patients and the "well" nonpatients. These items also predicted with fairly good accuracy the impairment ratings made by study psychiatrists on the basis of the entire questionnaire (Langner, 1962).

In the Stirling County study, comparisons of responses to an initial, large number of symptom questions were made between a group of patients diagnosed as neurotic and samples of community respondents (Macmillan, 1957). The 24 symptom items in the final questionnaire were selected from the larger pool primarily on the basis of their ability to discriminate between the patients and the community respondents. As in the Midtown study, scores on a scale including 20 of those 24 questions correlated substantially with the psychiatric evaluations of "caseness" and "impairment" made in that research (Leighton et al., 1963:208–14).

From this background it is apparent that both Macmillan's Health Opinion Survey and Langner's 22-item scale were constructed on the assumption that psychopathology is unidimensional. The assumption of multidimensionality is, however, far more widely accepted among experienced clinicians, as indicated, for example, in the World Health Organization's *International Classification of Diseases* and the American Psychiatric Association's *Diagnostic and Statistical Manual.* Our aim, therefore, was to construct an instrument composed of items selected to measure conceptually separate dimensions of psychopathology such as depression and anxiety. Such an instrument would, we reasoned, have greater construct validity than an instrument constructed from items chosen for their ability to discriminate psychiatric patients from nonpatients with little regard to their content. Moreover, it would refine the screening process by permitting the selection of persons with specific types of symptoms.

Method

Starting with a pool of fixed-alternative response questions selected from previous screening scales, our approach involved first a conceptualization of dimensions of psychopathology that they might represent and the construction of scales intended to measure individual differences along each of the underlying dimensions. The properties of these scales, their relations to each other and to measures of functioning, and the generality of these properties and relations across different populations were then examined.

The interview. Reports of symptoms and quality of functioning were obtained in response to a structured interview schedule (SIS) modeled on the questionnaires used in the Midtown (Srole et al., 1962) and Stirling County (Leighton et al., 1963) studies and in our own earlier work (B.P. Dohrenwend and B.S. Dohrenwend, 1969). Approximately 40 symptom questions were drawn, as in earlier research, principally from the Neuropsychiatric Screening Adjunct and from the MMPI.

Most of the symptom questions asked either whether the respondent had had the symptom or how frequently the respondent had experienced the symptom. The first had dichotomous response categories, either yes/no or true/false, and the latter had response categories of often, sometimes, and never. Subjects giving positive responses were further questioned about whether the problem was very serious, somewhat serious, not serious at all, or, where the temporal reference of the question was unclear, no longer a problem. This last bit of information was used to ensure that all positive responses referred to current problems rather than to problems in the past.

In addition, interviewers probed positive responses to 11 of these questions, including many from the Army's Neuropsychiatric Screening Adjunct, which our previous work had suggested might, in a community population of all ages, reflect poor physical health rather than psychopathology. For example, when a respondent reported a symptom such as "heart beating hard," the interviewer asked what the respondent thought was the cause of the symptom; if the respondent thought it was a physical illness such as heart disease, the respondent was asked how he or she learned that this was the case and especially whether the source of the information was a medical doctor.

Several ratings of respondent behavior during the interview were made by the interviewer. For example, one rating concerned the respondent's

tension level at the start of the interview, with categories of nervous, fidgety, sporadic nervousness, and mostly relaxed. Where relevant, such ratings were grouped with symptom questions in particular content areas.

Additional questions concerned the quality of functioning in a number of areas of life and covered aspects of work performance, job morale, marriage, social relations, housework, parenthood, and use of leisure (B.S. Dohrenwend, B.P. Dohrenwend, and Cook, 1973). Many of the questions were adapted from those used by other investigators (Bradburn and Caplovitz, 1965; Gurin, Veroff, and Feld, 1960).

The interviewers. The interviewers were 15 psychiatrists, who received several days of instruction on the use of the SIS. Although psychiatric training was irrelevant to the mechanics of administering and scoring most of the SIS, it did facilitate the earlier-described probing for causes of symptoms. In addition, the interviewers used their psychiatric training to make a Midtown "impairment" rating as well as a Stirling County "caseness" rating for each respondent on the basis of information gleaned from the respondent's answers to their questions and from their observations of the respondent during the interview. The interviewer also provided a tentative diagnosis for each subject. Except for those who were institutionalized, most respondents were interviewed in their homes. The great majority of interviews were tape-recorded; this permitted us to check the reliabilities of the clinical ratings, which proved quite satisfactory for caseness and impairment (B.P. Dohrenwend, 1974:279). Interviews averaged about 90 minutes in length.

Respondents. The subjects who provided the interview data on which the present analyses are based were 227 adults, aged 21–64. Of these, 124 constituted a community sample, while the remaining 103 were a sample of psychiatric patients.

The 124 community sample respondents were residents of the section of the Borough of Manhattan in New York City known as Washington Heights. They were drawn by full probability sampling procedures in such a way that four ethnic groups—black, Puerto Rican, Irish, and Jewish—were represented in roughly equal proportions. A fifth group of white Protestants of European ancestry made up a smaller proportion of the sample owing to their rarity in Washington Heights. Within each of these ethnic groups, an attempt was made to balance educational levels in order to permit subsequent analyses of the data with class and ethnicity unconfounded. The educational level of the husband was used for married women, since this was considered a better indicator of social class for them than their own educational level. The 124 community respondents were 66 percent of the community residents from whom interviews were sought. More detailed descriptions of these subjects and the procedures for drawing them have been presented elsewhere (B.P. Dohrenwend et al., 1970).

The remaining 103 respondents were psychiatric patients or prisoners, who, like the community sample respondents, were drawn to include blacks, Puerto Ricans, and other whites from contrasting social class backgrounds (B.S. Dohrenwend et al., 1973). Of these respondents, 60 were outpatients from various psychiatric clinics in or adjacent to Washington Heights, 32 were patients admitted to either of two New York City mental hospitals in the week prior to the interview, and 11 were prisoners convicted of crimes in New York City. The 92 patients were 76 percent of all patients from whom interviews were sought, and the 11 prisoners were 100 percent of those from whom interviews were sought.

All of the patients and prisoners received psychiatric diagnoses from the psychiatrists who interviewed them. Although independent reviews of the tape recordings of the interviews revealed that the interviewers' diagnoses were only moderately reliable, with weighted kappa (Fleiss et al., 1972) of .50, these diagnostic judgments by psychiatrists do suggest that a wide variety of disorders was present in the sample. According to the interviewers' diagnoses, 40 percent of the patients and prisoners were schizophrenic, 5 percent exhibited other functional psychoses, 21 percent were neurotic, 14 percent were sociopathic, 13 percent exhibited other types of personality disorder, and 7 percent received other diagnoses. In addition, there were formal psychiatric evaluations available in case records for 45 of the outpatients. Although probably very unreliable and differing somewhat in distribution from those provided by the psychiatrist interviewer for the total sample of patients and prisoners, these clinic diagnoses also included a wide range of disorders.

In order to interview Puerto Rican subjects whose facility with English was limited, the SIS was translated into Spanish. Thus, 57 percent of the Puerto Rican community respondents and 39 percent of the Puerto Rican patients were interviewed in Spanish by one of the Spanish-speaking interviewers, while others were interviewed in English.

Results

Because our sample was stratified by education, ethnicity, and sex, it is not appropriate to treat the total sample as representative of any population. In many of our analyses, we will actually be interested in within-strata statistics, and thus the sample definition will pose no problem. For other analyses, however, we will be interested in general statistics taken across the strata. Most of these general results involve correlations; these are computed pooling within-strata analyses, thus removing the possible bias introduced by between-strata variation.

A priori classification of symptoms. The first step in the analysis of the meaning of the symptom reports in different groups involved a conceptualization of underlying dimensions of psychological disorder to which each symptom item might be related. Accordingly, three board-certified psychiatrists on the study staff began by assigning each SIS symptom to one of a number of symptom types judged by them to be represented by these items. After some reconceptualization based on their initial assignment attempts, the three psychiatrists assigned through consensus almost all of the symptom questions derived from or similar in content to those in screening scales into five nonoverlapping groups.

Internal consistency of the classifications. The five item groups were then refined on two different bases. First, responses to interviewer probes concerning causes of symptoms were used to identify questions for which reported symptoms were likely to reflect physical illness rather than psychiatric disorder. For 5 of the 11 questions so probed, the majority of reported symptoms were said by the respondents to be due to the presence of physical illness confirmed by a physician. On this basis the five questions were judged inappropriate for use in measures of psychopathology and were discarded (Crandell, Cook, and Dohrenwend, 1971).

The first step of statistical analysis involved calculating the reliabilities of the five symptom scales, as indicated by coefficient alpha, a measure of the internal consistency of the scale (B.P. Dohrenwend and Crandell, 1970). Table 1 gives coefficient alphas for the final versions of each of the five scales (incorporating the revisions discussed below) for the two major subgroups of interest, the community respondents and the patients and prisoners.

Correlations were also calculated between each item in each of the scales and scores based on the remainder of the scale. These item-whole correlations were used to identify items that might be deleted. Thus, three items with particularly low item-whole correlations within some of the demographic subgroups were removed.

Assessment of the applicability of the scales in different subgroups. One indication of the inadequacy of a scale as a measure of psychopathology in a particular subgroup would be evidence of the scale's lack of internal consistency for that subgroup. Accordingly, the next stage in the analysis involved assessment of the reliabilities of the five scales for the various subgroups yielded by the stratifications according to sex, ethnicity, and education among community respondents and among patients and prisoners.

Table 2 gives the scale internal consistency reliabilities, calculated separately for the subgroups resulting from the stratifications. In general, these scale reliabilities remain reasonably high for the various subgroups of patients and prisoners and of community respondents. Clear exceptions occur at the highest educational level (college-educated for 16+ years of education) among community respondents for all but the Sadness scale and with black patients and prisoners for Perceived Physical Health. In all of these instances, the low alphas were accompanied by very low frequencies of reported symptoms, a factor that by itself can at least partially account on statistical grounds for the accompanying low levels of reliability. Thus, the low alphas for these instances cannot be interpreted as fully testing the applicability of the scales in question to these

Table 1
Coefficient alphas of SIS symptom scales for community sample and patient and prisoner sample

Scale	Sample	
	Community	Patient and prisoner
Perceived Physical Health (3)[a]	.63	.58
Psychophysiological Symptoms (7)	.62	.68
Anxiety (10)	.70	.78
Sadness (6)	.75	.64
Enervation (4)	.62	.51

[a] Figures in parentheses are number of items in scale.

NOTE: The number of respondents who completed all items on these scales ranged from 112 to 124 for the community sample with a median of 124; for patients and prisoners the range was 57 to 101 with a median of 97. The lowest frequencies were on the Anxiety scale because of missing data on one of the items, "Taking Sleeping Pills to Calm Nerves." This item was recoded, and those respondents whose answers were incomplete according to our coding standards were excluded from this scale.

particular subgroups. The Enervation scale provides the only instance of low alphas for some groups in which the accompanying frequencies of reported symptoms were at least moderate (i.e., each reported on the average by at least 15 percent of the sample). These occurred among patients and prisoners with 12–15 years of education, among black patients and prisoners, and among black community respondents. The results for these three subgroups provide the only clear suggestion that any of the scales fails to reflect a meaningful dimension for any of the subgroups.

Comparisons of patient and community respondents. It is reasonable to expect, as did the Midtown and Stirling County researchers, that psychiatric patients will, on the average, have higher scores than nonpatients on scales intended to measure psychopathology. This expectation was also realized with this data set. A multivariate analysis of variance was performed on the five symptom scales using sex, ethnicity, education, and patient status as factors. The multivariate F for patient status controlling for the other factors was significant beyond the .001 level (Wilks criterion $F = 13.69$, d.f. = 5,184). Univariate tests revealed that the differences were significant for all scales except Perceived Physical Health, which only showed a trend. The general difference was consistent in the ethnic, sex, and education groups; interactions of those demographic variables with patient status were tested and found to be insignificant. The differences are illustrated in Table 3, which shows the means and standard deviations for the five scales in the community and patient-prisoner samples.

Symptom scales and measures of functioning. On the assumption that psychiatric disorder is reflected in impaired ability to function in one or more social roles, one would expect scores on symptom scales intended to measure psychopathology to be correlated with scores on measures of quality of functioning. Table 4 gives the correlations between scores on the five symptom scales and scores on two measures of quality of functioning constructed from SIS questions covering social roles in a number of areas of life.

Table 2
Coefficient alphas of six SIS symptom scales for subgroups of community sample and patient and prisoner sample resulting from stratification according to sex, ethnicity, and educational level

Scale	Stratification basis: Years of education[a]			Ethnicity			Sex	
	<12	12–15	16+[b]	Black	Puerto Rican	Other white[c]	Male	Female
	Community sample (N = 124)[d]							
Perceived Physical Health	.65	.59	−.07	.65	.76	.42	.69	.60
Psychophysiological Symptoms	.60	.57	−.08	.60	.51	.68	.72	.59
Anxiety	.72	.71	.41	.50	.76	.70	.68	.71
Sadness	.73	.67	.80	.57	.70	.82	.82	.70
Enervation	.64	.66	.29	.34	.86	.60	.62	.62
N	(52)	(41)	(31)	(35)	(21)	(68)	(47)	(77)
	Patient and prisoner sample (N = 103)[d]							
Perceived Physical Health	.46	.66	—	.15	.72	.48	.55	.61
Psychophysiological Symptoms	.70	.64	—	.46	.73	.69	.46	.76
Anxiety	.73	.72	—	.70	.74	.80	.77	.79
Sadness	.65	.49	—	.54	.58	.65	.66	.63
Enervation	.56	.28	—	.25	.79	.40	.45	.56
N	(63)	(31)	(6)	(27)	(17)	(59)	(46)	(57)

[a] In the community sample, respondents were classified by educational level of the head of the household.

[b] Coefficient alphas were not computed for patients in this education category because of insufficient numbers.

[c] "Other white" includes Jewish, Irish, and Old American Protestant respondents.

[d] Many alphas, especially among the patients, are based on slightly fewer cases owing to missing data.

Table 3
Means and standard deviations for five SIS symptom scales[a] in community sample and patient and prisoner sample

Scale	Community sample (N = 124)		Patient and prisoner sample (N = 100)	
	Mean	S.D.	Mean	S.D.
Perceived Physical Health	.172	.287	.270	.317
Psychophysiological Symptoms	.118	.175	.241	.239
Anxiety	.226	.226	.493	.262
Sadness	.185	.255	.422	.283
Enervation	.185	.259	.314	.298

[a] Scale scores here are the sum of items endorsed divided by number of items in the scale.

The two role-functioning measures discriminated between, and were highly reliable for, patients and prisoners as a whole and community respondents as a whole (B.S. Dohrenwend et al., 1973). The Job Stability scale, relevant only to respondents in the job market, was based on reports of current employment status and the length of time unemployed in the last year. The Marriage scale, relevant only to currently married respondents, included eight questions covering feelings toward the spouse, satisfaction with the marriage, behavior of the respondent and the spouse toward each other, and stability of the marriage.

Table 4 shows that one of the two role-functioning measures, Job Stability, did not relate as strongly to the symptom measures in the patient-prisoner sample as in the community sample. Nevertheless, the results for the community respondents and for patients and prisoners are generally consistent. With the exception of the Job Stability scale for patients and prisoners, the correlations between the role-functioning measures and the remaining five symptom scales support the claim that the symptom scales measure different levels of psychopathology.

Table 4
Correlations between SIS symptom scales and role functioning scales for community sample and patient and prisoner sample

Scale	Role-functioning scale	
	Job Stability	Marriage
	Community sample	
Perceived Physical Health	.22*	.49**
Psychophysiological Symptoms	.25**	.28**
Anxiety	.31**	.37**
Sadness	.42**	.39**
Enervation	.20*	.35*
N	(98)	(77)
	Patient and prisoner sample	
Perceived Physical Health	.14	.40*
Psychophysiological Symptoms	.09	.35*
Anxiety	.17	.25
Sadness	.32**	.28
Enervation	.25*	.51**
N	(67)	(32)

* $p < .05$.

** $p < .01$.

Interrelations of the scales. The evidence is mainly consistent with the hypothesis that the five symptom scales, with the possible exceptions noted earlier, are generally applicable measures of psychopathology. An important question that remains is whether the five scales of Anxiety, Sadness, Enervation, Psychophysiological Symptoms, and Perceived Physical Health measure the dimensions of psychopathology that they were intended to measure by the clinicians who named them and adjudicated their content.

Taking into consideration that the internal consistencies of two scales limit the size of the correlation that can be obtained between them, one can see in Table 5 that the correlations among these five scales are extremely high in the community sample. Sadness and Anxiety, for example, are correlated .63 in the sample. Since both scales have alphas of about .70, one would expect their correlation to be about .70 if they were measuring the same thing. Actually, all five scales are intercorrelated about as strongly as their reliabilities permit in the community sample and, for several scales, in the patient-prisoner sample as well. Is it possible that the five symptom scales measure one common dimension?

The results of principal components analyses, reported in Table 6, suggest that this is the case. The principal components analysis for community respondents yielded one common factor underlying all five scales. For these re-

Table 5
Correlations among SIS symptom scales for community sample and patient and prisoner sample

Scale	Perceived Physical Health	Psychophysiological Symptoms	Anxiety	Sadness	Enervation
	Community sample (N = 124)				
Perceived Physical Health	(.63)[a]				
Psychophysiological Symptoms	.51	(.62)			
Anxiety	.51	.46	(.70)		
Sadness	.61	.46	.63	(.75)	
Enervation	.54	.45	.42	.53	(.62)
	Patient and prisoner sample (N = 96)				
Perceived Physical Health	(.58)				
Psychophysiological Symptoms	.38	(.68)			
Anxiety	.40	.41	(.78)		
Sadness	.14	.09	.31	(.64)	
Enervation	.51	.51	.50	.33	(.51)

[a] Figures in parentheses are the coefficient alphas for the scales.

NOTE: Based on pooled within-strata matrices; the effects due to ethnicity, education, and sex have been removed.

Table 6
Principal components analyses of five SIS symptom scales for community sample and patient and prisoner sample

Scale	Loadings on first two unrotated components			
	Community sample		Patient and prisoner sample	
	1	2	1	2
Perceived Physical Health	.81	−.06	.71	.27
Psychophysiological Symptoms	.72	−.43	.71	.39
Anxiety	.78	.47	.76	−.11
Sadness	.84	.30	.46	−.84
Enervation	.75	.33	.84	.00
Latent root	3.05	.61	2.50	.95
Percent of total variance	61%	12%	50%	19%

NOTE: Based on correlation matrices from which the effects of stratification levels have been removed (Table 5).

spondents, the first unrotated factor accounted for considerably over 50 percent of the total variance, and all five scales loaded extremely high on it. The second unrotated factor cannot be considered important, since its latent root is considerably less than one (Nunnally, 1967). Thus, for community respondents there is no evidence that any of the five scales measures a dimension distinct from that measured by any other. The situation is similar for the patients and prisoners, although here there is a possibility that the Sadness scale defines a second dimension.

What these findings on the interrelations of the five scales shown in Table 6 mean, for all practical purposes, is that we could substitute our measure of Perceived Physical Health for our measure of Anxiety, our measure of Anxiety for our measure of Sadness, our measure of Sadness for our measure of Enervation, and our measure of Enervation for our measure of Psychophysiological Symptoms with little effect on the results that we would obtain. Although they have been given different names and were intended to measure different phenomena, these five scales reflect a single underlying dimension. Does this result indicate a weakness of self-reports of symptoms as measures of psychopathology?

One possibility that would suggest such an intrinsic weakness is that scale scores were strongly influenced by extraneous factors that led respondents indiscriminately to report more or fewer symptoms. Individual variations in acquiescence, the tendency to agree or to give positive responses to questions regardless of their content, could be such a factor. Fortunately, we had scores on a 22-item measure of acquiescence available from a previous study for 48 of the community respondents. Illustrations of the types of items that were keyed "true" or "false" on the basis of a flip of a coin and included among the 22 are the following:

I am sometimes impatient with others. (True)

People never bother me just by being around. (False)

I am more critical of myself than other people are of me. (True)

When things go wrong it is rarely my own fault. (False)

When I disapprove of my friends' behavior I let them know it. (True)

It upsets me to think some thoughtless word or crack of mine might hurt someone's feelings. (False)

As it turned out, this measure of acquiescence showed only low correlations with symptom scale scores; these ranged from .08 for Sadness to .27 for Psychophysiological Symptoms, with a median of .13. Although this result does not rule out situation-specific acquiescence as a factor contributing to the lack of discrimination among the scales, it seems unlikely that such a response bias could account by itself for the lack of discrimination among the scales.

Another possible extraneous biasing factor is individual variation in the tendency to present a positive self-description owing either to differences in motivation or to differences in perceptions of symptom desirability or undesirability. We are investigating the possible influence of those factors in a second study. The results that we have obtained indicate that such possible biases cannot account for the present results.

Post hoc analysis and discussion

In order to investigate further the nature of the dimension being measured, we have combined the five SIS scales into one composite measure by standardizing each scale by its number of items and summing the scales as though they were themselves items. The resulting SIS composite had a reliability of .86 in both the community sample and the patient and prisoner sample and was also reliable (alpha greater than .50) in all population segments except college-educated community respondents. The very low frequency of symptomatic responses in this last group may account for its .41 reliability.

Relation of SIS to Macmillan's HOS and Langner's 22-item scale. The items that form the five scales in our composite overlap either explicitly or in their general character with the screening instruments developed by researchers in the Midtown and Stirling County studies. The composite shares 16 items with the 22-item Langner scale based on the Midtown study and eight items with the 20-item FLS and BHS scale used in the Stirling County study (Leighton et al., 1963:440-41). Statistically, it is indistinguishable from Langner's scale, with which we are able to compare it directly. The correlation between them is .91.

Do these scales measure neurosis? So far we have shown that the SIS composite is measuring a single dimension and that it is apparently the same dimension as measured by Langner's 22-item scale and probably by Macmillan's Health Opinion Survey. If psychopathology is multidimensional, we clearly have not succeeded in measuring the various dimensions. What we may have done, however, is to measure one dimension of psychopathology. Examination of the content of the SIS composite, as well as the items in Langner's and Macmillan's scales, at first seemed to suggest that this dimension might be neurosis. If so, we reasoned further, an unexpected difference between results for the community and the patient-prisoner samples might be explained by this conception of what the SIS is measuring.

Consistent with the Stirling County study and the Midtown study results in their general population samples, we found that our composite scale showed substantial correlations of .67 with our psychiatrists' caseness ratings and .66 with their impairment ratings. Neither the Stirling County nor the Midtown study researchers, however, tested their screening scales against ratings made in patient samples. When we did so, we found a sharp contrast with the results in the general population. In our patient-prisoner sample, the correlation of our composite scale with the caseness ratings was only .20; with the impairment ratings, it was little better, .32. How can we account for the difference between the patient sample and the general population sample?

There was some restriction of range on the caseness ratings for the patients, since only about 10 percent were judged unlikely to be cases by our psychiatrists, compared with about two-thirds of the community sample. However, on the impairment ratings and the composite scale, variances for the patients were as great as or greater than in the community sample. Thus, restriction of range cannot account for the lower correlations in the patient-prisoner sample.

There is, however, another difference between the patient and prisoner sample on the one hand and the community sample on the other that could account for the contrasting correlations. While almost two-thirds of the respondents judged to be cases in the community sample were diagnosed as neurotic, only about one-fifth of the patients and prisoners received this diagnosis. A plausible hypothesis, therefore, was that our scale is more strongly related to the caseness and impairment ratings in the commu-

nity than in the patient-prisoner sample because the scale is measuring neurosis rather than other types of disorders.

If this was so, we would expect a higher average score among respondents diagnosed as neurotic than among respondents given other diagnoses in both the community and the patient-prisoner samples. This expectation was not confirmed, however, in either sample. In both samples, the hypothesis was tested using the following categories of diagnosis: all psychoses, all neuroses, all personality disorders, and all other types combined. These classes defined one component of a four-way analysis of variance of the composite scale in which sex, ethnicity, and education were controlled. In the prisoner and patient sample, the F test for diagnosis was insignificant ($F = 0.669$, d.f. = 3,51) with the following group means: 1.96 for psychoses ($N = 41$), 1.88 for neuroses ($N = 19$), 1.51 for personality disorders ($N = 21$), and 1.91 for all other diagnoses combined ($N = 5$).

Since about two-thirds of the community sample were judged unlikely to be cases by the psychiatrists who interviewed them, we designated "no disorder" as a separate category for such respondents; for those community sample subjects rated as likely to be cases, we made the same diagnotic distinctions as in the analysis conducted with the patients and prisoners. In this analysis, the F ratio for diagnosis was significant ($F = 23.82$, d.f. = 4,85) owing to differences between the no disorder group and the others judged to be cases. Planned contrasts revealed no significant differences between the group diagnosed neurotic and the other groups diagnosed as cases. The group means on the composite scale were as follows: no disorder 0.45 ($N = 81$), psychoses 1.59 ($N = 3$), neuroses 1.99 ($N = 28$), personality disorders 1.36 ($N = 8$), and other diagnoses 0.43 ($N = 4$).

Do these scales measure a common denominator of functional disorder? Is it possible, therefore, that we have in this composite scale a measure of symptomatology that is frequently present in most, if not all, functional disorder with no distinction as to type or severity? If it is measuring a common denominator of all functional disorder, the scale would be an ideal screening instrument. Let us consider some further evidence on this question.

As Murphy (1974) has pointed out, most of the evidence on the relation of screening scales to clinical psychological disorder is of the kind we have presented—the ability of the measures to discriminate patients from nonpatients and to correlate in general population samples with clinical evaluations by psychiatrists that were not, as a rule, independent of the symptoms in the screening measures. Some additional evidence has recently become available on the relation of the scales to independent measures of clinical psychological disorders. Schwartz, Astrachan, and Myers (1973) found the Gurin Mental Status Index (Gurin et al., 1960), composed of 20 items from the Stirling County and Midtown study screening scales, to be only weakly related to two other measures of psychopathology in a sample of patients who had been diagnosed as schizophrenic. The measures, the Psychiatric Evaluation Form (Endicott and Spitzer, 1972) and the New Haven Schizophrenia Index (Astrachan et al., 1973) were more strongly correlated with each other (.67) than with the Gurin scale (.55 and .39, respectively). Moreover, Weissman, Myers, and Harding (1978) found that a subset of five of the Gurin scale items plus three similar items, all of which were meant to measure depression, were only weakly related to diagnosed depression in a community sample consisting of 515 subjects. Of the highest-scoring 100 respondents on this eight-item scale, only 28 percent were diagnosed as having a major or minor depression by Research Diagnostic Criteria (Endicott and Spitzer, 1978) on the basis of interviews with the Schedule for Schizophrenia and Affective Disorders (Spitzer, Endicott, and Robins, 1978).

Murphy (1974) administered a screening scale similar to those above the 1,170 freshmen at the University of Singapore and then followed them for from two to three years to learn who exhibited abnormal behavior, sought psychiatric aid, or made much greater than average use of health services. He concluded on the basis of this follow-up that "the symptom checklist . . . failed to identify vulnerable individuals" (1974:260).

In one of our own studies, psychiatrists used a structured clinical interview called the Psychiatric Status Schedule (PSS) (Spitzer et al., 1970) to follow up 55 adults from a community sample similar to the one used in the present study (B.P. Dohrenwend, 1973:485). All 55 respondents had previously been interviewed an average of four years earlier with the 22-item screening scale developed by Langner (1962) in the course of the Midtown study. At that time, these respondents had no history of prior treatment with members of the mental health professions. The psychiatrists who conducted these interviews were blind to the respondents' screening scores in the earlier interview when they made Stirling County caseness ratings and Midtown study impairment ratings of the respondents on the basis of their PSS interviews. Almost two-thirds of the 14 respondents who were iden-

tified as "cases" on the basis of the most frequently used cutting score on the Langner scale turned out not to be "cases" in the judgment of the psychiatrists who interviewed them four years later. At the same time, four of the nine respondents who turned out to be "cases" four years later would have been missed on the basis of their earlier screening scores.

In the face of the results from these studies that provide contemporary and prospective cross-checks with independent indicators of disorder, it is difficult to quarrel with Murphy's conclusion that these screening scales and, by extension, our composite scale are not direct measures of either clinical psychological disorder or vulnerability to such disorder. Unfortunately, the studies offer little evidence on what the screening scales do measure.

Alternative conceptualizations. A number of clinicians have described dysfunctional and distressing psychological states that could not be encompassed within the domain of diagnosable psychopathology. Let us see how well these concepts fit our results.

Dysthymic states. Foulds (1976) proposed the concept of "dysthymic states" consisting of anxiety, depressed mood, or elation, which he distinguished from neurosis and integrated and nonintegrated psychosis. In his formulation, the last three would be characterized by dysthymic states, to be sure, but other more distinctive and severe psychopathology would be present as well. The concept of "dysthymic states," then, also appears to fit some of the facts that we have discovered about our scale. There are, however, problems with dysthymic states as a description of what we have measured. The content of our scale does not cover elation; moreover, it is not limited to the other two types of dysthymic states. Among other types of items included in our scale, for example, are those measuring Perceived Physical Health, which focus on physical illness and poor physical health.

Helplessness and hopelessness. The work of researchers at the University of Rochester's Department of Psychiatry suggests one reason why physical illness might be integrated into a scale that also includes disturbed affect. In a series of retrospective studies of medically ill hospitalized patients, these researchers were impressed by evidence that feelings of helplessness and hopelessness and a "giving up–given up" complex often preceded development of a wide variety of physical illnesses (Schmale, 1972). This finding suggests, however, that along with dysphoric mood and perceptions concerning physical illness and poor physical health, helplessness and hopelessness are also part of what our composite scale is measuring, a point to which we will return.

Pseudo neurosis. Another possible conceptualization of what our scale is measuring is Schofield's "pseudo or quasi neurosis" (1964:62). Schofield argued that pseudoneurotics are "persons who are in some way psychologically uncomfortable and maladjusted (or maladapted), who are neither psychotic nor neurotic" (1964:62). Like most of us at some time or another, they are people who are experiencing normal anxiety and unhappiness in circumstances of marked situational stress (1964:63). The descriptive term itself, however, with its adjectives "pseudo" or "quasi" tends to put the emphasis more on what is *not* being measured than on what *is* being measured.

Frank's concept of demoralization. Consider another candidate, Jerome Frank's (1973) concept of "demoralization." "Demoralization" describes a state that Frank attributes to all persons seeking psychotherapy or other personal help regardless of diagnostic label (1973:314). He suggests that "a person becomes demoralized when he finds that he cannot meet the demands placed on him by the environment, and cannot extricate himself from his predicament" (1973:316). Among the sources of the predicament, Frank mentions environmental stresses, giving the example of wartime experiences (1973:316); constitutional defects (1973:316); learned incapacity (1973:317); existential despair (1973:317); all physical illnesses, especially chronic physical illnesses (1973:46–47); and "crippling (psychiatric) symptoms" (1973:315). Frank's theoretical formulation of the construct "demoralization" fits not only the facts discovered and questions raised in the present study leading to our composite scale of psychological and somatic distress but also the results obtained by others with screening scales that are similar to it.

In setting forth his concept of demoralization, Frank was drawing on his clinical experience with patients and with persons who had been in situations of extreme stress. He did not develop measures of his construct or suggest detailed specifications that might help accomplish the job. There is, thus, a considerable leap of inference as we move to Frank's construct from our analysis of symptom scales in the present research and from our interpretation of results obtained by others with psychiatric screening scales. However, the concept of demoralization as developed by Frank fits the facts that we have

discovered more precisely than the competing constructs of caseness, impairment, dysthymic states, helplessness and hopelessness, and pseudo neurosis that we also considered.

Moreover, the construct of demoralization as developed in Frank's monograph has heuristic value in its own right. The five scales that we investigated in the present study do not exhaust the meanings attributed by Frank to demoralization. For example, self-esteem and helplessness-hopelessness are central to the meaning of Frank's construct. It is quite possible, through further research, to test whether measures of self-esteem, helplessness-hopelessness, and other variables that are theoretically relevant to the construct of demoralization as Frank describes it are in fact strongly related to each other and to the measures of Sadness, Anxiety, Enervation, Psychophysiological Symptoms, and Perceived Physical Health that we have investigated in the present research. Such further investigation would test the boundaries of demoralization as a dimension of people's psychiatric condition.

Perhaps even more important for community studies of psychopathology, such investigations of the boundaries of demoralization should tell us what types of psychological symptomatology are empirically distinct from demoralization. Although measures of demoralization, like measures of body temperature, may be useful in pointing to the existence of problems, their lack of specificity regarding the nature of the problems limits their usefulness. Progress in community studies will depend in large part on our ability to develop measures of other dimensions of psychopathology that are as reliable and as generally applicable as our measure of demoralization but that point to the presence of distinctive types of clinical psychological disorders. Our current work is directed toward these further goals.

Discussion: What brief psychiatric screening scales measure

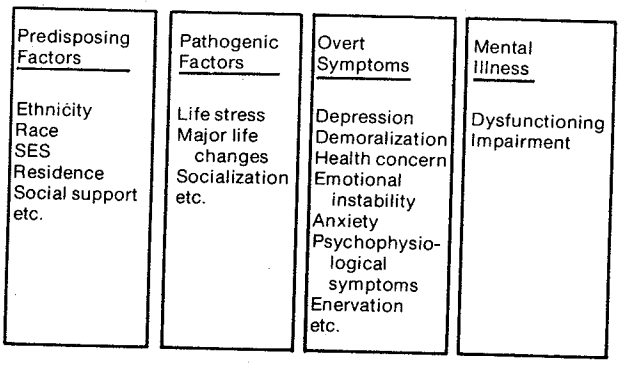
Thomas T.H. Wan, Department of Sociology, University of Maryland/Baltimore County Campus

In recent years there has been considerable interest in formulating measures of mental health status based on self-reported data. Many studies have attempted to provide evidence of the validity and reliability of subjective well-being as an important component of mental health. However, fewer studies have systematically examined the applicability of a composite index of psychiatric symptoms to reflect accurately the type of psychiatric disorder and to discriminate precisely the levels of impairment. It may be the case that the screening instrument has not been validated for its sensitivity and specificity in screening community members. The sensitivity of the scale is defined as the capacity to identify positively those who actually perceive themselves as having psychiatric illnesses, and the specificity of the scale is defined as the capacity to identify those who perceive themselves as having no illness. Obviously, it is relevant to raise the question of why a sensitive and specific instrument has not yet been developed. Does it represent the failure of psychiatric researchers to conceptualize "mental dysfunctioning" as a unidimensional concept? Or is it a limitation of current research methodology in dealing with measurement problems associated with the latent structure of the construct "mental dysfunctioning" or "mental illness"?

Mental health research is a difficult area in which many basic concepts are not agreed on by researchers. However, by merely formulating the pertinent questions in a brief psychiatric assessment instrument, the authors have moved the state of the art forward. Overall, the paper is very well written. Findings were interesting and have enhanced our understanding of the difficulties in the development of reliable and valid psychiatric instruments.

In reading the paper, I learned that the measurement of mental illness based on self-reported symptoms still needs greater refinement and standardization. Several problems pertaining to validity of psychiatric assessment require further discussion.

1. *Clarification of the relationship between psychiatric symptomatology and diagnosis.* There is a need for establishment of a new research paradigm to focus on the causal relationship of psychiatric dysfunctioning or impairment to common pathogenic factors, psychiatric symptoms, and predisposing factors. What are the linkages among these factors?



The overt symptoms can be identified precisely by the screening instrument, and the diagnosis of mental dysfunctioning or impairment can be determined by clinical assessments.

2. *Scale construction.* The authors selected major psychiatric symptoms that represent five distinct dimensions of psychopathology. Scales were constructed based on the scoring of the presence or absence of specific symptoms or items. A total SIS symptom scale was than computed from the summation of the five subscales. Although the analysis of internal consistency between the item and subscale revealed that the construction of subscales was reliable, the validity of this procedure was not fully demonstrated. In order to validate the dimensionality of the SIS symptom scale, two

additional procedures are recommended: (a) to include all 40 symptom items in a factor analysis using oblique rotation and (b) to use the fixed parameters of five distinct psychiatric symptom dimensions in a multiple indicator model and to confirm the desired properties of the scale. In the latter approach, the goodness of fit of the maximum likelihood solution under the hypothesis (e.g., five symptom scales measure a unidimensional aspect of mental dysfunctioning or mental illness) can be examined by the confirmatory maximum likelihood factor analysis (see Jöreskog, 1969).

3. *Test of sensitivity and specificity of the scale.* Assessment made by the screening instrument cannot be considered valid if it cannot be confirmed by psychiatric diagnosis. Based on this principle, we can derive (a) the sensitivity test: the ratio of the number of true positives to the total number of persons who had certain levels of symptomatology, i.e., positive cases with symptoms identified from the screening instrument; and (b) the specificity test: the ratio of the number of true negatives to the total number of persons who reported no, or a low level of, symptoms. The criterion set in the SIS screening scale determines the level of sensitivity and specificity. Generally, a low limit results in a more sensitive test, and a high limit results in a more specific test (Lilienfeld, 1976). It is desirable to set an intermediate limit that may produce minimum total errors in the classification of cases. Furthermore, I recommend that both sensitivity and specificity tests be performed for each of the psychiatric disorders (neurosis, psychosis, personality disorders, etc.).

In conclusion, the utility of the brief psychiatric screening instrument as a classifier of individuals for such purposes as allocation of mental health services has yet to be proven by further research on its validity and applicability in community psychiatry.

Development of social support scales*

Nan Lin, Department of Sociology, State University of New York at Albany

Alfred Dean, Department of Psychiatry, Albany Medical College

Walter M. Ensel, Department of Sociology, State University of New York at Albany

The concept of social support has drawn increasing research attention for its potential contribution to the epidemiological explanation of illness. This is especially evident within the framework of the stressor-illness model. In this model, the relationship between stressors (usually measured with a stressful life events scale) and illness (especially psychiatric symptoms and depression) has been well documented (B.S. Dohrenwend and B.P. Dohrenwend, 1974). Yet, the stressors usually account for less than 10 percent of the variation of any illness measure studied. Thus, calls are frequently heard (B.S. Dohrenwend and B.P. Dohrenwend, 1978; Rabkin and Struening, 1976) for incorporating additional factors to increase the explaining power of the model. It has been further suggested that some of these factors may serve as mediating or buffering factors between the stressors and illness (Cassel, 1974; Cobb, 1976; B.S. Dohrenwend and B.P. Dohrenwend, 1978; Gore, 1978; B. Kaplan, 1975; B. Kaplan, Cassel, and Gore, 1977; Rabkin and Struening, 1976). Social support has clearly emerged as one such important factor (Dean and Lin, 1977; B.S. Dohrenwend and B.P. Dohrenwend, 1978).

Presently, the specific role played by social support in the stressor-illness model is unclear. Theoretical expectations suggest that it may either reflect structural resources (Leighton et al., 1963) against the onset of stressful life events or serve as a reactive and coping mechanism buffering the potential stressing effects of life events (Cassel, 1974; Cobb, 1976). These causal issues are currently being debated and examined by a number of investigators, including ourselves (Dean and Lin, 1977; Lin et al., 1979).

A critical requirement in the systematic examination of the role of social support in the stressor-illness model is the availability of sufficiently reliable, valid, and precise measurements of variables. Important strides are being made in improved measurements for stressful life events (B.S. Dohrenwend et al., 1978; Rahe, 1975; Ross and Mirowsky, 1979) and illness (e.g., for psychiatric symptoms and depression, Dohrenwend et al., 1978; Myers, 1976; Weissman, Myers, and Harding 1978). On the other hand, existing social support scales tend to be ad hoc in nature and used for their predictive validity (Gore, 1978; Lowenthal and Haven, 1968; Medalie and Goldbourt, 1976; Nuckolls, Cassel, and Kaplan, 1972). We have not been able to identify social support scales that have been systematically and empirically tested for their reliability and validity properties. Other studies have resorted to the use of surrogate indicators such as marital status and other sociodemographic variables (Myers, Lindenthal, and Pepper, 1975; Pearlin and Johnson, 1977). Few studies using the same social support indicators have been reported, and these were not designed for scale assessment or development (Berle et al., 1952; Holmes et al., 1961; Moriwaki, 1973).

The purpose of this paper is to present some preliminary results of our efforts in developing reliable, valid, and precise measures of social support. These efforts are part of a research program designed to investigate systematically the roles of social support and stressors in the explanation of psychiatric symptoms and physical illness. Thus, the results to be reported here are tentative and will be updated periodically in the next few years.

Our work is guided by contributions made by other researchers as well as conceptualizations of our own. Our intentions were to cast a wide net over a variety of existing and potential scales and to subject them to rigorous testing.

*The work reported here was supported by a grant from the Center for Epidemiological Studies, the National Institute of Mental Health (MH 30301). We acknowledge the participation of Ronald Simeone and Irene Farrell in data collection and analysis.

Selection of social support items

Basically, we included four groups of social support items for examination. They can be identified as (1) the Medalie-Goldbourt family problem items, (2) the Lowenthal-Haven-Kaplan confidant items, (3) the neighborhood and community satisfaction items, and (4) some newly constructed instrumental-expressive support items. The first three groups of items were scales used or proposed by others. The last group of items was derived from our own conceptualization that social support should reflect the primary functional areas. A brief introduction to each group of items is in order.

Medalie and Goldbourt (1976) regarded their scale as a measure of "family problems." Conversely, it is assumed to measure good family relationships and love and support provided by the wife. They found that positive scores were associated with a significantly lower risk of anginia pectoris among men, even in the presence of biological risk factors.

Lowenthal and Haven (1968) regarded the presence of a "confidant" as an indicator of the availability of an "intimate relationship." Studying an elderly population, they concluded that the presence of an intimate relationship served to reduce the risk of depression in the context of gradual role losses as well as the traumas of retirement and widowhood. Subsequently, Moriwaki (1973), in a community population of retired persons, found a direct relationship between the number of confidants and psychological well-being. Kaplan's (1975) proposed scale items direct attention to other potentially significant attributes of confidant relationships.

The supportive environment for an individual in the neighborhood and the larger community is the third area that we examined. The structural effects on illness have long been observed and documented (Leighton et al., 1963). Yet most of the past efforts considered communities as the units of analysis. When such structural effects were measured at the individual level, the items tended to be combined with other types of social support items to form an overall scale. In a recent study of Chinese-Americans in Washington, D.C. (Lin et al., 1979), we found that satisfaction with neighborhood and community constituted the most effective social support items for predicting (fewer) psychiatric symptoms. Thus, it seemed feasible to construct a scale of neighborhood and community satisfaction for the individual level of analysis.

Finally, implicit in most existing social support scales are both the expressive and the instrumental dimensions. Conceptually, social support serves both functions. We define an instrumental relationship as one in which the relationship is used to achieve an end that is distinguishable from the relation itself. Thus, relationships used to seek a job, to look for a doctor, or to get financial help are instrumental in nature. An expressive relationship, on the other hand, serves as both the means and the end. It does not have any extrinsic purpose other than what it may mean to the individuals maintaining such a relationship. Friendship is a typical example of an expressive relationship. Much of the existing literature on social support mixes the two types of relationships. Yet there is a need to distinguish the two functions, in case they show differential effects on any illness measures studied. Since there is a lack of empirical measures, we decided to explore items that might tap these two dimensions of social support.

In the remainder of this paper, we describe our efforts in developing these social support items into precise, reliable, and valid scales.

The study was conducted with a sample of adults, aged 20 and over, in the Albany-Schenectady-Troy SMSA. The 99 respondents were drawn from a modified area probability sample in which two consecutive households came from each sampled block. Those interviewed were predominantly white (90 percent), and the majority were women (74 percent). The age distribution of the respondents was approximately normal, with a mean age of 42.

The majority (58 percent) of those interviewed were married, 11 percent were either separated or divorced, 20 percent were widowed, and 10 percent had never been married. A large number of them had lived in their county of residence over ten years (79 percent), and over one-half (55 percent) had lived at their present residence for more than five years.

One-third (33 percent) of the respondents had not completed high school, and slightly less than that (31 percent) had gone beyond high school. Among all respondents, 45 percent were employed, 8 percent were unemployed, 12 percent were retired, and 33 percent considered their primary occupation to be keeping house. The occupational positions of those employed ranged from professional/technical to service workers, with a median prestige score of 36.5. Income distribution was fairly even, with a median family income between $10,000 and $14,999.

Development of scales

Medalie-Goldbourt items. These items, taken from Medalie and Goldbourt's (1976) study of angina pectoris among men, focus on family problems: (1) family problems in the past, (2) family problems at present, (3) effects of spouse/children not listening or opposing, and (4) whether spouse shows his/her love. Although these sames items were used, the original response categories were modified. In the original Medalie and Goldbourt study, each item had two response categories: "0" for no serious problems or no problems at all and "1" for very serious or serious problems. In our study, we constructed four response categories for each item. For the first two items, they were "no problems at all," "no serious problems," "yes, serious problems," and "yes, very serious problems." Responses for the third item were "never happens," "does not effect me especially," "upsets me quite a bit," and "upsets me very much." The response categories for the fourth item were "loves me and shows it often," "loves me and shows it occasionally," "loves me but never shows it," and "does not love me." The rationale for having more response categories was simply to increase the sensitivity of the measure by increasing the number of categories. A total scale score was constructed for each respondent by summing the responses over the four items. Table 1 shows the means, standard deviations, and inter-item correlations.

As can be seen, all items correlate highly (between .590 and .765) with the total score. All inter-item correlations are in the positive direction, and all except two of the correlation coefficients are significant at the .05 level. The fourth item (spouse not showing love) seems to have a slightly weaker relationship with the total score, as compared with other items. Nevertheless, as a whole, there seems to be internal consistency, as reflected in the convergent validity of the scale items. However, since most of these items concern married persons, their application to a general population is limited. In our study, only 57 of the 99 respondents qualified to respond to these items.

Lowenthal-Haven-Kaplan items. For this battery of 11 items, three dealing with the availability of a confidant were taken from the Lowenthal-Haven study (1968). The scale included (1) "Is there someone you confided in or talked to about yourself or your problems?" (2) name and relationship of this person, and (3) "In the past year, has there been any change in your relationship with this person?" The other eight items, taken from Kaplan (1975), dealt with various aspects and relationships with confidants (size, reachability, density, content, directedness, durability, frequency, and intensity). In our study, respondents were asked to identify as many confidants as they wished, i.e., "During the past 12 months, have you had anyone that you could trust and talk to?" and "How many people have you been able to trust and talk to?" Those who identified one or more confidants were asked to write down on a piece of paper the names of up to three persons whom they were most likely to talk to. The interviewer did not ask to see the names. Then, a series of questions was asked relative to each confidant listed. Thus, the data yielded three sets of responses to the items relating to the three persons.

In the analyses conducted so far, we have concentrated on responses regarding the first confidant named. Here, we will report the internal consistency among the seven Kaplan-type items. Responses for these seven items are as follows: (1) durability (number of years known); (2) frequency of contact ("most or all of the time," "occasionally or a moderate amount of time," "some or a little of the time," "rarely," "never"); (3) density—"How often have you talked with this person when you had a problem?" ("most or all of the time," "occasionally or a moderate amount of time," "some or a little of the time," "rarely," "never"); (4) directedness—"How often has this person talked over their problems with you?" ("most or all of the time,"

Table 1
Inter-item and item-total correlations of the Medalie-Goldbourt items
(N = 57)

Item	1	2	3	4	Total scale	$\bar{X}$	S.D.
Family problem—present	1.000	.703	.259	.126*	.6904	1.28	.49
Family problem—past		1.000	.260	.209*	.7335	1.43	.59
Spouse/kids not listening			1.000	.3726	.7647	2.12	1.26
Spouse not showing love				1.000	.5901	6.09	1.73

*Not significant at the .05 level.

"occasionally or a moderate amount of time," "some or a little of the time," "rarely," "never"); (5) reachability—"How easy has it been to get a hold of this person?" ("very easy," "easy," "somewhat easy," "not very easy," "not easy at all"); (6) content—"How freely have you been able to talk about anything you wished with this person?" ("very freely," "freely," "somewhat freely," "not very freely," "not freely at all"); and (7) importance—"How important would you say this person is to you?" ("very important," "important," "somewhat important," "not very important," "not important at all").

All items, except durability, had five ordinal response categories. Before a summated scale could be constructed, we grouped the durability-item responses (number of years) into four categories, each of which covered 25 percent of the responses (2–15 years, 16–30 years, 31–45 years, and 46–60 years). For larger samples, they probably should be grouped into five categories so that the number of response categories would be completely comparable to those of the other confidant items. Again, a total scale score was computed for each respondent by summing the scores over the seven items.

The results are shown in Table 2. Item-total correlations range from .279 to .822. Among the inter-item correlations, 12 of the 21 coefficients are not significant. Obviously, these items do not constitute a unidimensional scale. The decision was to examine the relationships between these items and the dependent variables individually. We hoped that such an analysis would identify specific confidant characteristics that contribute to the prediction of the dependent variables.

Neighborhood and community satisfaction items. Two items on satisfaction with neighborhood and community were incorporated in the study: (1) "In general, how satisfied are you with this neighborhood?" ("very satisifed," "somewhat satisfied," "somewhat dissatisfied," "very dissatisfied") and (2) "On the whole, how satisfied are you with living here in this community?" ("very satisfied," "somewhat satisfied," "somewhat dissatisfied," "very dissatisfied"). The means and standard deviations of the two items were 1.59 and 1.57, and .89 and .90, respectively. The responses tended to concentrate in the positive categories, as expected. The decision was to use the two items (the zero-order correlation between them being .67) to construct a summated scale of community-neighborhood satisfaction.

Instrumental-expressive support items. Incorporated in the study was a set of 26 items focusing on the activities and aspects that might provide (or jeopardize) either instrumental or expressive support to the respondent. Since we felt that there was a need to construct new scales, we composed a number of items, based on their face validity, that described the instrumental and expressive support systems of respondents. One objective in constructing these items was to make them capable of describing the various modes of support despite differences that might be attributable to sociodemographic characteristics (e.g., marital status, employment status). In other words, we wanted to make the items applicable across demographic subsets and status and role characteristics of respondents.

Following a general introduction ("Would you tell me how often you have been bothered by these problems over the past 12 months?"), a list of 26 items was asked of each respondent. There were four response categories: "most or all of the time," "occasionally or a moderate amount of time," "some or a little of the time," "rarely or none of the time." The responses were subjected to a factor analysis (orthogonal solution, varimax rotation, and a limiting eigenvalue of one or higher), which resulted in a

Table 2
Inter-item and item-total correlations among the Kaplan items
(First confidant; N = 84)

Item	1	2	3	4	5	6	7	Total scale	$\bar{X}$	S.D.
Durability	1.000	−.124	−.156	.069	−.022	.156	.139	.279	2.97	1.01
Frequency		1.000	.522	.338	.319	.244	.052	.620	1.44	.86
Density			1.000	.750	.164	.088	.094	.753	1.83	1.14
Directedness				1.000	.246	.146	.180	.822	1.92	1.25
Reachability					1.000	.200	.033	.457	1.27	.70
Content						1.000	.212	.411	1.23	.50
Importance							1.000	.353	1.21	.56
Total score								1.000	11.88	3.42

five-factor solution. The five factors identified were (1) monetary problems, (2) lack of companionship, (3) demands, (4) communication problems, and (5) no children. The items highly loaded on each of the five factors are presented in Table 3. The last factor had only a single highly loaded item.

We could have constructed factor scores by using a regression formulation with beta weights assigned to each contributing item. That approach would have used all of the information contained in the data matrix. However, it would also have assumed that the items in the matrix had substantive reasons to be self-contained. There was no reason to assume that the items we constructed were self-containing. Thus, we decided to identify the items most representative of each factor and to construct a summated indicator from these items for each factor. (See the items used for each factor in Table 3.) These computations resulted in five constructed variables tapping the instrumental and expressive support factors. It seemed clear that monetary problems and demands were instrumental dimensions and that lack of companionship, communication problems, and having no children were expressive dimensions. The decision was to use these five constructed instrumental-expressive support scales either separately or in the functional groups (instrumental versus expressive).

Scale validation procedures

The dependent and control variables. The validation process began with the identification of the dependent variable, proceeded to an examination of the relationship of each set of independent variables with the dependent variable, and concluded with a tentative construction of a model in which all of the independent variables were examined simultaneously for the dependent variable. We also incorporated several sociodemographic variables and stressful life events as other independent variables in validating the social support scales.

Selection of the dependent variables.[1] Two instruments were used to measure psychiatric symptomatology: the Center for Epidemiologic Studies Depression (CES-D) Scale (Markush and Favero, 1974; Radloff, 1977) and the Gurin Scale (Gurin, Veroff, and Feld, 1960), a general psychiatric symptom inventory.

The CES-D Scale consists of 20 items, each of which was answered on a four-point scale ("none of the time" to "all of the time") based on the last week. Scores have a possible range of 0 to 60, with higher scores indicating depressed mood.

The Gurin Scale is a 20-item index with a four-point response ("often," "sometimes," "hardly ever," "never") and a possible score range of 20 (maximum severity) to 80 (complete absence of symptoms).

The zero-order correlations between the independent variables and these two dependent variables were consistent and in the same direction. For parsimony, subsequent analyses focused on the depression scale as the dependent (criterion) variable.

The sociodemographic variables. Selected sociodemographic variables included sex, age, marital status (married versus not married), occupational prestige, and family income. Only marital status and income had significant relationships with the depression scale. Thus, it was decided to explore marital status and income further, along with other independent variables in the modeling process.

Table 3
Factor loadings of instrumental-expressive support scales

Items	Loading on instrumental/expressive factors*
	Factor I
Monetary problems:	
Problems managing money	.809
Deciding how to spend money	.790
Not enough money to do things	.875
Not enough money to get by	.828
	Factor II
Lack of companionship:	
No close companion	.720
Not happy with marital status	.834
Not enough close friends	.664
Problems with spouse/ex-spouse	.811
No one to show love/affection	.823
Too dependent on others	.543
	Factor III
Demands:	
Too many responsibilities	.833
No one to depend on	.782
Too many demands	.793
Unsatisfactory sex life	.722
	Factor IV
Communication problems:	
Problems communicating	.627
Problems with children	.805
Unsatisfying job	.753
No one to understand problems	.738
Conflicts with those who are close	.781
	Factor V
Not having children	.794

* All coefficients significant at the .001 level.

Stressful life events. The other set of independent variables crucial in our research concerned stressful life events (SLE). Much analytical work has been done on stressful life events, with more recent discussions focusing on the issue of negative (undesirable) events versus total events, number of events to be studied, and subjective versus objective evaluations. In our study, the strategy was as follows: (1) include the original Holmes and Rahe (1967) items (excluding Christmas) so that we could replicate the original findings; (2) add items recently proposed by Rahe (1975), along with items used by Myers and others (1972); and (3) expand certain items to reflect positive or negative effects. Further, stressful life events were tapped for each respondent for two time periods (last six months and the previous six months) to obtain the temporal sequence of events. Finally, these questions were asked relative to the respondents and their significant others ("members of your family, or other important people in your life").

The data were simply summed for each of the four variables: (1) total unweighted score of stressful life events that occurred to the respondent in the last six months (SLE-S6) and (2) in the previous six months (SLE-S12); (3) total unweighted score of stressful life events that occurred to the respondent's significant others in the last six months (SLE-O6) and (4) in the previous six months (SLE-O12). Past research has indicated that weighted and unweighted total scores do not show substantial differences in their relationships to the illness symptoms and that negative-event total scores and all-event total scores have about the same amount of effect on illness symptoms (i.e., the zero-order correlation remains about .19 to .23). In view of the above, we did not construct weighted or negative-item scores, focusing only on unweighted sums. Eventually, such scales will be constructed and studied in detail.

Also, in the pretest we did not attempt to measure subjective definitions of desirability or magnitudes of events. If the ultimate interest is to gauge the causal relationships between stressors and illness, it would be more efficient, as pointed out by B.S. Dohrenwend and B.P. Dohrenwend (1978), to construct scales that reflect environmental input rather than individualized resultant evaluations. It seemed to be a reasonable argument to adopt. However, in the future, we intend to incorporate subjective evaluations of respondent's events.

Initial validation. The zero-order correlations between selected independent and dependent variables are presented in Table 4.

Stressful life events. Four scales of stressful life events were analyzed: (1) SLE to self, last 6 months; (2) SLE to self, 6–12 months ago; (3) SLE to significant others, last 6 months; and (4) SLE to significant others, 6–12 months ago. Only the first two scales were significantly related to the depression scale. Since these two scales correlated significantly (.36) and the first scale yielded a correlation slightly higher than the second scale in its relationship to the dependent variable, we decided to focus in further modeling on the first scale (SLE to self, last 6 months) as the indicator of stressful life events. This decision is consistent with many of the previous studies in which life changes to the respondent in the last six months were used as the measure of stressors.

Social support scales. A large number of social support scales and items were examined in conjunction with the depression scale. The Medalie-Goldbourt scale, tapping family problems with a focus on spouse and children, showed a substantial relationship with the depression scale (.42). Among the instrumental and expressive scales, monetary problems, demands, community and neighborhood satisfaction, communication problems, and lack of companionship all showed significant relationships with the depression scale. Two of the seven confidant items showed significant relationships with the depression scale: the durability of the confidant (number of years knowing the confidant) and directedness with the confidant ("How often has this person talked over their problems with you?"). None of the other Lowenthal-Haven-Kaplan items (including number of confidants and former confidants) showed significant correlations with the depression scale.

These results are, in general, encouraging. We were especially satisfied with the effectiveness of many of the instrumental-expressive support scales. The Medalie-Goldbourt scale, in fact, was highly correlated with some of the instrumental-expressive support scales (for example, its correlations with monetary problems, demands, communication problems, and lack of companionship are all greater than .55). It was apparent that the instrumental-expressive support items tapped a dimension similar to that tapped by the Medalie-Goldbourt scale. However, the instrumental-expressive support scales seemed to (1) tap specific areas of support (or the lack of it) rather than general problems and (2) apply to most respondents rather than to just the married respondents. Thus, we decided to focus on the instrumental-expressive support scales in the modeling process. This is not to reject the validity of the Medalie-

Table 4
Zero-order correlations between selected independent variables and the dependent variable (CES-D Scale)

Independent variables	*Coefficient*	*N*	*p*
Sociodemographic variables:			
Sex	.08	99	n.s.
Age	−.02	98	n.s.
*Marital status (not married vs. married)	−.21	98	.02
Occupational prestige	−.17	76	n.s.
*Income (X_{10})	−.45	78	.001
Stressful life events:			
*To self, last 6 months (X_9)	.31	99	.001
To self, 6–12 months ago	.20	99	.02
To significant others, last 6 months	.02	99	n.s.
To significant others, 6–12 months ago	−.02	99	n.s.
Social support scales:			
The Medalie-Goldbourt scale	.42	57	.001
The instrumental-expressive support scales:			
*Community and neighborhood satisfaction (X_5 & X_6)	−.38	97	.001
*Monetary problems (X_1)	.46	97	.001
*Demands (X_2)	.43	90	.001
*Lack of companionship (X_3)	.32	68	.004
*Communication problems (X_4)	.37	69	.001
No children	.03	64	n.s.
Confidant characteristics:			
Durability of confidant (X_7)	−.11	84	n.s.
*Directedness with confidant (X_8)	−.22	84	.02

*Scales and items retained for model validation; see text.

Goldbourt scale. We believe that it is probably a valid scale tapping general family problems as applied to married respondents. However, for our purposes, specific support functional areas as applied to most respondents seem to have been tapped satisfactorily by the instrumental-expressive support scales.

Validation in the modeling process. After we made the decisions regarding the specific independent and dependent variables to be included in the modeling process, we proceeded to consider all of the relationships between the selected independent variables and the dependent variable (the depression scale) simultaneously. These variables are indicated by asterisks in Table 4.

The modeling procedure involved the construction of a regression model for the dependent variable, with the selected independent variables. The models were refined as we eliminated independent variables that did not exceed a regression coefficient of .10. Further, it became necessary to eliminate some independent variables (e.g., occupational status) because of their high correlation with another independent variable (e.g., family income), so that the problems of multicollinearity could be minimized. The final regression model is presented in Table 5.

Table 5
Final regression model

Independent variables	*Dependent variable (depression)*		
	Metric coefficients	*S.E.*	*Standardized coefficients*
Family income (X_{10})[a]	−.915	.240	−.366
Stressful life events (X_9)	.519	.289	.175
Social support:			
Monetary problems (X_1)	.252	.217	.138
Demands (X_2)	.487	.240	.235
Community and neighborhood satisfaction (X_5 & X_6)	−.961	.526	−.180
Constant	42.32		
Error of estimate	6.66		
R^2			.450

[a]The variable legends refer to those in Table 4.

The model suggests four aspects of social support, along with family income and stressful

life events, as the significant predictors of depression. Family income is the most significant contributor, accounting for more than a quarter of the explained variance (.3 of .46) in depression. The four social support scales account for a combined 66 percent of the explained variance, independent of family income. Stressful life events contribute an additional 3 percent to the explained variance.

To assess the direct and indirect effects of the various independent variables and the effects of social support, independent of those from the sociodemographic variables and stressful life events, we constructed the standardized reduced-form equations (see Table 6) and decomposed for the depression scale the direct and indirect effects from the various independent variables (see Table 7). In so doing, we made the stringent assumption that both the sociodemographic variables and the stressful life events *causally* preceded the social support items. Thus, the estimates of the effects of social support on depression are conservative ones.

As can be seen, family income has mostly a direct effect on depression, while stressful life events affect depression both directly and indirectly through social support (or the lack of it).

Table 6
Coefficients of standardized and reduced-form structural equations for depression (CES-D)

Predetermined variables	Structural equation 1	2	3
Family income	−.449	−.453	−.366
Stressful life events		.313	.175
Social support:			
Monetary problems			.138
Demands			.235
Community and neighborhood satisfaction			−.180

Excluding the indirect effects of family income and stressful life events through (the lack of) social support, the indirect and direct effects of social support remain substantial. If we ignored the signs of the coefficients, the total independent effect of social support on depression would be .328.

Since the two instrumental support scales (monetary problems and demands) are strongly correlated (r=.583), the collinearity has reduced the significance of their independent contributions to the dependent variable (see the relatively high standard errors in Table 5). We then reconstructed the model, using an analysis of covariance structures. This approach allows the incorporation of multiple indicators of each variable. Because it estimates both the measurement errors and the equation (relational) errors, it probably constitutes the most powerful and precise statistical tool for social scientists today. In this model, monetary problems and demands were considered to be indicators of the unobserved variable, instrumental support. Also, in order to confirm the weak contributions of the expressive support scales (lack of companionship and communication problems) and directedness with a confidant, we allowed them to reappear in the model (where expressive support was indicated by the lack of companionship and communication problems). The result of this analysis (using a maximum likelihood solution) is presented in Figure 1.

As can be seen, this structural equation model essentially confirms our regression results: the significant independent variables are instrumental support, community and neighborhood satisfaction, and income. The contribution of instrumental support has increased because of the grouping of monetary problems and demands as its indicators. Stressful life events are not significant in their effects on depression. The goodness of fit, as reflected in the chi-

Table 7
Decomposition of effects for depression (CES-D)

Predetermined variables	Total effects	Indirect SLE	Effects via social support	Direct effects
Family income	−.449	.004	−.087	−.366
Stressful life events	.313	—	.138	.175
Social support:				
Monetary problems	.138	—	—	.138
Demands	.235	—	—	.235
Community and neighborhood satisfaction	−.180	—	—	−.180

square statistic, shows that this model is remarkably consistent with the raw data.

The final model suggests that both objective support (income) and social support (mostly instrumental, but also satisfaction with the community and neighborhood) have a much greater effect on depression than do stressful life events.

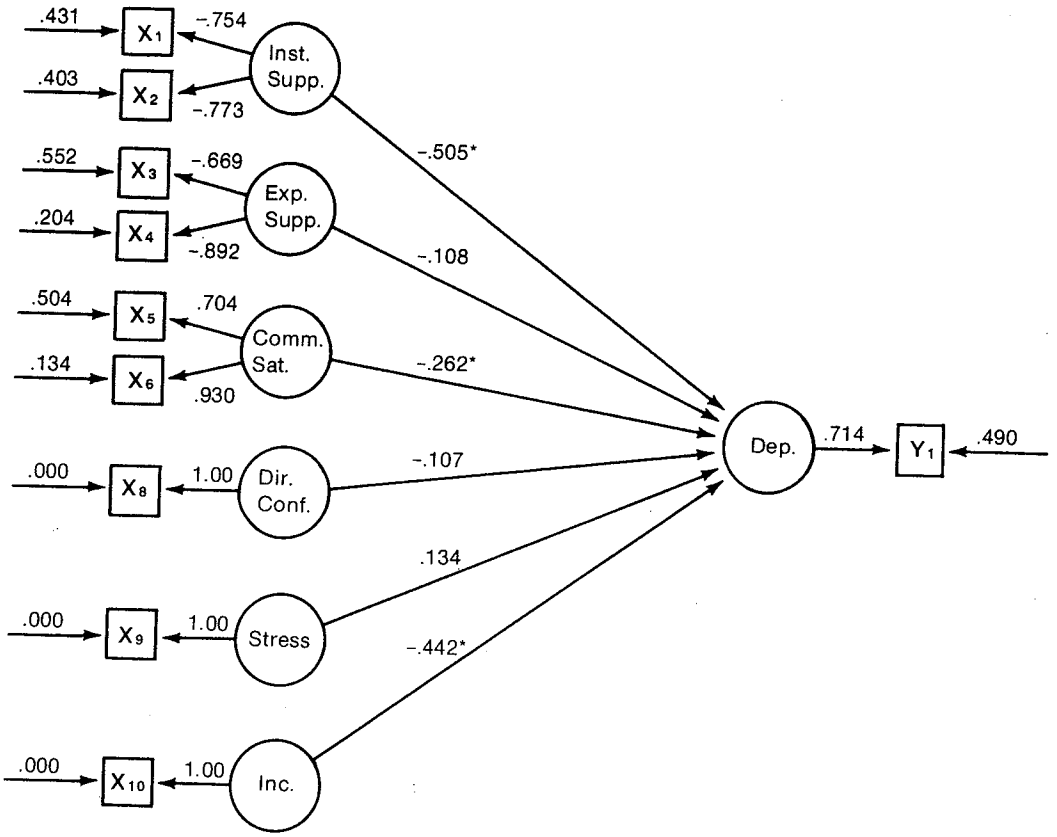
Discussion

The data, while preliminary, suggest that our efforts at constructing social support scales are yielding promising results. Social support measures clearly exert strong effects on depression and other psychiatric symptoms (we have obtained similar results for the Gurin Scale as the dependent variable). Nevertheless, we must offer the following cautions as we proceed to improve and use the social support measures. These cautions apply to the development of any precise, reliable, and valid scales.

First, *the scales that one constructs are inevitably restricted to the items selected for investigation.* Items that form clusters and thus "scales" are not necessarily the only or the "best" scales possible. Other items that do not fall into the clusters either may in fact be substantively meaningless or may simply reflect the investigator's selection of items in the first place. There is no reason to assume that items which do not fit in any scales are automatically "bad" items. It may merely reflect the fact that not enough items tapping the same substantive dimensions were included for scaling. Thus, it is incumbent on the investigator to examine further these "isolated" items and to caution the users of the scales of their nonexclusive nature.

For example, of the 26 instrumental-expressive support items, only 20 items loaded highly on the five dimensions. The other six "isolated" items were either "bad" items or representative of dimensions not tapped well by other items. If the latter case was true, then these items may be unreliable (in the convergent sense) but nevertheless valid indicators of some unknown dimensions of expressive-instrumental support. To examine this possibility, we looked at the zero-order correlations between these items and depression. We found *two* of the items were significantly correlated with depression. We then

Figure 1. The structural equation model of depression



*Significant at .05 level.

Chi-square with 21 d.f=13.0568
Probability level = .9068

entered these two items into the regression equation along with the other independent variables as shown in Table 5. The results showed that the addition of these two items increased the explained variation of depression from 45 percent (Table 5) to 51 percent, as shown in Table 8. Since these estimates show substantial standard errors, they are rather unreliable. However, the analysis warrants the further exploration of other instrumental-expressive support dimensions suggested by these items.

Second, *scales that do not predict or explain one criterion variable may predict or explain well other criterion variables.* In our study, the Lowenthal-Haven-Kaplan confidant items did not predict depression well either as a group or individually. Obviously, further conceptualization is needed and new items must be explored. Also, even when revised items form scales, we doubt that they will show any explaining power in terms of depression when they are considered in conjunction with other support scales and sociodemographic variables, as we have done. Nevertheless, these items are derived from sound conceptualization. It would be premature to discard them from further epidemiological analyses. It may well be that they predict and explain other types of illness and/or help-seeking measures.

Similarly, our study included a large number of items on the respondents' involvement in primary and secondary group activities and their relationships to these groups in times of need (which persons, groups, clubs, or organizations the respondents would go to regarding financial matters, illness, work problems, transportation, etc.). Preliminary analysis did not uncover any significant contribution from these items to the dependent variables. Again, we are refining and retaining some of these items in the research program for further examination.

Third, *scaling should not be restricted to certain analytic strategies.* We have reported results from analyses based on linear relationships. This does not mean that one should not examine curvilinear relationships. We have compiled extensive cross-tabulations to explore possible curvilinear relationships. These efforts did not generate any systematic findings. We will continue to explore the complex alternative relationships, applying different transformations. It is important that we not be bound by the readily available patterns of responses forced by the items and response categories that we constructed.

Finally, *significant statistical relationships do not automatically indicate causal relationships.* Although family background (such as income) seems to temporally and causally precede the illness measures, we are less sure about the causally preceding nature of stressful life events and social support measures. In our study, the respondent was asked to recall the life events and social support activities for a period of time before the current state of illness symptoms and indicators. Such an approach, while inevitable in any cross-sectional investigation, is subject to measurement errors related to actual recall errors and the impact of current-state perceptions and interpretations of past events and activities. We are especially concerned with the causal order of social support measures and illness. It is conceptually as well as empirically viable to argue for their mutual influences. The near-ideal test of temporal causality requires longitudinal data. It is our hope that the social support measures presented here will be examined with longitudinal data to verify their causal effects on illness measures.

Table 8
Regression analysis with two additional social support items

	Dependent variable (depression)		
Independent variables	*Metric coefficient*	*S.E.*	*Standardized coefficient*
Family income	−.923	.241	−.369
Stressful life events	.544	.282	.183
Social support:			
Monetary problems	.107	.219	.059
Demands	.358	.237	.173
Community and neighborhood satisfaction	−.976	.512	−.183
"Not enough responsibility"	1.875	.858	.213
"Too controlled by others"	.717	.789	.102
Constant	49.36		
Error of estimate	6.41		
R^2			.506

Footnote

[1] Our study incorporated a number of other illness measures as potential dependent variables.

History of illness was monitored with a checklist of 55 diseases and conditions covering the major organ systems of the body. Residual categories were used to tap disorders not included on the list. Scores for prior history ranged from 0 to 13, with a mean score of 1.32. Among the respondents, 37 percent acknowledged no previous history of illness, 30 percent stated that they had had one illness, and 32 percent reported two or more illnesses.

A substantially modified version of the Cornell Medical Index (Brodman, Erdmann, and Wolff, 1956), developed through consultation with medical specialists at Albany Medical College, was used to measure physical symptomatology. This index consisted of 81 symptoms (73 for

women and 63 for men) covering the body's major organ systems. Total scores for last month ranged from 0 to 37. Among the respondents, 36 percent experienced two or less symptoms, 35 percent had between three and seven, and 29 percent had nine or more of the symptoms.

Help-seeking behavior was assessed with a series of questions asking respondents how often they had sought treatment for illness from health professionals and health facilities in the last 12 months.

Validation of social support scales relative to these illness measures is being carried out.

Discussion: Development of social support scales

Lisa F. Berkman, Department of Epidemiology and Public Health and Institution for Social and Policy Studies, Yale School of Medicine

The extent to which an individual maintains certain social and community ties is gaining recognition as a part of the social environment that may be related to disease causation. However, as Lin has noted, we know very little about precisely which ties have critical health consequences; investigators differ greatly in how they conceptualize social supports and in how they think of them fitting into a model of disease causation; and finally, we have few, if any, measures or methods of data collection that have proven to be either reliable or valid. I would like to discuss these issues by addressing three questions that concerned me in reading Lin's paper.

What are social supports?

A basic question is precisely what are social supports. Social support generally connotes the amount of emotional support that one receives from friends and relatives. Most implicit assumptions concerning the importance of social ties derive from the concept that social support is a function of the amount of time spent in a relationship with someone and the emotional intensity, intimacy, and reciprocity that characterize the tie. This leads most investigators to deal with small, well-developed groups or dyadic ties between people. Thus, most investigators have, a priori, limited the field of social support to the expressive and emotional ties that an individual maintains with his or her primary relations, most notably spouse, family, and close friends.

Although these intimate relationships may be a major source of emotional support, another aspect of social ties would suggest the importance and cohesive power of what are usually defined as "weak ties." Such ties are characterized by lack of intimacy and by the limited time spent in the relationship. Lin has recognized the importance of these relationships by including items that concern satisfaction with neighborhood and community and a set of items that tap needs for instrumental support. Granovetter (1973) argues that these "weak ties," ignored in most research, may be important in the diffusion of influence and information, mobility opportunity, and political and community organization. In terms of personal support, an individual who has such weak ties to others may obtain information socially distant from them and may have access to a circle of people, should they be needed, that is wider than that for an individual with only strong, closely knit ties.

Two examples will illustrate this point. In a study of how people found new jobs, Granovetter discovered that 16.7 percent of those finding a job through a contact reported that they saw their contact often, 55.6 percent saw this person occasionally, and 27.8 percent rarely. In many cases, a chance meeting or a mutual friend reactivated ties with these contacts, who were only marginally included in the current network. A second example, reported by Boswell (1969) in a study of personal crisis in an African town, found that distant and often unknown relatives of an individual may be expected to take charge of burial ceremonies and finances, again suggesting that intimacy may not always be the critical factor in mobilizing social support.

The relationship between nonintimate weak ties and health status has not been well investigated. However, it is likely that migrants, as well as other mobile and lower-socioeconomic groups who are at high risk of developing a wide range of diseases, have difficulty establishing such ties (Axelrod, 1956; Litwak, 1960a, 1960b, 1961; Moore, 1971). The positive associations between church attendance, organizational activity, and mental and physical health also suggest that extended social and community ties may have some impact on health status (Comstock and Partridge, 1972; Graham et al., 1978; Palmore and Luikart, 1972).

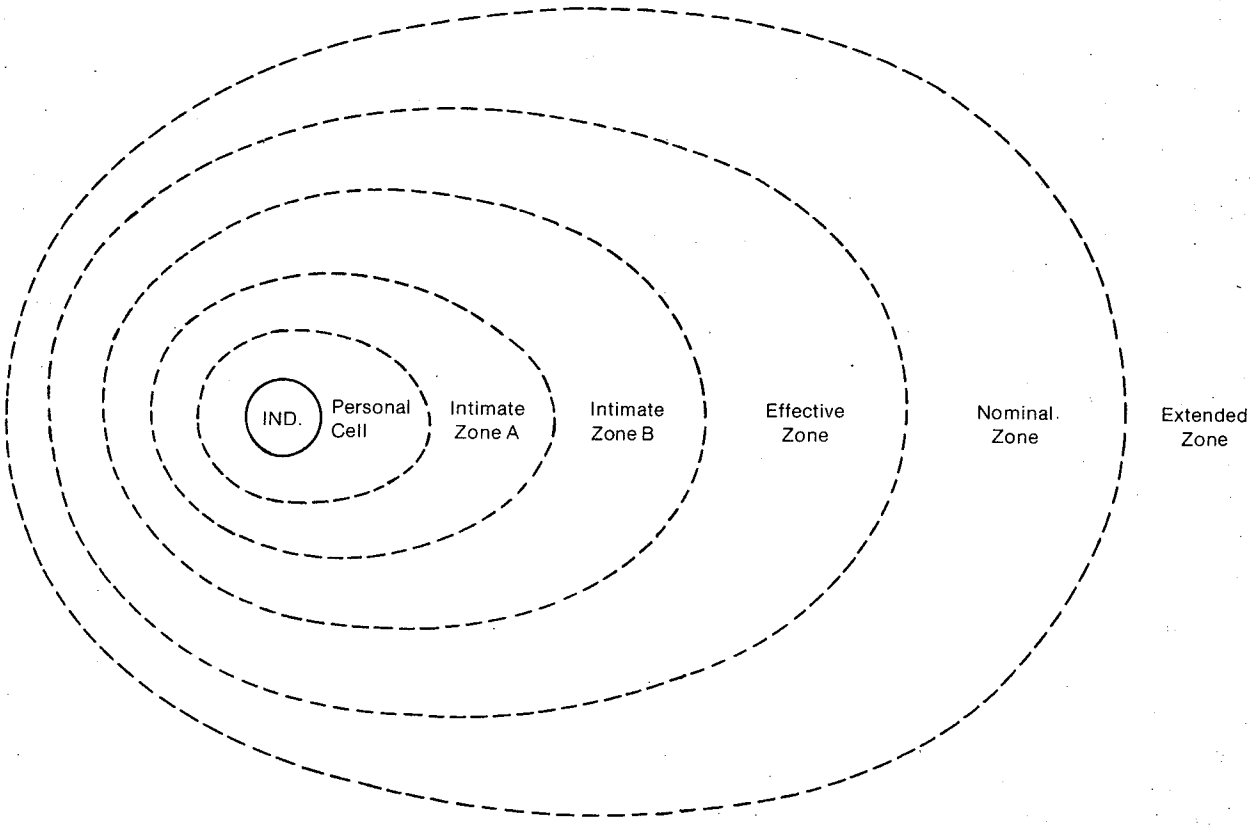
My point concerning the potential importance of extended nonprimary group ties was not meant to negate the important effect of close and emotional contact. Rather, it was presented with the hope that at this early point in the development of measures of social ties, we can develop models and measures that will be broad enough to include both primary and extended contacts and measure both emotional or expressive and instrumental kinds of support. I think Lin and I are in agreement on this point.

In order to do this, I would like to propose (1) a method of analysis that can deal with many kinds of social ties and the varied qualities of such ties and (2) a model of social contacts that I think is sufficiently broad. A useful framework for viewing and conceptualizing the kinds of social relationships that a person has with others is provided by social network analysis. Mitchell (1969), one of the proponents of network analysis, noted that this approach focuses not on the attributes of people in the network but rather on the characteristics of the linkages in their relationships to one another. Several other investigators, notably B. Kaplan, Cassel, and Gore (1977) and Cobb (1976), have also proposed the use of network analysis to assess the extent of an individual's social support.

One of the clearest theoretical models of a personal network with differing levels of interaction or social contact is presented by Boissevain (1974). In this model (see Figure 1), the center circle represents the person being studied. The Personal Cell includes closest relatives and perhaps most intimate friends. Intimate Zone A, the second sphere, includes very close relatives and friends with whom there are active intimate relationships. Friends and relatives with whom one has more passive but still emotionally important relationships comprise Intimate Zone B. The Effective Zone is filled with people who are important in a pragmatic sense for economic and political purposes and for the logistics of daily life. The Nominal Zone contains acquaintances and people who are known by the individual but who have no great importance. The Extended Zone includes individuals who are on the fringe of primary relationships. They may simply be faces that are recognized by the individual.

Social network analysis and the Boissevain model provide us with a way of looking at a

Figure 1. Social network model



Source: Boissevain (1974).

wide array of social and community ties. Now I would like to suggest that instead of talking about social supports, we refer to social networks, since this term is more neutral in value whereas social support builds in the desirability of the circumstances. Also, social support carries the connotation of a focus on intimate relationships. "Networks" is a broader term. I would also like to suggest that the definition of such networks be the web of social and community ties maintained by an individual and the quantity, quality, and morphology of those ties. Although the term "social networks" is very broad and encompasses many relationships and the quality of those relationships, I think it is important that social networks or supports do not become catchall terms for many social and psychological circumstances. For instance, I do not think social networks or supports should include economic, personal, or religious resources or problems; nor should they include daily problems or hassles, psychological states or traits of loneliness, inability to maintain contacts or make friends, extroversion-introversion, or overall measures of life satisfaction. I say this not because these factors may not have health consequences but because it is essential to develop a working definition of social networks that will allow us to form and test hypotheses concerning their association with physical and mental health status. It is also important that we increase our understanding of the ways in which social networks are related to other social circumstances and psychological factors. This task will be enormously complicated if we cannot distinguish among variables.

What are some important dimensions of social networks?

In order to measure relationships or linkages, network analysts look at two major categories of networks: the structure or morphology of networks and the types or quality of interaction occurring among linked people. These dimensions have been described in detail elsewhere. I will review them only briefly here. The first category, concerning the structure or morphology of the network, includes such characteristics as size, range, and symmetry of links. Also included are the degree to which people connected to the individual are known to one another (density) (Barnes, 1969) and the number of steps that it takes to contact a specified person from a given starting point (reachability). In the second category, concerning the quality of interaction, qualities such as intensity, durability over time, frequency of contact, and content of interaction are measured. All of these measures might be applied to a person's family, friends, co-workers, neighbors, church members, and casual associates.

Another approach involves differentiating between actual behaviors or relatively objective contacts and the perception of having certain contacts. In terms of social networks, it may be equally important to be able to distinguish between behaviors or actual contacts (i.e., How often do you see any friends, neighbors, or relatives? Last time you needed some sugar or nails, what did you do? How many close friends do you have? Do you know anyone who lives in your neighborhood?) and the personal perception or assessment of networks (i.e., If you had a personal problem, with whom would you talk the problem over? Are you satisfied with your neighborhood?). The latter dimension adds a psychological component that may prove to be a more sensitive measure and/or may seriously confound the measure of networks with other factors such as depression and dissatisfaction. Since we know so little about how networks function, it is critical at this stage that surveys include both perceived contacts and actual contacts or reported behaviors and that investigators are able to distinguish between them as much as possible.

Finally, in the analysis of social networks, it is important to understand the relationships among different sources of social contact. Most network studies have failed to assess the independent effects of different sources of contact or the interaction effects among them.

I would like to take a moment to discuss some research that I have been conducting on the impact of social networks on mortality risk. In this work we have attempted to deal with some of the issues just discussed. The relationship between social and community ties and mortality for all causes was assessed using the 1965 Human Population Laboratory survey of a random sample of 6,928 adults in Alameda County, California, and a subsequent nine-year mortality follow-up (Berkman and Syme, 1979). Thus, the study was a historical prospective or longitudinal one.

Four sources of social contact were examined: (1) marriage, (2) contacts with close friends and relatives, (3) church membership, and (4) informal and formal group associations. With few exceptions, respondents having each type of social tie had lower mortality rates than respondents lacking such connections. In separate multivariate analyses, three of the four sources of contact were found to be independent predictors of mortality. Only group membership did not contribute independently to mortality risk. In terms of interactions among sources of

contact, of particular interest was the number of possible relationships that place an individual in a particular risk category. For instance, people who were not married but who had many friends and relatives were found to have mortality rates equal to those who were married but who had fewer contacts with friends and relatives. Similarly, it did not seem important whether contacts were among friends or among relatives; only in the absence of either of these sources of contact was there a significant increase in the risk of death during the follow-up period. Tradeoffs and substitutions such as this ensured that about 60 percent of the sample, through one kind of contact or another, managed to maintain reasonably low mortality risk. It was only in the presence of mounting social disconnection, when individuals failed to have links in several different spheres of interaction, that mortality rates rose sharply.

To summarize the effects on mortality of increasing social isolation, a Social Network Index was constructed. Briefly, the Social Network Index considers not only the number of social ties but also their relative importance. Thus, intimate contacts are weighted more heavily than church affiliations and group memberships. Four network categories were developed to reflect differences in type and extent of social contact. The age-adjusted relative risks for those most isolated when compared with those having the most social contacts were 2.3 for men and 2.8 for women. The association between social ties and mortality was found to be independent of self-reported physical health status at the time of the 1965 survey, year of death, socioeconomic status, and health practices such as smoking, alcoholic beverage consumption, obesity, physical activity, and utilization of preventive health services, as well as a cumulative index of health practices and a range of psychological factors.

What is the role of social networks in the causation of disease?

Lin has noted that while it seems likely that the concept of social support will make a contribution to an epidemiological explanation of illness, it is also clear that we do not understand precisely how it fits into a model of disease causation. Several theoretical positions have been taken. One position with regard to the role that psychosocial factors play in the etiology of disease suggests that one set of psychosocial factors act as "stressors" that enhance disease susceptibility. The second set act as cushions or buffers that protect the organism from the effects of noxious stimuli. Social networks or supports have generally been placed in the latter category. (This classification has been, in part, based on the finding by Nuckolls, Cassel, and Kaplan [1972] that only when low psychosocial-asset scores were coupled with high life-change scores—the noxious stimulus—were there dramatic increases in risk of pregnancy complications.) A second theory does not categorize social supports as buffers but suggests that lack of social support, or social isolation, is a stressful circumstance in itself and capable of enhancing disease susceptibility. Data from the Lin, Dean, and Ensel study and from the Human Population Laboratory provide preliminary evidence that this latter position may be correct. However, this hypothesis will need to be rigorously tested in further investigations.

As a brief summary, I would like to suggest that we do the following:

1. Conceptualize social networks in a way that includes both emotional and instrumental contacts.
2. Define networks as a characteristic of the social environment having to do with ties between people and not as a catchall term for all psychosocial resources or problems, hassles, or other attitudes, feelings, or beliefs.
3. Use both behavioral and perceptual measures but attempt as best we can to differentiate between them.
4. Make few assumptions concerning where social networks fit into a model of disease causation but design studies to test the adequacy of various models.

Standardization of comparative health status measures: Using scales developed in America in an English-speaking country

Donald L. Patrick, St. Thomas's Hospital Medical School, University of London

The Highway Code in Great Britain advises the motorist on a dual carriageway to take extra caution when overtaking a large lorry, approaching a roundabout, or anticipating a stop at an open level crossing. The mechanic who services your English vehicle may refer to the wings, boot, petrol indicator, or bonnet and advise you to purchase a new model. To an American driver or mechanic, this advice might prove perplexing. Like the Highway Code and English mechanic, a manual on the content and use of standardized health status measures, if it existed, might well be a cautionary tale. The concepts and measures themselves might need adjustment depending on the setting in which they were to be applied. The model set of procedures to be used in applying the measures—for example, question wording, coding, or scaling methods—might require revision to meet the specific situation of the independent researcher.

The issues examined in this paper arise when an investigator wishes to use a scale or measure developed in one culture in another culture, either within one nation or between nations. Several questions concerning the standardization of content and method immediately confront the investigator. What are the objectives of using standardized measures in comparative studies? What are the problems and pitfalls in applying these measures? How might the investigation best contribute to the developing research tradition on comparative measures? In discussing these issues, I shall be concerned with measures of health status, in particular sociomedical measures, and the use of such measures in the study of chronic illness and disablement. As an illustration, I shall use our experience in England of adapting the Sickness Impact Profile (SIP) for a longitudinal study of disablement to investigate the role of social support in the physical, social, and emotional functioning of disabled persons.

Standardized measures of health status

Health research has probably attained a higher level of standardization than have many other branches of the social sciences. Standardization is evident in codified terminology, in the classification of causes of death, in epidemiological studies of disease, and in studies of health services utilization. A large number of standardized measures have been proposed for measuring the health status of individuals and populations at various levels of aggregation (Acheson, 1965; Berg, 1973; Elinson, 1976). Some of these measures have also been applied in a comparative framework—within one nation or between nations and in various combinations of micro- and macro-health studies (Holland, Ipsen, and Kostrzewski, 1979; Kohn and White, 1976). An ideal set of questions and methods rarely emerges, however, to meet a particular objective or series of objectives. Investigators have instead chosen an existing set of items, and procedures, or indeed have attempted to develop a new set that fits their particular research objectives and situation, modes of data collection, hypotheses, analytic plans, research budgets, funding agents, or convenience.

Aday and Andersen (1978) have suggested that more explicit use of criterion validity might aid in suggesting standards for health measures. For some health measures, criterion validity is a useful test. No clear criteria exist, however, for most sociomedical measures of health or illness constructs (Patrick and Elinson, 1979). Instead, problems of *defining* "health" have to be solved by construct validation. Construct validity involves specifying the factors or constructs that account for the variance in the proposed measures as well as the hypothesized relations among constructs. Evidence must be collected (for instance, mortality and morbidity data), and this evidence evaluated according to the hypotheses for the relationships between the constructs (for example, the hypothesized relationship between

morbidity and mortality). If different measures of the same construct are logically related and highly correlated, then convergent validity has been achieved; if a logically different measure is not as highly correlated as a more logically related measure, then discriminant validity is shown. Agreement among researchers on a proposed construct, consensual validation, may eventually follow construct validation (R.M. Kaplan, Bush, and Berry, 1976).

It is precisely this lack of clear criteria for health status that provides one impetus for standardized items in comparative studies. Like ecologists, health researchers wish to understand why there are so many ways of perceiving, reacting, and responding to disease, impairment, and their associated signs and symptoms and whether some ways are more evident in one environment than in another. Fabrega (1975) has urged an ethnomedical science to study how illness is defined by societies or smaller groups and the influences of cultural patterns. Such a science will be necessary unless one accepts that the diversity in health/illness phenomena is irreducible and that one can only record the diversity with systematic descriptive research methods. Equally, if we are to accumulate evidence about the factors influencing health status, then we must describe and explain the ways in which different cultures define and value health and illness. Again, like ecologists, social researchers will have to endure the tension between diversity and the search for repeated patterns and the regularities that may underlie all of the diversity.

One major objective of comparative studies of health status using standardized items, therefore, is to discover and explain real differences in the definitions of health and illness among different populations and cultures. Standardization maximizes progress in achieving this knowledge, and the chief way of being scientifically reasonable is to do whatever we can to maximize the progress of scientific research traditions (Laudan, 1977).

To date, the universal descriptors of health status have mainly fallen within the behavioral paradigm or the activities and tasks in which we engage routinely in our daily lives. Behavioral categories for codifying and measuring a person's social functioning have been easier to use than more subjective feeling states. Furthermore, the behavioral categories currently used have covered illness, dysfunctional, or sickness behaviors rather than more positive and preventive behaviors, because of the difficulties in specifying and operationalizing agreed-on positive dimensions of health and in testing the reliability and validity of positive health measures. The number of behavioral dimensions employed has varied, although many investigators have used variants of the six sociobiological categories of the activities of daily living: bathing, dressing, toileting, transfer, continence, and feeding (Katz and Akpom, 1976).

The second major objective of standardized health status measures follows from an understanding of the social and cultural content of health notions: to further the comparative examination of health systems and the many other social factors that affect the well-being of populations. The activities of health systems—prevention, education, diagnosis, treatment, prognosis, caring, and rehabilitation—must be compared cross-culturally in relation to health status if there is to be an equitable distribution of access or a better understanding of how health services fit into the link between economic, political, or cultural factors and population health status. Comparisons between economic development and well-being, in general, and between collective public health measures and health status, in particular, require standardized measures of health status, including more sophisticated use of mortality indicators. Elling (1974) and De Miguel (1974) have specified frameworks for the study of national health systems using health status indicators. More attention is now being paid to the study of developing countries and to non–English-speaking societies and groups. Comparative examinations of health systems also include regional differences within a country. Equitable allocation of health service resources among the regions of a nation may rely increasingly on proxy measures of need. In England, for example, a resource allocation formula is used to secure equal opportunity of access to health care for people at equal risk (Great Britain Resource Allocation Working Party, 1976). The allocation formula uses the standardized mortality ratio as a proxy indicator of morbidity and thus of regional differences in need when related to resource inputs. In such allocation formulas, standardized sociomedical health indicators may eventually replace or supplement mortality indicators in the subnational geographical redistribution of resources to remove inequity.

The final major objective of developing the standardized approach is to provide a methodologically tested data collection instrument that will save the time, effort, and money of the investigator and user of the information. Information exchange is a long and costly business, from deciding "what we need to know" and "what is available or practical to collect" to "what the message is" and "what use it is." Standardized measures can cut the cost at all

points of the information exchange process, particularly in the design of data coding and processing.

Problems and pitfalls in applying standardized measures of health status

Answers are necessarily limited by the questions asked and herein lies the major weakness of the standardized approach. A rigid application of set measures may not serve the purposes of either the investigator or the data user. Thus, the advantages of economy and comparability in standardization must be traded off against the purpose of the data collection enterprise. The choice of a health status measure will primarily depend on (1) the proposed use for the information — to compare populations, evaluate the interventions, or allocate resources; (2) the nature of the population to be studied — individual, community, region, or nation, and the language, medical condition, age, sex, religion, or other distinguishing characteristics of the respondents; (3) the known validity and reliability of the research instrument; and (4) the time, money, and manpower resources available to the investigator for administration, data collection, coding, and analysis.

As mentioned earlier, the actual number of behavioral descriptors of health status differs widely across measures and investigations. The selection of these descriptors can be a major obstacle to standardization. Measures based on routinely available data are limited to the categories of descriptors already in the information system, e.g., disability days in the Health Interview Survey. Health status indicators based on ad hoc surveys or data collection systems do not have this limitation. Nevertheless, many of these measures have emphasized physical functioning rather than dimensions of psychosocial and mental functioning. Some investigators have suggested the use of factor analysis to reduce the large number of possible descriptors on the basis of their ability to discriminate (Glasser and Forthofer, 1972). As R.M. Kaplan et al. (1976) point out, "such a procedure subtly substitutes variation in frequency for variation in social importance." Some behavioral descriptors may not occur often in the population under consideration and yet are important when they do occur, e.g., intravenous feeding at home. Thus, selecting the descriptors cannot be accomplished by data reduction techniques alone. A broad and inclusive number of descriptors are needed to cover the entire range from death to optimum well-being.

The selection of standard indicators and their classification in a conceptual framework will ultimately depend on the proposed use for the indicators. For example, the development of an international classification of impairment, disability, and handicap has been difficult, in part, because different groups of experts as well as different countries and agencies within countries use different definitions and indicators of these constructs, particularly in determining the benefit or compensation due to an impaired individual. Political disagreement about *who* gets *what kind* and *how much* compensation for *which* disabilities has made the selection of standard indicators for disablement a process of negotiation and compromise. Proposed users and uses for the classification system have continued to shape the selection and classification of the descriptors (Wood, 1975).

In our study of the Health and Care of the Physically Handicapped in Lambeth (Patrick, 1979), we have set out to describe and compare disabled and nondisabled respondents in a longitudinal health interview survey conducted in an Inner London Borough. We have tried to make comparisons on a comprehensive set of daily activities and tasks. Any resulting intervention in the population will be evaluated using the same set of behavioral descriptors. We chose to adopt the Sickness Impact Profile (SIP) designed at the University of Washington as our interview measure of disability (Bergner et al., 1976). The SIP comprises 136 statements, within 12 different categories of physical, social, and mental dysfunction, which describe a broad number of limitations or changes in behavior related to impairment and sickness. These behaviors were selected from a catalog of sickness behaviors obtained by self-report, observations, and existing measures and were reduced by item analysis to cover the range from maximal to minimal dysfunction. In our case, the SIP met the need for a comprehensive and tested measure to compare and evaluate individuals with varying types and degrees of dysfunction in a single community using a household interview survey. For other purposes, another scale or measure might be more appropriate.

Once the behavioral descriptors have been selected through the choice of a health status measure, a number of methodological issues arise in applying the measure in a comparative study. A major problem in cross-cultural studies is undoubtedly language, sometimes beginning with the label of the standardized measure itself. How we think and perceive depends to some extent on the structure of the languages we speak. Notions of health no doubt have some psycho- and socio-linguistic relativity, and a social grammar of health has yet to be written.

Although the SIP was developed in the United States in American English, translation

of the statements to conform to British English has been a formidable task, particularly because the apparent similarity often masks subtle but significant differences. British staff on the project translated the American items in the form of diagrammed sentences. Then I, as an American living in England, edited these sentences, keeping in mind the behavioral orientation and other objectives of the instrument. The items were then circulated to members of the local Lambeth community and the Department of Health and Social Security for further revision and translation into British English, in general, and into the local language of Lambeth residents, in particular.

A final conference was held in which project staff argued the exact wording of each item. For example, the SIP item, "I often act irritable toward my work associates, for example, snap at them, give sharp answers, criticize easily" was finally translated to "I often get irritable with my workmates, for example, I snap at them or criticize them easily." The final instrument was then sent to Seattle for review and confirmation that the item content for the British version (Functional Limitations Profile) was comparable to the original American version (Sickness Impact Profile).

The translation of standardized measures into different languages, particularly languages with varying expression of tense or voice of verbs that describe actions, or differing rules for nominalizing, e.g., "his" behavior, requires care to maintain the meaning of the original item while adapting it to the culture of the respondent. Bilingual interviewers are necessary when the cultural mix is high and there is no shared language.

Variation in the wording of questions is not only confined to problems of translation. Talcott Parsons (1958) defined personal health as "the state of optimum *capacity* of an individual for the effective performance of the roles and tasks for which he has been socialized." Investigators conducting health inquiries have varied their procedures in asking questions about *capacity* ("can" or "am able") or about *performance* ("do" or "do not" do an activity because of health). In an important paper, Anderson et al. (1978) have reported that this variation in wording may account for much of the widespread 15–20 percent underreporting of health-related dysfunction in well populations. Since performance is more readily observable and objective than capacity, perhaps the capacity mode of questioning should be abandoned or, at least, compared with responses obtained from the performance mode of questioning.

Adaptation of standardized indicators to the comparative research setting involves most other issues of reliability and validity in survey or measurement methodology. Two issues are of particular importance, however, to the comparative use of health status measures. The first is interviewer training and interviewer variability. Particular items in a standardized set may be influenced by interviewer effect. Different interviewers may behave differently on different sets of items. Thus, the contribution of interviewer variability to total survey error needs to be assessed in comparative studies to identify the potentially sensitive items and correct for them. This task is formidable because the proportion of response variation due to interviewer effect is difficult to estimate. Interviewer effect studies are also expensive to mount because of increased costs in interviewing, i.e., scheduling and traveling costs. In Lambeth we are applying and extending models for analyzing interviewer variability on our longitudinal data set on disability in hopes of contributing to the analysis and understanding of this phenomenon.

The second issue specifically related to survey methodology in comparative health status measurement is respondent burden and the use of proxy respondents. Respondent burden may vary across populations or cultures when using standardized questions. The response task itself, or items irrelevant to the respondent, may be more or less burdensome depending on the respondent, e.g., interviews with severely ill or well respondents. For example, in administering the translated SIP items in England, we have modified the task of the respondent during the interview. In previous administrations of the SIP, respondents have been asked to reply only when they hear an item that applies to them *and* is, in their opinion, related to their health. In England, we have asked for a response to each item, and where respondents indicate that an item applies to them, we probe to determine whether or not the behavioral change or description is health related. In some populations or cultures, greater use of proxy respondents may be necessary because of age, illness, language problems, comprehension of the respondent task, or acquiescence to the interview process itself. Such issues have to be considered in the context of specific data collection activities when adopting a standardized measure.

A final problem in applying standardized measures to comparative studies is the question of cross-cultural assessment of the severity or relative importance of the descriptors of dysfunction or disability. A composite profile or index requires the aggregation of the multiple

individual descriptors into summary scores. The investigator may impose his or her own value-weighting system onto combinations of descriptors (Great Britain Social Survey Division, 1971) or use data-analytic techniques to score items according to their frequency of occurrence (R. Williams, 1979). Explicit attention to the weights, social preferences, or measures of relative importance that members of the sociocultural group under study attach to descriptors of dysfunction seems preferable, and methods for obtaining these weights have been developed (Patrick, 1976). Indeed, the social and cultural content of health notions can be examined in comparative scaling studies to investigate the relativity/universality of the social values associated with states of health.

In Lambeth we are assessing the severity of dysfunctional behaviors in a comparative value-scaling study. The translated SIP items are being weighted by a British sample according to the perceived severity of the dysfunctions or limitations described in the statement. The British weights will then be compared with those obtained in America to test the social consensus in the two cultures.

Developing research on standardized comparative measures

A number of issues should be considered as we develop and use standardized health status measures cross-culturally. The international highway code should perhaps address the following issues:

1. *The match between proposed use for the measure and selection of indicators.* Agreement on the concepts of health status depends on the proposed use for the indicator. We need to develop an information matrix that suggests which indicators best fit the comparing, evaluating, and allocating purposes of the users.
2. *The cultural relativity of standardized items.* The translation of standardized items into different languages requires consideration of the specific cultural connotation of the items. We need to test the procedures for translating measures and for studying the psycho- and socio-linguistic relativity of items.
3. *Variation in modes of questioning.* The wording of standardized items and the experimental procedures used affect results. Experiments in wording questions are needed to identify preferred modes of questioning. Investigators developing measures should be encouraged by funding agents to provide manuals on their use to encourage standardization.
4. *Scaling of indicators.* The relative value of standardized health status items may differ across cultures. The extent to which social consensus exists regarding the severity of dysfunctional behaviors will come only through comparative studies.
5. *Sociology of survey research.* Barriers to the adoption of standardized measures is one topic suitable for a developing sociology of survey research. Such studies might contribute to the understanding of progress in social research and lead to solutions of the problems that impede our progress.

Discussion: Standardization of comparative health status measures

Marilyn Bergner, School of Public Health and Community Medicine, University of Washington

The principal concern of Patrick's paper is the transferability of measures developed in a particular place and a particular language to other places and languages. Generally, the issue is that of reliability and validity across population subgroups. Language differences complicate the issue and heighten the awareness of cross-cultural differences. I would like to add that reliability and validity across time also need to be considered in this framework. In longitudinal studies using the same measures, test-retest interactions are often noted as threats to validity (Campbell and Stanley, 1966). Less obvious are the changes that may occur to affect the interpretation and comparison of one-time surveys. Changes in language and usage may change one question into another. For example, ten years ago questions about gaiety or smoking may have provided different information than when the same questions are asked today.

Patrick may be correct in crediting health survey research with attaining higher levels of standardized measurement than other social sciences; I would note that the superiority may extend to the nonsocial sciences as well. Yet the fact that this topic is part of this conference is evidence of our dissatisfaction with our achievement.

Others have summarized the problems of the validity of measurement across subgroups and the particular problems of the interaction of language differences and response bias. I would commend an article by Tom Bice (1976) as a particularly good summary of the issues.

In the way of solutions, the road is not so clear. I would be the first to acknowledge and often insist that investigators use measures and instruments that have already been used. Reinventions of the wheel are not cost-effective. But when working with groups different from those on whom existing measures were developed, caution is the byword. Social groups, ethnic groups, and even specific patient or disease groups may not provide equally valid data.

One way of assessing validity, as Patrick notes, is provided by scaled measures. Rescaling furnishes a "natural" test of the social value of particular survey items, statements, opinions, etc. The Sickness Impact Profile provides an opportunity to assess validity in this way. Patrick is taking this opportunity to compare the English and American SIPs. I regret that there is no independent assessment of the translation equivalence of the two SIPs so that, if there are value differences, we could attribute them more accurately to culture, not to language.

Another, more complex assessment involves translation validation. Although validated translations of tests are rare enough from one language to another, the methodology suggested by Brislin, Lonner, and Thorndike (1973) for translating surveys for use in cross-national research could be generalized to rewording a survey so that it is valid across subgroups within a single language. Generally the method involves translation by bilinguals from a source language to a target language, the back translation by other bilinguals, a comparison of these back translations with the source language version, and some reconciliation of any differences. Tests of reliability should then follow.

My colleagues and I have used a variant of the Brislin et al. (1973) methodology to translate the SIP into Chicano Spanish. Briefly, we asked four bilingual Chicanos to translate the SIP. Four other bilinguals translated from Chicano to English; the result was 16 back translations. The 16 back translations were compared with the original English, and each item was rated in terms of its similarity to the original by three monolingual raters and three bilingual raters. The ratings were used to choose the most acceptable translations. Items that presented problems were reconciled at a conference attended by three staff members and all the bilinguals.

Next, reliability was assessed by having 31 bilinguals complete Chicano and English SIPs.

Test-retest and alpha reliabilities obtained were comparable to those obtained in reliability tests of the English version of the SIP. Finally, the Chicano version of the SIP was scaled by 29 Chicanos who considered Spanish their first language. The scale values obtained for the Chicano version were so highly correlated with the English version scale values that no change in item weights was deemed necessary. Thus, a reliable and valid translation of the SIP into Chicano Spanish is now available (Gilson et al., forthcoming).

In closing, I would like to return to Patrick's opening remarks. I would underscore his comments that the methodology for standardizing and validating measures depends on what you want to achieve, that the problems of standardization and translation pervade much of what we do, and that the solutions to the problems may raise more issues of comparability than we may have expected.

To bring us back to more pedestrian measurement considerations, I would also like to note that I will know that the millenium has arrived when I can walk into an American hardware store and purchase a nut and bolt that will fit my Peugeot; when I can buy fabric at Rodin's on the Champs-Élysées without having to frantically convert yards into meters; and when I know instinctively that if I was having my VW Rabbit repaired by a French auto mechanic, it would be unwise, at best, to ask that he fix *mon lapin.*

Open discussion: Session 4

Psychiatric screening scales

In opening the general discussion on the paper by Dohrenwend and others, Dean commented that the paper depended on judgments of psychiatrists, and he questioned the validity and reliability of their judgments. Barbara Dohrenwend replied that if a test purporting to measure psychopathology is unrelated to what psychiatrists consider psychopathology, a necessary but insufficient evaluation of it has been made. If, in psychiatrists' judgments, a test was related to psychopathology, further investigation would be needed. Unless constrained by data base and decision rules, psychiatrists' ratings are unreliable and their validity is questionable. These ratings would have to be subjected to a number of alternative hypotheses—for example, genetic or stress—and generally put in a nomological net and subjected to construct validation.

de la Puente commented that he had a different problem, namely, different interpretations of similar instruments according to the populations studied. Some instruments yield many more false positive diagnoses of mental illness than others. He recommended that studies of the reliability of diagnoses be undertaken. Dohrenwend replied that extensive work is being done in the U.S. and England on reproducibility of diagnoses. She cited the Biometrics Section of the Psychiatric Institute of Columbia University as a place where such research is taking place. In England the Medical Research Council Unit at Maudsley Hospital under John Wing is conducting similar research. Clinicians have advanced quite far in developing standard data bases and decision rules for diagnoses for research purposes. The number of false positives is a function of decision rules. The instruments, however, have been used in many varied populations rather than in patient populations.

Pope suggested that the tests may be screening for impaired social functioning. Many people show up in mental health clinics because they are functioning inadequately but are not mentally ill. He wondered whether the scales that Dohrenwend tested would provide measurements of that impairment. In reply Dohrenwend said that it is as (or more) difficult to screen for impaired social functioning as for mental illness and that the tests evaluated in the study are not good for measuring impaired social functioning either. Impaired social functioning is very complex, and mental illness is one, but only one, domain of it. (For example, people may be functioning badly because of a bad boss or spouse with whom no one can get along or because they are in an unfavorable environment.) Pope felt, nevertheless, that scales measuring demoralization were valuable for scales of mental illness. Dohrenwend agreed that social functioning is part of the picture but taps many other domains than mental health.

Berkman noted Dohrenwend's hesitancy to call measures of demoralization mental illness or to use them as screening devices or measures of psychopathology. But having developed a scale, she wondered whether, nevertheless, the test has value or is an important predictor of any other pathology in which researchers are interested.

Dohrenwend replied that—for use in community surveys—the test has no value unless there is interest in rates of demoralization in the community. If rates of mental illness are desired, it will be necessary first to recognize that it will require multidimensional scaling and then to decide how many dimensions and which dimensions to screen for. Many people are focusing on depression as an aspect of mental illness. Other scales measure depression as distinguished from demoralization and focus on insomnia, suicidal ideations, enervation, and guilt feelings. A review of studies of demoralization and clinical estimates of mental illness provides rough estimates of 25–30 percent demoralized, of whom at least half are not mentally ill. Fif-

teen percent have problems relating to mental illness, and some of them are not demoralized. On the basis of such evidence, she concluded that the demoralization scale should not be used as a proxy for mental illness measurement.

Harold Dupuy felt that Dohrenwend loaded the dice against herself and made an unfair test of the measures that she was evaluating. He pointed out that the data showed that structured interview scales (SIS) discriminated between people in the community and mental hospital patients. He also asked why the study started with an attempt to measure neurosis. Dohrenwend thought that ability to discriminate was an inadequate test. The scales that distinguish between psychiatric patients and persons in the community would also discriminate between chronically physically ill patients, recent widows or widowers, and the general community. The point is that the tests are *not specific* to psychiatric disorders.

In answer to a question about why the research started with efforts to measure neurosis, Dohrenwend explained that, based on their clinical judgments, psychiatrists grouped the scale items into categories and hypothesized that these would measure neuroses. Researchers then looked for evidence about whether they would do so, independently of psychiatric diagnoses. She considered that to be a reasonable test of whether the SIS measured a diagnostic category of mental illness. Clinicians agree that, on the face of it, anxiety, sadness, and similar reactions are indications of neurosis.

Andrews reported that in his work he repeatedly found better and fuller reporting by more highly educated people than by less well educated people; he cited Yaffe and Shapiro's paper, which showed results similar to his. He wondered whether Dohrenwend had any comments about why her research findings showed the opposite results—i.e., fuller reporting by low-education respondents.

Dohrenwend disagreed and thought that the differences by education in her tables did not seem to be sufficiently large or consistent to be of interest. The five scales that she used do not consistently show the highest educated to have the lowest alpha coefficient, except for a number of physical symptoms that are consistently more often reported by less well educated people.

Morton Israel asked whether there was a theoretical framework for any of the concepts measured in the SIS scales and suggested that they might come from some of the constructs, such as sense of identity, self-esteem, or others. Dohrenwend said that the scales did pick up self-esteem, and she wondered whether Israel was interested in that. In a scale developed later than the ones reported on, a measure of self-esteem was developed that turned out to be on the same dimension as demoralization.

Social support scales

The general discussion on Lin's paper and Berkman's comments began with a remark by Elinson on the rarity of a nine-year follow-up of a study.

Sudman expressed a serious concern about which are the causes and which the effects—whether, for example, illness contributes to social isolation or social isolation to illness; whether illness prevents marriage or affects other demographic, social, and economic characteristics. He felt that only panel or longitudinal studies could measure the variables. Elinson pointed out that the Alameda study *was* a follow-up or longitudinal study.

McDowell asked for advice about evaluating social networks. What is significant is that they be available when needed; but frequency and quality of contacts, as well as availability or ease of making them, are important in unknown ways. Lin said that he could not answer the questions but believed that it is important to separate the content and structure of networks. He went on to say that he disagreed with Berkman that research should focus on the network itself. Networks are useful only to identify the right resources. The ultimate factor is the kind of resources you reach rather than the way you find them. It is, however, premature to decide to focus on one or another approach; both should be explored.

As an example of the function of social support, Berkman, in response to a comment by McDowell, said that not everybody has to know a lawyer. You only need to know a lawyer if you are in trouble and are going to jail. Then it becomes imperative that you find one, and that is the critical part of support. It is not knowing everyone all of the time but having the capacity to deal with the issues that is important.

Berkman also commented on Sudman's point, agreeing with him that there is much confounding. In a study done at the Alameda Human Population Laboratory, researchers looked at health status in terms of the number of chronic conditions, symptoms, and disabilities at the time of the survey in 1965 and controlled for that. You could see a network gradient among different categories in the physical health spectrum. You could even see that social isolation predicated mortality among people who reported being in very good health—not even reporting a single symptom or complaint.

Of course, mortality rates were low for the group as a whole but differed between those who were socially isolated and those who were not. The second way of checking on confounding was to examine year of death. The mortality gradient stretched over the entire nine years and was not observed in the first two years. There were no physical examinations so they relied solely on self-reports and did not know how sick people really were.

Yaffe wondered to what extent professional special support by psychiatrists, psychologists, social workers, and other therapists is a factor. From the point of view of health status and health services utilization, can the professions improve the social support system? Lin replied that his scales did incorporate some items describing utilization of formal health groups in the community. Unfortunately, they did not work, probably because the limited sample did not provide enough incidence. There were not enough data points to make the items work. The items are being retained, and the search will continue.

Berkman said the same thing happened in the survey that she reported about, that is, there was a variety of questions about professional and formal health groups, but there were too few users of them to provide reliable statistics. She added, responding to another comment by Yaffe, that health care questions about whether respondents went to a doctor or dentist did not explain much either.

Patrick described a study that he is conducting in Britain which specifically addresses the question of the extent to which formal services substitute for informal services for the disabled. He is trying to distinguish between formal services, informal services, and network frequency of contact and meaning of the contact. He agrees that researchers should try to keep the structure or network content separate from the meaning content, although he would not want to see one measure without the other.

Bergner agreed and said that unless you can keep the concepts separate, you cannot address the policy issues of whether provision, in a more community-related fashion, of support systems would ameliorate the need for medical care. They can only be addressed if network issues are separated from other quantitative and qualitative issues.

Pope commented that studies have found networks to be related to diseases with high emotional components. It appeared to him that high users of services are using networks for support. (Lack of relationships in the past deprive other people of networks now.) He thought much more clarification was needed.

Dean agreed with Sudman's earlier comment and concluded that conceptually and operationally research is at an early stage of development. He is trying to do a panel study to see how the same subjects deal with stressful life events.

Singer asked Berkman whether she had controlled for age and sociodemographic variables when she made predictions based on network relationships. In reply Berkman said that all of the analyses were age adjusted. She also looked at age-specific mortality rates. Analyses were confirmed for every age group. Analyses were also done separately by sex and by socioeconomic status (income and education). She also controlled for every level of physical status and for every level of health practice.

Standardization of health status measures

McDowell began the general discussion on Patrick's paper with the recommendation that, as a group, mental health researchers undergo the discipline of adhering to common standards, explain their conceptual framework to other potential users, and be more helpful to one another. He thought that they should follow the American Psychological Association guidelines for tests and presentations and should provide more structuring of their validation papers in more standardized fashion. Bergner agreed wholeheartedly and felt that the problem has to do with resources. If many investigators become interested in your method and you are fortunate in having them replicate your uses, you are unfortunate in having to provide a good deal of advice and consultation by phone—hours of it. There is a problem in reconciling competing demands for time. Elinson commented that Dupuy has spent so much time consulting that he has not had time to publish results of his research.

Jones noted that in Canada they must do research in English and in French. One of the reasons they did not use the Foldberg Scale was that when they translated the French back to English, two items came back identically worded. Patrick said that it would be difficult to do comparisons between American and English. Few consider themselves bilingual. He thought that there was a translation problem between New York and New Mexico American language. Verbal structures and nominalizing in American Indian make it almost impossible to translate something like the SIP to it.

In response to a question by Axelrod on the problems of translating among English, Irish, Scottish, and Welsh, McDowell said that you have to see whether the concepts exist in the other culture. That was the first step in trans-

lating semantic differential questions when researchers did not understand whether the problem lay in the culture or the language.

Bergner said that if you do not reword the question, you will never know if the problem is in the language. If you reword it and have no way of assessing the similarity of your translation—as has happened here—you end by not knowing at all what you have, assuming there are some differences. If there were an independent measure of the equivalent of the translation, you could feel a little less concerned and judge that what you had was purely a language problem.

McDowell recommended that less emphasis be put on a literal word-for-word translation and more attention be paid to concepts. Elinson said that he once tried to get Bradburn's happiness questions translated into Puerto Rican Spanish and the leading poet in Puerto Rico gave him the translations. His translation of "In the last week or so, were you feeling on top of the world?" was "feeling like a newborn baby." From some teenagers who were helping on the study, he got, "feeling like being in mashed potatoes."

Miller commented that some concepts may not have salience in another culture. It is not a translation problem. In developing and Third World countries, they accuse us of ramming our culture through their language. Patrick agreed and said that it is happening to him in England, where they accuse him of pushing American concepts and versions. Bergner also agreed and returned to McDowell's earlier remarks, saying that she was concerned about and wants to measure culture differences but does not want them contaminated by semantic differences.

Recommendations

The following recommendations emerged from the open discussion in this session:

1. Additional studies of reliability of diagnoses are needed, particularly in the area of mental health.
2. Mental health researchers should follow APA guidelines for tests and presentations and should provide more validation of methods.
3. In cross-cultural health research, more attention should be paid to concepts and less to literal translation.

SESSION 5: Methodological implications of the National Medical Care Expenditure Survey

Chair: Daniel G. Horvitz, Research Triangle Institute

Recorder: Mimi Holt, Research Triangle Institute

The use of Summaries of previously reported interview data in the National Medical Care Expenditure Survey: A comparison of questionnaire and Summary data for medical provider visits*

Mimi Holt, Research Triangle Institute

Preface

The data presented in this paper were originally tabulated for the National Medical Care Expenditure Survey—Methods and Analysis project conducted for the National Center for Health Statistics (NCHS) by Research Triangle Institute (RTI). The purpose of the Methods and Analysis study was to address a number of operational, procedural, and methodological issues of relevance to the design of the National Medical Care Utilization and Expenditure Survey (NMCUES) by analyzing data and experience gained in the National Medical Care Expenditure Survey (NMCES). The NMCES is a joint project of the National Center for Health Services Research and NCHS involving a panel survey of a national sample of families on their health care utilization and expenditures during 1977. The NMCUES is a joint project of NCHS and the Health Care Financing Administration; health care utilization and expenditure data for a national sample of families will be collected in a panel survey in 1980.

One of the research issues included in the project was the use of the computer-generated Summaries of previously reported interview data, which were sent to interviewers and respondents for review at each NMCES interview after the first. Of particular interest was an evaluation of the Summary as a means of improving the quality of the originally reported data. The comparative analysis presented in this paper was designed and executed in this context. It should be noted that the matching operation described here was based on manually coded entries from hard-copy documents (questionnaire and Summary) rather than computer matches of data in the respective questionnaire and Summary files. This approach was taken for a small sample of visits because of schedule and budget constraints on computer matching on such a small scale and because the final NMCES data base construction activities are ongoing. Provision has been made in the NMCES data files for more extensive and sophisticated direct matching operations based on formal data linkages. It should also be noted that variances may occur between the results of the manual matching described here and the contemplated computer matching operations for NMCES analysis files.

The purpose of this paper is to present a preliminary indication of the apparent utility of the computer-generated Summaries.

Introduction

The design of the National Medical Care Expenditure Survey (NMCES) required the production of cumulative computer Summaries of certain data reported in the initial and subsequent interviews. Two copies of the Summary were produced—one was sent to the interviewer assigned to the case and the other was sent to the family. The computer-generated Summaries used in NMCES had three basic purposes:

1. The Summaries allowed interviewers and respondents to review data reported in previous interviews and to add to, change, or delete incorrect or incomplete reports of medical care utilization and expenditures.
2. The Summaries provided necessary cross-round continuity for such data as flat fees[1] and health insurance coverage in effect.
3. The Summaries aided the bounded recall period approach to collecting data by providing a cumulative record of all previously reported visits and services.

The NMCES Summary was viewed as a means of maintaining and improving the quality of data collected during five interviews covering calendar year 1977. In theory, information not known to respondents at the time of a previous interview might have become known through

* Work supported by NCHS Contract No. 233-78-2102, "National Medical Care Expenditure Survey—Methods and Analysis."

receipt of provider bills, statements from third-party payers, and the like. Also, utilization events not reported during a previous interview might have been recalled in a retrospective appraisal of what data were initially reported. During the Summary review at each interview round, interviewers were instructed to probe for information previously unknown and to ask if utilization and cost data reported earlier were, in fact, correct and complete.

The Summary concept is clearly an innovation in survey research methodology, particularly when used in a survey with the magnitude and complexity of NMCES. During a given interview, the interviewer used the computer Summary as a data collection instrument. Based on responses to general and specific probes about data contained on the Summary, interviewers corrected, deleted, or added details of medical care received during dental, hospital, and medical provider visits or expenses incurred for items such as eyeglasses and medical appliances. Changes were also made in health insurance policies reported to be in effect and in individuals reported as covered by those policies.

From statistics generated during the coding of Summary changes at RTI, it is clear that a large number of changes were made on the Summaries. Over 445,000 changes were coded for updating the Summary data file. These changes included substituting a new data element for a previous one, adding a data element to an entry, or deleting a data element from an entry.

It is useful to review the *frequency* of changes in the context of the instructions given to the interviewers during each survey round. In the four rounds during which the Summary was reviewed, two overall review procedures were required at all times; several other procedures were specified during one or more rounds. The two required procedures were (1) a specific query for all data not known or missing from previous reports and (2) a general probe for each type of visit or expense to determine if all visits or services for each family member had been reported. Special review procedures that were requested in one or more rounds included (1) questions to secure complete, consistently spelled medical provider names and addresses; (2) probes to confirm that visits or services reported as free from the provider were, in fact, free; (3) instructions for correcting errors resulting from keying and processing problems; and (4) a final major reconciliation of all charges, sources of payment, and amounts paid by all payers.

It should not be surprising to see that the *types* of changes made in each interview round have a direct relationship to the review specifications given to the interviewers; that is, when a specific data element was identified for special treatment, the frequency of changes to the data element increased. Table 1 summarizes, for Rounds 2 through 5, the interviewer instructions for Summary review and the related frequencies of changes requiring updates to the Summary file.

A few observations are made regarding the display in Table 1. While the later Summaries—Rounds 4 and 5—clearly contained more data than the earlier ones, the percent increases in frequencies of change appear to reflect purposeful changes based on interviewer review specifications. That is, one can speculate that the *rates of change* should have increased only slightly as more data were added to the Summary. In theory, data elements corrected in Round 2 should not have required additional changes in all subsequent rounds. However, since the Summary review specifications were modified from round to round, data elements that were considered acceptable in an earlier round were scrutinized in a different light in later rounds and, in many cases, changed.

Consider the percent increase in frequency of changes in four key update categories (provider name, provider address, service types, and payers)[2] across the four interview rounds, keeping in mind the degree of emphasis placed on each update type in the interviewer instructions. Table 2 shows the percent increase of change from the preceding round.

It is clear, both from pure frequencies of change and from relative or percent increases of change, that changes in interviewer specifications from round to round led to higher numbers of changes than might otherwise have been expected. One could reasonably argue that the frequency and percent increases of change should have increased only slightly or remained about the same from round to round if the Summary review specifications had remained constant, since both the interviewers and the respondents became more efficient at their respective tasks. The Summary update statistics, however, reflect dramatic increases in the number of changes made, culminating in the large number of changes made in Round 5, where the most detailed review of the Summary occurred. Almost 63 percent of all Summary changes were made in Round 5.

Unfortunately, the *frequencies* of change do not provide any indication of the *nature* and *direction* of changes that were made on the Summary. A change, per se, by no means implied better or more acceptable data. A change, as measured by the update frequencies, could have meant that (1) an unknown response was

Table 1
Summary updates by type in relation to interviewer instructions

Round	Interviewer instructions	Types of updates: Person name	Provider name	Provider address	Service types	Payers	Flat fee	Insurance policies	Persons covered	Refund	Total updates per round
2 (personal interview)	Item-by-item review of every entry Probes to determine if all visits and services were reported Specific queries about unknown or missing data from previous round	1,917	6,729	2,476	3,626	16,887	965	8,927	6,444	—	47,971
13,258 Summaries											(3.6 per Summary)
3 (telephone interview)	Page-by-page review Probes to determine if all visits and services were reported Specific queries about unknown or missing data from previous rounds	2,505	6,990	2,976	3,201	18,361	824	2,032	1,438	121	38,448
12,745 Summaries											
4 (telephone interview)	Page-by-page review Probes to determine if all visits and services were reported Specific queries about unknown or missing data from previous rounds Confirmation of all reports of visits/services as being "free from provider" Attempts to secure complete, consistently spelled provider names and addresses	742	24,124	8,138	5,293	37,213	1,431	795	736	216	78,688
9,514 Summaries											
5 (personal interview)	Item-by-item review of every entry Probes to determine if all visits and services were reported Specific queries about unknown or missing data from previous rounds Confirmation of all reports of visits/services as being "free from provider" Attempts to secure complete, consistently spelled provider names and addresses Reconciliation of all charges and sources and amounts of payment	966	83,627	18,255	26,477	134,785	8,650	3,927	2,909	424	280,020
13,497 Summaries											(20.7 per Summary)
Total updates		6,130	121,470	31,845	38,597	207,246	11,870	15,681	11,527	761	445,127

changed to a known value, (2) a known value was changed to an unknown response, or (3) an essentially editorial change was made that had no impact on the overall quality of the data. For the latter category of change, "Main" may have been changed to "Maine," or "Dr. E. L. Banks" may have been changed to "Dr. Eugene L. Banks." Hence, the outcome of change could be *positive* (better data), *negative* (lower quality or loss of data), or *neutral* (no apparent difference in overall data quality).

The method of comparison of questionnaire and Summary data

In an attempt to measure the nature and direction of changes made on the Summary, a comparison coding operation was designed. A random systematic sample of 450 visits was drawn from the medical provider visits reported in Round 1. The visit record from the main data base was printed out in its entirety so that the corresponding visit could be located in the hard-copy Round 1 questionnaire. The questionnaire report of the visit was then compared with the visit entry on the final Summary generated for respondent and interviewer review. In the majority of cases, the final Summary was the one reviewed in Round 5.

In examining the questionnaire and the Summary at the same time, the coders first located the sample visit in the questionnaire. Once the visit had been identified with certainty in the questionnaire, they next searched the Summary for the visit record. Exhibit 1 illustrates medical provider visit data as they were initially recorded in the questionnaire; Exhibit 2 shows how the same visit data were displayed on the Summary.

The document-matching operation was confounded by two factors: (1) after almost two years of storage, a number of documents could not be located; and (2) additional documents, primarily Round 5 Summaries, were being used by other staff members for final data base corrections.

Table 2
Percent increases of Summary changes

Round	Update category			
	Provider name	*Provider address*	*Service types*	*Payers*
2	—	—	—	—
3	3.9	20.2	−11.7	8.7
4	245.1	173.4	65.3	102.7
5	246.7	124.3	400.2	262.2
Percent increase from Round 2 to Round 5	1,142.8	637.3	630.2	698.2

Consequently, the document-matching operation produced the results shown in Table 3.

The 409 visits for which both required documents were available were then coded. The purpose of the comparison coding was to identify the similarities and/or differences in common data elements between the initial questionnaire report and the final Summary entry for the visit.

The first step in the coding process was to determine if the visit could, in fact, be matched with certainty between the questionnaire and the Summary. Of the 409 visits that could be coded, 368, or almost 90 percent, could be matched with certainty. In literal terms, this match meant that on at least two of three critical elements, the two visit reports were identical. The three critical data elements were provider name, date of visit, and charge(s) associated with the visit. No match with certainty could be made for 11 of the 409 visits, or 2.7 percent. In most of these cases, the visits could not be matched because critical data elements were originally reported as unknown in the questionnaire. The remaining 30 visits (7.3 percent) were found in the Round 1 questionnaire but not on the Summary. Of the 30 visits that were not found on the Summary, 18 visits had been recorded on Medical Provider Visit continuation pages.[3] A documented data processing problem apparently prevented these visits from ever appearing on the Summary for review by the respondent and interviewer. It should be noted, however, that these visits have been successfully merged into the final Summary data file. The other 12 visits that could not be located on the Summary apparently were deleted from the Summary by respondents and interviewers prior to the final review. No attempt was made to track the sample visits through all interview rounds to determine what happened to them between the initial questionnaire report and the final Summary.

After the questionnaire and Summary reports of a visit were matched with certainty, the common data elements were examined and coded. The coding scheme involved the coding either of actual values (e.g., dates and dollar amounts) or of relationships between the common data elements (e.g., variations in provider names and total charge responses). The codes also identified situations where a data element appeared on one document but not on the other. The coding operation was performed by RTI Data Services staff members who had extensive experience in NMCES Summary update coding. All coding was checked by the author. The coded data were keyed and then completely re-

Total updates ..... 6,130 121,470 31,845 38,597 207,246 11,870 15,681 11,527 761 445,127 (20.7 per Summary)

Exhibit 1
Medical Provider Visit section from NMCES round 1 questionnaire

MEDICAL PROVIDER VISIT

Person # 2 First Name: Jane

PREPARE A SEPARATE PROVIDER VISIT PAGE FOR EACH VISIT REPORTED IN PROVIDER PROBES. FOR EACH PERSON WITH VISITS IN PROVIDER MP-20 & MP-22 BOXES, DETERMINE AND ENTER FIRST/NEXT TYPE OF PROVIDER, PLACE, OR SERVICE FROM PAGE MP-20 OR PAGE MP-22.

P

Provider: Medical doctor

Service: _____

Place: _____

You told me earlier that (PERSON) had [seen or talked to (PROVIDER)/(SERVICE)/gone to a (PLACE)] (NUMBER) times since (REF. DATE).

1. On what date during that period did (PERSON) (first/next)

visit or talk to (PROVIDER)?
have (SERVICE)?
go to (PLACE)?

02 / 18
Month / Date

2. What is the name of the medical person (PERSON) saw or talked to on (DATE)?

Name: Dr. R.S. French

Don't know 94

A. In what city and state is (MEDICAL PERSON) located?

Raleigh / NC
City / State

IF ANY KIND OF NURSE, ASK B & C.

B. What is the name of the doctor (NAME OF NURSE) works for?

Name: _____

Don't know . . . . . 94(3)

C. In what city and state is the doctor located?

_____ / _____
City / State

IF PLACE ENTERED IN "P", CODE Q.3 WITHOUT ASKING.

3. Where did (PERSON) [see or talk to (PROVIDER)/GET (SERVICE)]--at a clinic, hospital, doctor's office, laboratory, or some other place?

IF HOSPITAL, ASK:

A. Was it the outpatient clinic or the emergency room?

IF CLINIC, ASK:

B. Was it a hospital outpatient clinic, a company clinic, or some other kind of clinic?

While inpatient in hospital. 01(STOP)
Doctor's office (group practice or doctor's clinic). . (02)(5)
Telephone. . . . . 03(5)
Home . . . . . 04(5)
Hospital outpatient clinic . 05
Hospital emergency room. . . 06
Company/Industry clinic. . . 07
Laboratory 08
Other (SPECIFY) _____ 09

4. What is the name of this place? IF NAME NOT KNOWN, TRY TO GET ADDRESS OR OTHER IDENTIFYING DATA.

Name: _____

A. In what city and state is it located?

_____ / _____
City / State

5. For what condition did (PERSON) [see or talk to (PROVIDER)/have (SERVICE)/go to (PLACE)] on (DATE)? Any other condition?

Checkup. . . . . 01(Re-ask 5)
No condition . . . . . 02(A)

Condition		Cond.#
Pain in left side	CC	02 (6)
	CC	(6)

A. What was the reason for the [visit/telephone call]? Any other reason?

CODE ALL THAT APPLY

Pre or post natal exam . . . 01(CC)#
General physical exam/checkup 02
Eye exam for glasses 03
Diagnostic tests 04
Shots/immunization 05
Hearing test 06
Counseling 07
Prescription 08
Other. 09

Don't know . . . . . 94(3)

C. In what city and state is the doctor located?

City / State

[illegible] post natal exam	01(CC)#	Hearing test	06
General physical exam/checkup	02	Counseling	07
Eye exam for glasses	03	Prescription	08
Diagnostic tests	04	Other	09
Shots/immunization	05		

IF "TELEPHONE" IN Q. 3, SKIP TO Q. 9.

IF X-RAYS, LABORATORY TESTS, OR DIAGNOSTIC PROCEDURES IN "P", CODE "Yes" IN APPROPRIATE QUESTIONS(S) 6 THROUGH 8 WITHOUT ASKING. ASK ALL QUESTIONS NOT CODED.

	Yes	No
6. Were any X-rays taken during this visit on (DATE)?	(01)	02
7. Were any laboratory tests such as a blood test, urinalysis, culture, or other kind of test done as a part of this visit?	01	(02)
8. Was an EKG, EEG, (a pap smear)or any other diagnostic procedure done on this visit?	01	(02)

9. How much was the total charge for this [visit/telephone call] on (DATE) including any amounts that may be paid by health insurance or other sources? (Include any separate bill for [use of emergency room/X-rays/laboratory tests/diagnostic procedures].)

$ 47.00 *(11)*

No charge . . . . . 01*(10)*

Included in charges for other visits. . . . . 02*(FF__)*

Don't know. . . . . 94*(A)*

A. Do you expect to receive a bill for this visit?

Yes . . . . . 01*(18)*

No. . . . . 02*(B)*

B. Why don't you expect to receive a bill?

Free from provider. . . . . 01*(18)*

Included in charges for other visits. . . . . 02*(FF__)*

Other source(s) will pay. . 03*(11)*

10. Why was there no charge for this visit?

Free from provider. . . . . 01*(18)*

Included in charges for other visits. . . . . 02*(FF__)*

Other source(s) will pay. . 03*(17)*

ASK Q's. 11, 12 & 13 BEFORE ASKING 14 THROUGH 17.

COMPLETE QUESTIONS 14 THROUGH 17 IN ONE COLUMN BEFORE GOING TO NEXT APPLICABLE COLUMN.

	11.	12.	13.
	IF VISIT WAS IN EMERGENCY ROOM, ASK: Was any of (CHARGE IN Q.9) identified as a separate charge for the use of the emergency room? Yes . . 01 (A) No. . . 02 (12) A. How much was that? Charge $_____ (12)	IF X-RAYS, LAB TESTS DIAGNOSTIC TESTS ON THIS VISIT, ASK: Was any of (CHARGE IN Q.9) identified as a separate charge for [X-rays/tests]? Yes . . (01) (A) No. . . 02 (13) A. How much was that? Charge $ 32.00 (13)	ENTER CHARGES A)- Q. 9 [47.00] B)- Q.11A [ ] C)- Q.12A [32.00] D)- 11A PLUS 12A [32.00] E)- A MINUS D [15.00] Zero amount. . . . 00 (14) Charge for visit $ 15.00 (14)
14. How much of the (CHARGE) for the [SERVICE/VISIT] did you or your family already pay?	Partial $_____ All. . . . . 01 (16) None 00	Partial $_____ All. . . . . 01 (16) None . . . . . (00)	Partial $_____ All. . . . . (01) (16) None 00
15. How much (more) of this charge will you or your family pay?	Partial $_____ All. . . . . 01 None . . . . . 00 D.K. 94	Partial $_____ All. . . . . 01 None . . . . . 00 D.K. . . . . (94)	Partial $_____ All. . . . . 01 None . . . . . 00 D.K. 94
IF FAMILY PAID /WILL PAY ANY AMOUNT (Q's. 14 OR 15), ASK: 16. Do you expect any source to reimburse you or pay you back?	Yes. . . . . 01 (A) No 02	Yes. . . . . (01) (A) No 02	Yes. . . . . 01 (A) No . . . . . (02)
A. Who will reimburse or pay you back? ENTER UNDER "SOURCE." Anyone else? B. How much will (EACH SOURCE) reimburse or pay you back?	SOURCE / AMOUNT	SOURCE / AMOUNT	SOURCE / AMOUNT
	_____ %	BC/BS of NC 80 %	_____ %
	_____ %	_____ %	_____ %
	_____ %	_____ %	_____ %
IF "ALL" IN Q's. 14 OR 15, GO TO NEXT COLUMN OR Q. 18.	No other source . . 01 (Q.14, next col.)	No other source . . (01) (Q.14, next col.)	No other source . . 01 (Q.18)
17. Who else paid or will pay any part of the charge for the [SERVICE/VISIT]? ENTER UNDER "SOURCE." Anyone else? A. How much will or did (EACH SOURCE) pay?	SOURCE / AMOUNT	SOURCE / AMOUNT	SOURCE / AMOUNT
	_____ %	_____ %	_____ %
	_____ %	_____ %	_____ %
	_____ %	_____ %	_____ %

IF RESPONDENT VOLUNTEERED THAT THIS VISIT IS ONE

IF TELEPHONE, GO TO RV.

18. For this particular visit, did (PERSON) have an appointment or just walk in?

Appointment . . . . (01)(A)
Walk in . . . . . 02(19)

A. Did the (PROVIDER) tell (PERSON) when to come in during an earlier visit or did (PERSON) just call up for an appointment?

Set by provider . . 01(19)
Patient called. . . (02)(B)
Other . . . . . 03(19)

B. About how long did (PERSON) have to wait to see the (PROVIDER) after making the appointment--how many days, weeks, or months?

[2] days [ ] weeks [ ] months

IF MORE THAN ONE DAY, ASK C:

C. Was this wait longer than (PERSON) thought it should be?

Yes 01
No. . . . . (02)
Don't know. 94

19. About how long did (PERSON) have to wait before seeing the (PROVIDER) after (PERSON) arrived at (PLACE)--about how many minutes or hours?

[20] minutes [ ] hours

20. Was this wait longer than (PERSON) thought it should be?

Yes 01
No. . . . . (02)
Don't know. 94

RV	IF RESPONDENT VOLUNTEERED THAT THIS VISIT IS ONE OF REPEATED VISITS BY (PERSON) TO SAME (PROVIDER/PLACE), ASK Q's. 1 THROUGH 5.

You mentioned that (PERSON) had (NUMBER) [visits to this (PROVIDER/PLACE)/(SERVICE)].

1. Of these [visits/telephone calls], how many were for the same condition as the one we just talked about?

_____visits/calls

None . . . . . .00(NV)

2. Of the (NUMBER IN Q.1) [visits/telephone calls], how many cost the identical amount as the [visit/telephone call] we just talked about?

_____visits/calls (3)

#____visits included in same FF____. . . . . (3)

None . . . . . .00(NV)

3. What were the dates of these (NUMBER IN Q.2) visits (telephone calls)?

_____ _____ _____ _____
Month/Date Month/Date Month/Date Month/Date

_____ _____ _____ _____
Month/Date Month/Date Month/Date Month/Date

IF TELEPHONE, SKIP TO INSTRUCTIONS ON BOTTOM OF PAGE.

4. Did you have appointments for all these visits?

Yes. 01
No 02

5. About how long did (PERSON) usually have to wait before seeing the (PROVIDER) after (PERSON) arrived at (PLACE)--about how many minutes or hours?

[ ] minutes [ ] hours

CONTINUE WITH ANY ADDITIONAL VISITS NOT COVERED HERE. IF THIS IS LAST PROVIDER VISIT FOR ENTIRE REPORTING UNIT, GO TO RECONCILIATION PAGE, R-25.

Exhibit 2
Summary entry for Medical Provider Visit

The first entry below represents the questionnaire data as they would have appeared on the first Summary generated after the report of the visit. The second entry represents the visit data as they might have been updated by the respondent at a later point in time.

RU # 1000018
FOR JANE CURTAIN

MEDICAL CARE AND EXPENSES SUMMARY
FROM 01/01/77 to 03/18/77

RU # 1000018
PID # 9999994
ROUND 02
COMPUTER ID #

1

PROVIDER NAME / ADDRESS / CITY, STATE	DATE OF CARE	TYPE OF SERVICE OR ITEM PROVIDED	CHARGE	SOURCE OF PAYMENT	AMOUNT OF PAYMENT	
*** III. MEDICAL PROVIDER EXPENSES						
DR. R.S. FRENCH RALEIGH, NC	02/18/77	X-RAYS/AND/OR TESTS	32.00	FAMILY BC/BS of NC	NOT KNOWN 80%	01301 020199 01301 019901
		PROVIDED SERVICE	15.00	FAMILY INSURANCE?	15.00 .00	01301 009902 01301 019902

RU # 1000018
FOR JANE CURTAIN

MEDICAL CARE AND EXPENSES SUMMARY
FROM 01/01/77 to 11/28/77

RU # 1000018
PID # 9999994
ROUND 04
COMPUTER ID #

2

PROVIDER NAME / ADDRESS / CITY, STATE	DATE OF CARE	TYPE OF SERVICE OR ITEM PROVIDED	CHARGE	SOURCE OF PAYMENT	AMOUNT OF PAYMENT	
III III. MEDICAL PROVIDER EXPENSES						
DR. R. S. FRENCH RALEIGH, NC	02/18/77	X-RAYS AND/OR TESTS	32.00	FAMILY BC/BS of NC	6.40 25.60	01301 020199 01301 019901
		PROVIDED SERVICE	15.00	FAMILY	15.00	01301 009902

Table 3
Availability of documents for coding

Document	Number
Both questionnaire and Summary available	409
Only required questionnaire available	17
Only required Summary available	20
Neither questionnaire nor Summary available	4
Total	450

keyed for 100 percent verification and resolution of errors.

Results of the comparisons

Definitional considerations. In considering comparisons of data elements from the NMCES questionnaire and the Summary, two factors should be stated. First, given the nature of the data collected about medical provider visits, there are certain data elements that should be viewed as *constant*; that is, no change in values should be expected over time unless the original report was inaccurate. Constant data elements are judged to be date of visit, provider name, and total charge for the visit. Conversely, certain data elements should be expected to change over time; that is, their validity is temporal rather than absolute. The data elements where change is expected are viewed as *dynamic*. Dynamic data elements are defined as sources and amounts of payment for charges associated with the visit.

In observing change among constant data elements, the only judgment about improvement or deterioration of quality that can be made concerns the substitution of a known value for an unknown one and vice versa. That is, change from an unknown value of a known value is viewed as "better data" or "improved quality." Likewise, a change from a known value to a "don't know" response is viewed as "poorer data" or "loss of data." For dynamic data elements, where change is expected, no judgments are made here when one known value is changed to another. Either value can be considered of "equal quality," since some change over time in relative values was expected. As with the constant data elements, "improved quality" is measured only by the substitution of a known value for an unknown one. In the following presentation of the results of comparisons, the distinction between constant and dynamic data elements will be referred to when rates of change are discussed and where improvements in data quality are cited.

Consistency versus change in responses over time. The first results to be examined focus on *lack of change,* that is, a comparison of responses that did not change from the time of the original questionnaire report to the final Summary review. The frequency and percentage of identical responses are shown in Table 4.

Table 4
Matches of questionnaire and Summary responses
(Percent matching)

Variable	Type of response		Total matching reports of variable	N[a]
	Known value	Don't know		
1. Visit date				
(a) Month and day	88.6	0.8	89.4 } 98.1	368
(b) Month only	8.7	0.0	8.7 }	
2. Provider name	87.5	0.5	88.0	368
3. Total charge	62.8	21.4	84.2	199
4. X-ray/tests charge	61.5	0.0	61.5	26
5. X-ray/tests family amount	46.7	0.0	46.7	15
6. X-ray/tests payer A[b] amount	40.0	20.0	60.0	5
7. Visit charge amount	73.7	18.3	92.0	175
8. Visit family amount	80.7	1.7	82.4	119
9. Visit payer A[b] amount	31.5	63.0	94.5	54

[a] The total number of visit records examined was 368. For individual variables in the table, N represents the frequency with which the variable was available both in the questionnaire and on the Summary for comparison of responses, either as a known value or as a "don't know."

[b] The phrases "payer A" and, later, "payer B" are used to identify sources of payment for services and visits in lieu of specific designations of third-party payers.

As might be expected, high percentages of matches of initial and final reports occurred for the three variables classified as constant. Visit date reports matched in 98 percent of all cases, while provider names matched in 88 percent of all cases. The third data element considered to be constant, total charge for the visit, matched in 84.2 percent of all cases. Determining the frequency of matches for this variable was somewhat confounded by the variety of possible responses in the questionnaire. The matches for the total charge variable included in this table are of known values (dollar amounts) and "don't know" responses. A further examination of variation in total charge responses is given later in this paper.

It should also be noted that pairs of "don't know" responses are counted as matches. In the case of the amount paid by a third party for a visit charge ("visit payer A amount" in Table 4), 63 percent of the cases remained unknown for the duration of the survey. Thus, for this variable, the Summary review did not improve the quality of the initial response for a large percentage of the sample visits.

The remaining variables in Table 4 (numbers 4–9) are all classified as dynamic variables—that is, some change in values over time was expected. However, for visit charge and related amounts of payment, the percentage of matches was uniformly high—92 percent for the visit charge, 82.4 percent for the amount of the visit charge paid by the family, and 94.5 percent for the amount paid by a third party.

In considering these rather high percentages of identical responses, it should be noted that the percentages are based on cases where the variable appeared both in the questionnaire and on the Summary for comparison. In a later table, additional data are presented that show the reporting frequencies of the dynamic variables by source of report—questionnaire or Summary or both documents.

Table 5 shows the frequency and percentage of nonmatches of questionnaire and Summary responses on common variables. The highest percentages of nonmatches occurred for the dynamic variables associated with charges and amounts paid for x-rays and tests. As indicated earlier, change was expected for these dynamic variables.

Provider name responses. The variable "provider name" was considered critical during the Household Survey because of the need to secure permission to contact specific medical providers during the Medical Provider Survey. During interview Rounds 4 and 5, a great deal of emphasis was placed on securing complete, consistently spelled provider names. In 88 percent of the cases, the provider name reported in the questionnaire matched the name on the final Summary. A distribution of these matches by reporting category is shown in Table 6.

For the cases where provider name reports did not match, the variation in responses is shown in Table 7. A measure of improvement in data quality is reflected in the eight cases where the questionnaire response was "don't know" but a name appeared on the Summary. This gain was offset by an apparent loss of data in seven cases where a name had originally been reported in the questionnaire, but "don't know" appeared on the Summary. Thus, a net gain of only 3 percent was achieved in reporting known provider names. The last two variations in Table 6 are of interest because they represent fairly drastic changes in reporting from the questionnaire to the Summary. In almost 30 percent of the cases where provider names did not match, the person or place name on the Summary was completely different from the initial questionnaire report. The reason for a change of this type is not clear, nor is it within the scope of this paper. However, one must remember that the purpose of the Summary review was to permit respondents to appraise the data initially reported and to change, add to, or delete their original responses.

An effort was made to assess the quality of provider names as reported both in the questionnaire and on the Summary. The criterion for determining whether or not a provider name was "acceptable" can be stated as follows: acceptable name = Dr. Banks, Rex Hospital Emergency Room, etc. Not acceptable are "doctor, emergency room, clinic, x-ray techni-

Table 5
Nonmatches of questionnaire and Summary responses (Percent)

Variable	*Total nonmatching reports of variable*	*N*[a]
1. Visit date		368
(a) Month and day	1.6	
(b) Month only	0.2	
2. Provider name	12.0	368
3. Total charge	15.8	196
4. X-ray/tests charge	38.5	26
5. X-ray/tests family amount	53.3	15
6. X-ray/tests payer A amount	40.0	5
7. Visit charge amount	8.0	175
8. Visit family amount	17.6	119
9. Visit payer A amount	5.7	54

[a]The total number of visit records examined was 368. For individual variables in the table, N represents the frequency with which the variable was available both in the questionnaire and on the Summary for comparison of responses, either as a known value or as a "don't know."

cian" or other nonspecific titles or types of places.

Of all provider names, 92.1 percent were deemed to be acceptable on both documents. Of the remaining provider names, slightly more were acceptable in the questionnaire but not on the Summary (3.5 percent) than were acceptable on the Summary but not in the questionnaire (3.3 percent). The 1.1 percent of names that were not acceptable on either document included "don't knows" and nonspecific names such as "nurse" and "technician."

Table 6
Frequency of matches of provider name by reporting category
(N = 368 visits)

Provider name reporting category	Frequency of matches	
	Number	Percent
Person name	292	79.3
Place name	28	7.6
Title	2	0.5
Don't know	2	0.5
Total	324	88.0

Table 7
Provider name reporting variations
(N = 44)

Response variation			
Questionnaire	Summary	Number	Percent
Person name	Place name	3	6.8
Person name	Title	1	2.3
Place name	Person name	6	13.6
Place name	Title	4	9.1
Title	Person name	1	2.3
Title	Place name	1	2.3
Don't know	Name	8	18.2
Name	Don't know	7	15.9
Person name	Different person name	5	11.4
Place name	Different place name	8	18.2
Total		44	100.0

Comparison of total charge responses. As indicated earlier, comparisons of total charges reported in the questionnaire and on the Summary were confounded somewhat by the variety of possible responses allowed in the questionnaire (cf. Qs. 9–10, Medical Provider Visit Section in Exhibit 1). Comparison of reported known values was complicated by the fact that the actual total charge amount, when reported in the questionnaire, was not directly transferred to the Summary as a separate line entry. On the Summary, the total charge was either the sum of its component costs (emergency room charge, x-ray/tests charge, and visit charge) or equal to the visit charge where no separate service costs were reported. In the questionnaire excerpt shown in Exhibit 1, the total charge ($47.00) is shown in Question 9, a separate service charge for x-rays or tests ($32.00) is shown in Question 12, and the visit charge ($15.00) is shown in Question 13. In the corresponding Summary entry in Exhibit 2, the service charge of $32.00 and the visit charge of $15.00 are itemized, but the $47.00 total charge does not appear on the Summary.

Insofar as direct comparisons are possible, Table 8 shows a cross-tabulation of variation in total charge responses. As shown, six possible types of responses were possible in the questionnaire, while only four responses could appear on the Summary. In the data processing for production of the Summaries, three of the questionnaire responses were translated to "don't know." These responses were (1) don't know, with no bill expected, (2) other source will pay where the total charge was unknown, and (3) don't know, but a provider bill was expected.

Several comments are in order regarding this table. If one examines marginal percentages, several improvements in reporting are reflected. While specific dollar values represented 43.8 percent of all questionnaire total charge re-

Table 8
Variation of total charge responses
(Percent; N = 368 visits)

Summary response	Questionnaire response						
	Dollar value	Free from provider	Flat fee	Don't know (no bill expected)	Other source will pay	Don't know (bill expected)	Total
Dollar value	42.1	0.5	0.0	0.3	1.9	8.2	53.0
Free from provider	0.0	17.9	0.3	0.0	0.5	0.3	19.0
Flat fee	0.5	0.0	3.8	0.0	0.0	0.3	4.6
Don't know	1.1	2.7	0.5	0.8	15.5	2.7	23.4
Total	43.8	21.2	4.6	1.1	17.9	11.4	100.0

sponses, Summary responses of specific dollar values increased to 53 percent, a gain of over 9 percent. When percentages were added for the three questionnaire responses processed as "don't know" (see above), the true percentage of questionnaire "don't know" responses was 30.4 percent of all responses, while the Summary "don't know" responses were reduced to 23.4 percent of all responses. The flat fee percentages were identical on both documents. A slight reduction (2.2 percent) is also shown in the free-from-provider responses, which project staff believed were reported with higher frequencies than expected. In early rounds of the survey, respondents were apparently reporting visits and services as "free" when there was no direct charge to them, although a third party may, in fact, have been billed for the visit or service.

It is possible to identify very precisely how improvements in reporting were achieved by considering the same cross-tabulation in a different way. In Table 9, the percentage in each cell shows the relationship of the cell value to the column total, or the percentage distribution of questionnaire responses to Summary responses. The relationship of several total charge responses to maintained or improved data quality are of interest:

1. Of questionnaire dollar value responses, 96.3 percent remained as dollar values on the final Summary.
2. Of questionnaire free-from-provider responses, 84.6 percent remained as free-from-provider entries on the Summary; 2.6 percent changed to dollar values, while 12.8 percent changed to "don't know" responses, probably as a result of interviewer probes to determine if the services really were "free."
3. Of questionnaire flat fee responses, 82.3 percent were unchanged on the Summary, 5.9 percent were changed to free from provider, and 11.8 percent were changed to "don't know" responses.
4. Of questionnaire responses of "don't know" where no provider bill was expected, only one (25 percent) changed to a dollar value, while three (75 percent) remained unknown.
5. Likewise, where the questionnaire response was that another source would pay the bill, only 10.6 percent were changed to dollar values, while 86.4 percent remained unknown.
6. When the questionnaire response was "don't know" but a provider bill was expected, a significant 71.4 percent of the responses changed to dollar values, while 23.8 percent remained unknown.

If the questionnaire "don't know" responses are combined, the direction of change becomes clearer. A total of 33.9 percent of the questionnaire "don't know" responses were changed to dollar values on the Summary; 2.7 percent of the questionnaire "don't know" responses were changed to free-from-provider responses on the Summary; .9 percent of questionnaire "don't knows" were changed to flat fee response on the Summary; and 62.5 percent remained as unknown on the Summary.

One can use the same cross-tabulation to determine the components of types of Summary responses. Table 10 shows cell values as percentages of row totals so that the percentage distribution of the ultimate Summary responses from questionnaire responses is known. Aside from the high percentages associated with identical pairs of responses (dollar value, free from provider, and flat fee), it is interesting to note that 15.4 percent of the Summary dollar values came from questionnaire "don't know" responses where a provider bill was expected. Of all Summary "don't know" responses, 66.3 percent resulted from questionnaire responses that another source would pay the bill. Finally, it should be noted that combined questionnaire "don't know" responses accounted for 81.4 per-

Table 9
Variation of total charge responses—
percent distribution of questionnaire responses to Summary responses
(N = 368 visits)

Summary response	Questionnaire response: Dollar value	Free from provider	Flat fee	Don't know (no bill expected)	Other source will pay	Don't know (bill expected)
Dollar value	96.3	2.6	0.0	25.0	10.6	71.4
Free from provider	0.0	84.6	5.9	0.0	3.0	2.4
Flat fee	1.2	0.0	82.3	0.0	0.0	2.4
Don't know	2.5	12.8	11.8	75.0	86.4	23.8
Total	100.0	100.0	100.0	100.0	100.0	100.0
N	(161)	(78)	(17)	(4)	(66)	(42)

cent of the Summary "don't know" responses. Thus, of all questionnaire "don't know" responses, 19.6 percent were changed to "better" responses on the Summary.

Differences in charges and amounts of payment. Table 11 shows the ranges of difference between questionnaire and Summary reports of charges and amounts of payments. The difference was calculated as questionnaire value minus Summary value. Hence, the ranges of difference preceded by minus signs signify Summary values greater than questionnaire values. Considering the combined proportion of differences for all variables, questionnaire values were greater in 57.3 percent of the cases where differences are shown, while Summary values were greater in 42.7 percent of the cases. (The two reports where the Summary report exceeded the questionnaire report by $100 or more [x-ray/tests charge and x-ray/tests family amount] can be attributed to an interviewer/respondent failure to correct a data processing error that inflated the values by a factor of 10.) The implications of these differences are not clear, and no judgments are made here about the quality of responses from either source. Analyses planned by NCHS/NCHSR in conjunction with NMCES Medical Provider Survey data should clarify this issue.

Improvements in reporting from the questionnaire to the Summary. As stated earlier, the only measure of improvement deemed appro-

Table 10
Variation of total charge responses—
percent distribution of Summary responses from questionnaire responses
(N = 368 visits)

Summary response	Questionnaire responses: Dollar value	Free from provider	Flat fee	Don't know (no bill expected)	Other source will pay	Don't know (bill expected)	Total	N
Dollar value	79.5	1.0	0.0	0.5	3.6	15.4	100.0	195
Free from provider	0.0	94.3	1.4	0.0	2.8	1.4	100.0	70
Flat fee	11.8	0.0	82.3	0.0	0.0	5.9	100.0	17
Don't know	4.6	11.6	2.3	3.5	66.3	11.6	100.0	86

Table 11
Differences between questionnaire and Summary reports of charges and amounts of payment
(N = 368 visits)[a]

Ranges of difference (questionnaire value minus Summary value)	Variable: Total charge	X-ray/tests charge	X-ray/tests family amount	X-ray/tests payer A amount	Visit charge	Visit family amount	Visit payer A amount
−$100 or less	0	1	1	0	0	0	0
−$25 to −$99	0	0	0	0	0	0	0
−$20 to −$24	1	0	0	0	0	1	0
−$15 to −$19	1	0	0	0	1	1	0
−$10 to −$14	0	0	0	0	0	1	0
−$5 to −$9	7	3	2	1	2	4	0
−$1 to −$4	4	1	1	0	2	3	0
Zero	123	16	7	2	129	96	17
$1 to $4	9	5	4	1	6	4	1
$5 to $9	3	0	0	0	2	4	1
$10 to $14	3	0	0	0	1	2	1
$15 to $19	1	0	0	0	0	0	0
$20 to $24	1	0	0	0	0	1	0
$25 to $29	1	0	0	0	0	0	0
$30 to $99	0	0	0	0	0	0	0
$100 or more	0	0	0	0	0	0	0
N	(154)	(26)	(15)	(4)	(143)	(117)	(20)

[a] The total number of visit records examined was 368. For individual variables in the table, N represents the frequency with which a known value greater than zero appeared both in the questionnaire and on the Summary.

priate for use in this analysis of questionnaire and Summary responses is viewed as the replacement of an unknown response with a known value. That is, a known value is considered a better response than a "don't know," regardless of the apparent reliability of the known value. Table 12 provides a comparison of known and unknown values reported for visit date, charges, and amounts of payment by family and a third party. The frequencies and percentages in Column 1 reflect improvements in reporting, since the questionnaire report was unknown but the Summary report was a known value (greater than zero). The highest relative percentage of improvement was for the total charge variable, where almost 20 percent of the unknown questionnaire reports were changed to known values on the Summary. Column 2 represents an anomaly that is referred to as "loss of data"; that is, a known value was recorded in the initial questionnaire report, while the final Summary report of the same variable was unknown. The reason for this loss of data was not tabulated, although a few cases were observed where respondents had changed known values to "don't know" responses on the final Summary. Column 3 represents the "don't know status quo"; that is, no improvement in reporting was achieved in those cases where both the questionnaire and Summary reports were unknown. In 63 percent of the cases where a third party was to pay some amount of a visit charge, the amount of payment remained unknown. Some comments on this problem are made in the conclusions to this paper.

Table 13 includes both comparative and individual document data. Certain variables were reported both in the questionnaire and on the Summary, only in the questionnaire, or only on the Summary. The "net gain" column shows the percent increase of reports achieved on the Summary, although a caveat is appropriate here. The net gain percentage was computed by subtracting the number of reports that appeared only in the questionnare from the number that appeared only on the Summary; the product was then divided by N, the combined total reports of the variable on all combinations of documents. Calculating the net gain in this manner is clearly a conservative calculation, since the process assumes that an apparent "loss" of questionnaire data has equal weight with added Summary data. This assumption may be inappropriate for several of the dynamic variables, where validity is presumed to be temporal rather than absolute. In particular, the "visit payers A and B" reported only in the questionnaire may represent third-party payments that were expected but simply did not materialize. Consequently, the true net gain of reports on the Summary may have been considerably higher if one calculates the net gain as

Table 12
Comparisons of known and unknown values in the questionnaire and on the Summary
(Percent)

Variable	(1) Questionnaire report = don't know, but Summary report = known value	(2) Questionnaire report = known value, but Summary report = don't know	(3) Both questionnaire and Summary reports = don't know	N[a]
Visit date				368
(a) Month	1.1	0.5	0.5	
(b) Day	1.6	0.0	7.1	
Total charge	19.9	2.0	21.4	196
X-ray/tests charge	11.5	0.0	0.0	26
X-ray/tests family amount	0.0	6.7	0.0	15
X-ray/tests payer A amount	20.0	0.0	20.0	5
Visit charge amount	4.6	2.3	18.3	175
Visit family amount	3.4	1.7	1.7	119
Visit payer A amount	14.8	3.7	63.0	54

[a]The total number of visit records examined was 368. For individual variables in the table, N represents the frequency with which "don't know" responses or known values greater than zero appeared both in the questionnaire and on the Summary for comparison.

the percent increase of Summary only reports over reports appearing on both the questionnaire and the Summary.

Consistency in reports of constant variables. A final comparison was made to measure *consistency* rather than change in reports of visit data. This comparison examined all visit records for identical questionnaire and Summary values on the following variables:

1. Visit date—month and day.
2. Provider name—same person, place, or title.
3. Provider name quality—acceptable on both documents.
4. Total charge—known value greater than zero or both documents cited charge as free from provider or covered by a flat fee.

Of the 368 visit records, 165 (44.8 percent) met the match criteria for consistency between the initial questionnaire report and the final Summary report. It should be noted that only constant variables were matched, since it was believed that inclusion of the dynamic variables in the match criteria would have inappropriately reduced the percentage of matches because change was *expected* in the dynamic variables.

Conclusions

The purpose of the comparative analysis of questionnaire and Summary responses was to determine the nature and direction of changes that were made during one or more retrospective reviews of previously reported data. The focus was to identify frequencies of change among key variables and to observe the outcome of change as it affected the quality of the data ultimately derived from the Summary. The sole criterion for judging improvements in the quality of the data was the change from an unknown value to a known one. The principal results of the comparisons are summarized below:

1. Almost 45 percent of all visit records reflected identical reports on constant variables—visit date, provider name, quality of provider name report, and total charge for the visit.
2. The use of the Summary had almost no impact on improving the reporting of provider names. Similarly, the relative quality of provider names reported was about the same on both documents.
3. Improvements in the total charge variable can be attributed to the Summary. Almost 34 percent of all questionnaire "don't know" responses were changed to dollar values on the Summary.
4. Known value charges and amounts of payment were identical on both documents in over 81 percent of total reports of these variables. Questionnaire values were greater than Summary values in almost 11 percent of total reports, while Summary values were higher in almost 8 percent of total reports.
5. With the exception of provider names, some improvement in reporting was observed in all variables examined where the questionnaire report was "don't know" but the Summary yielded a known value.
6. Small numbers of known questionnaire responses were changed to "don't know" responses on the Summary; the reasons for this

Table 13
Total reporting frequencies of charges and sources of payment by source of report
(Percent)

Variable	Source of report			Net gain	N*
	Report in questionnaire and on Summary	Report in questionnaire but not on Summary	Report on Summary but not in questionnaire	Summary only minus questionnaire only reports	
X-ray/tests charge	76.5	8.8	14.7	5.9	34
Family as X-ray payer	71.4	9.5	19.0	9.5	21
X-ray payer A	55.5	0.0	44.4	44.4	9
Visit charge	68.9	3.5	27.5	24.0	254
Family as visit payer	88.1	4.4	7.4	3.0	135
Visit payer A	63.5	17.6	18.8	1.1	85
Visit payer B	30.8	30.8	38.5	7.7	13

*The total number of visit records examined was 368. For individual variables in the table, N represents the frequency with which a known value greater than zero or a "don't know" response was reported either in the questionnaire, on the Summary, or on both documents.

apparent loss of data were not explored in this analysis.

7. Certain variables—total charge, visit charge, and payer amounts—were more likely than other variables to be reported initially as unknown and to remain unknown in spite of repeated attempts to secure known values.
8. For all variables examined, there were gains in reporting where responses were found on the final Summary but not in the initial questionnaire visit report.

The results of comparisons of questionnaire and Summary responses presented earlier in this paper do not offer irrefutable evidence for or against use of a document such as the Summary in subsequent surveys of medical utilization and expenditures. One can look back at Table 4 and argue that a fairly high percentage of matches of questionnaire and Summary reports occurred on most visit data elements. In looking only at matches of known values, matches between questionnaire and Summary reports were observed for six of the nine key variables in over 60 percent of all reports of the variable. On the other hand, there is clear evidence that use of the Summary led to improved data, particularly where charges and amounts of payment were initially reported as unknown (cf. Table 12) and that data collection methodology did not require extensive probing of such initial reports.

One can draw a general conclusion that the overall effect of the Summary was positive rather than negative. Clearly, some data initially reported as unknown or missing from the questionnaire report of the visit would not have been secured had not the Summary been used. However, as shown in Table 12, both questionnaire and Summary reports of two key variables—total charge and visit charge amount—remained unknown in 21 and 18 percent, respectively, of total reports of the variable. A comment based on NMCES experience with measurement of medical expenditures may help explain this problem.

It appears from NMCES experience that there is a certain subset of respondents from whom reliable data on costs of medical care will never be available, regardless of how many times an interviewer asks, "Do you now know how much the charge was for this visit?" These respondents are primarily those enrolled in public programs where benefits are assigned to providers, such as Medicaid and, in some cases, Medicare. These respondents rarely have access to provider bills, third-party-payer statements, or other documents that state concrete facts about the costs of services. In addition, persons who are enrolled in prepaid health plans and health maintenance organizations can usually report only the nominal charge that they pay on a per visit basis.

In retrospect, the NMCES questionnaire and Summary never really acknowledged the reporting limitations of these groups of respondents. Both the questionnaire and the Summary were based on the measurement concept that the gross cost of a visit or service was some amount greater than the actual cost to the respondent, who was really a copayer with some third party (Medicare and Medicaid) or who paid certain service charges above and beyond any fixed prepayments to PHPs and HMOs. Consequently, charges and proportionate amounts paid by all payers in these cases were treated as unknown, since the gross cost could not be reported by the respondents. In some instances, it was noted that cases involving unknown charges and amounts of payment occurred where the individuals were Medicaid or Medicare participants.[4] Thus, despite up to four reviews of previously known data, certain respondents simply could not improve the quality of their initial report because the data they needed were not available to them.

The purpose of the preceding discussion is to put further comments about the utility of the Summary in a proper context. Quite simply, it does not seem unrealistic to expect improvement and change over time where respondents may gain access to information that will enable them to improve their initial report. It should be pointed out that the comparisons in this paper are based on what might be considered "worst case data"—the Round 1 questionnaire data were probably of poorer quality than those collected in subsequent rounds because the respondents were unaware of their expected reporting and record-keeping roles. Thus, similar comparisons of questionnaire data from later rounds with corresponding Summary entries might well reflect better initial reports and higher percentages of improvements (i.e., changes from "don't knows" to known values) as respondents performed their roles better.

It is clear from the comparisons in this paper that certain variables are reported more completely and with greater face validity than others. Specifically, the date of the visit, the provider's name, and the amount that the family paid for the visit charge were identical in 87, 88, and 81 percent, respectively, of cases where known values could be compared. Third-party-payer amounts for all charges reflected the least consistency in reporting as well as rather high percentages of final "don't know" responses (cf. Table 12). Thus, one can con-

clude that the use of the Summary did not resolve the "don't know status quo" problem entirely.

The ultimate methodological evaluation of the utility of the Summary will require extensive statistical analyses of changes that were made across all survey rounds, in combination with a careful cost-benefit analysis. Until additional cost-benefit analyses are undertaken, no definitive conclusions can be drawn about the net advantages (or disadvantages) of using a computer-generated Summary to maintain and improve the quality of interview data.

The analytic implications of the utility of the Summary will become known as a series of matches of data from three sources are made. The initial questionnaire report, the final Summary report, and the independent provider report from the Medical Provider Survey will be matched to create a "best estimate" file. The results of these matches will be of value not only analytically but also methodologically as perhaps the most revealing evidence of the outcome of using computer-generated Summaries.

Appendix: Definitions of summary update types

Update: A single line of code used to make an addition, change, or deletion to a single summary line of information.

Correction: An addition, change, or deletion of one item in an update.

Type: There are nine update categories that permit corrections for the Medical Provider Visit and Prescribed Medicine pages, the Flat Fee page, and the Health Insurance page. Each type of update can have one or more corrections as follows:

Person name: Only "name of respondent" spelling changes can be made.

Provider name: The "provider name" for doctors, hospitals, dentists, and prescribed medicines and the "date of visit" can be changed, added, or deleted with this update type.

Provider address: The provider address variables can be corrected with this type.

Service types: The "service" phrase (i.e., provided service, x-rays and tests, emergency room) can be corrected using this type. In addition, an indicator to code the kind of payment for service is included. Different kinds of payments are (1) payment for a service, (2) payment for the visit, or flat fee payments. The flat fee letter may also be corrected using this type of update.

Payers: There are four corrections that can be made using this update type: (1) the "charge" is the total charge for payment, (2) the "source" is the source of payment, (3) the "amount" is the amount paid by each source, and (4) the $/% is an indicator for the amount, being dollars or percent.

Flat fee: This update pertains to charges, sources, and amounts on the Flat Fee page in the same manner as the "payers" update.

Insurance policies: There are four corrections possible on the Health Insurance page using this update:

1. "Plan letters" are the assigned double alpha characters.
2. "Category" is dental plan only, public, private, or special plans.
3. "Plan name" is the policy name shown.
4. "Not plan name" is an indicator to be used when the plan name is not available but company names or some other title is used.

Persons covered: The "PID" is the person identification number of the respondents covered by a given insurance policy. The "Most know" is an indicator of whether or not that person is the most knowledgeable about the insurance policy.

Refund: This update allows coding of an indicator to reflect reporting of an insurance refund that covers multiple services where specific amounts cannot be allocated to individual visit/service payer amounts.

Footnotes

[1] A "flat fee" was defined as a single charge for multiple visits and/or services. For example, a single charge of $800 might cover a patient's preoperative office visits, inpatient surgery and physician services, and postoperative office visits.

[2] Definitions of all Summary update categories are included in the Appendix. The four categories shown in Table 2 are defined here to aid the reader.

Provider name: The proper or common name reported by a respondent to identify a dentist, physician, facility, or other type of medical provider. This data field also included the names of prescribed medicines and dates of provider visits.

Provider address: City and state where provider is located.

Service types: A descriptive legend identifying the type(s) of service reported as being received during a provider visit (e.g., x-rays and/or tests, eye examination, shots/immunization). This data field also contained indicators for corresponding service, visit, or flat fee payments.

Payers: This update category included each charge, source of payment for a charge, the amount of the charge paid by each source, and an indicator for the amount paid to differentiate between dollar amounts and percent values.

[3] Continuation pages were used to record interview data when all available recording space in the questionnaire had been used.

[4] This observation is based on visual examination of the questionnaires and Summaries by the author while checking the comparison coding. Unfortunately, the com-

parison codes did not identify specific sources of payment, so that tabulation of the frequency of Medicare and Medicaid as payers was not possible.

Discussion: The use of Summaries of previously reported interview data in the National Medical Care Expenditure Survey

Judith Kasper, National Center for Health Services Research

I would like to talk about two separate aspects of Holt's paper. One of them is her Table 1, which shows changes that were made on the Summary across all five rounds of interviewing with all the families in our survey; then I want to address some comments to her study of 450 visits.

The striking thing, of course, about Holt's Table 1 is the very large number of changes that we observed our respondents made on the Summary. To understand this table, it is useful to look at Exhibit 2. The first category in Table 1, person name, refers to changes to what you see in Exhibit 2 as Jane Curtain. The second category, provider name, includes changes both to the name of the provider and to the date of care. Provider address is self-evident. Service types refers to changes on the Summary to "type of service or item provided." The column in Table 1 headed "payers," which probably is the most important category to understand, refers to changes to any of the following: total charge, source of payment, or amount of payment. In looking at Table 1, you see that almost half of all the changes that were made on the Summary affected the items in this category. The other categories in Table 1—insurance policies, persons covered by insurance policies, and refunds—had essentially separate pages on the Summary and were handled in a slightly different way from the regular visit information.

Holt's study focuses on changes of data from known to unknown and from unknown to known. The kinds of changes that we see in Table 1 really are much less dramatic than the changes that Holt considered. She mentioned an example involving doctors' names. A respondent in Round 1 may have reported to us that a visit was made to Dr. Jones; then in Round 4, when we were trying to get more complete information about names, the respondent may have said it was Dr. Burt Jones; in Round 5, it may have become Dr. Burt L. Jones. Each of these variations on the provider's name would be recorded in Table 1 as a change. But in Holt's study, where she is looking at changes of data from a known status to an unknown status and vice versa, these types of changes are not taken into account. The most severe test of the Summary's effectiveness is whether unknown data became known. It is also possible, however, that many of the more minor changes, which are probably reflected in Table 1, improved originally reported data in less dramatic ways, e.g., the example of Dr. Burt L. Jones. Ultimately, of course, the important issue is whether all of these changes that were made on the Summary resulted in better data. There are a number of methodological studies being planned to address this issue; these are going to be discussed later in this session in a little more detail. One of the major analyses to answer this particular question concerns comparing the data as first reported by a respondent with the final Summary version of this information and then making a further comparison with the medical provider data that we acquired from hospitals and physicians.

Turning to Holt's study of a sample of 450 visits, I would first like to comment on the distinction that she makes between constant and dynamic data elements. Although there may have been some expectations, early in this process, that some data items were more likely than others to be changed by our respondents, in fact they had the opportunity to change anything once we put it on that Summary. Sometimes very drastic kinds of changes occurred, such as when a respondent in one round said that a visit was made by her husband and later changed the person by whom the visit was made to a daughter or son. Table 4, where Holt tries to look at matches for constant versus dynamic data elements, demonstrates that perhaps this distinction is not going to work very well for us. Excluding Categories 4, 5, and 6, which cover x-rays and deal with some very small numbers, the rate of matches for the first three categories and then for 7, 8, and 9 really

are not very different (the three constant variables had match rates of 98, 88, and 84 percent, respectively, while the dynamic variables had match rates of 92, 82, and 94 percent).

I would also like to discuss briefly the idea of evaluating the Summary by looking at changes in data items from unknown to known and known to unknown. There are some difficulties with using this particular way of evaluating the effectiveness of the Summary. I have already mentioned the kinds of changes that we observed with regard to provider names, which really were not changes from an unknown to a known status but were, instead, refinements on a first report. In her study of 450 visits, Holt observed a very high degree of match on provider names—88 percent—but if you look at Table 1, you see that 27 percent of all of the changes on the Summary were made in this category. Clearly, a lot of the kinds of changes that were being made probably were clarifications or refinements.

It might be useful to think about changes on the Summary as being of three types. The first type picks up data not originally reported, which would be Holt's change from unknown data to known data. The second type are other dramatic alterations of originally reported data, such as deleting a visit or assigning a completely new provider to a visit, one aspect of which Holt tried to examine by looking at changes from a known to an unknown status. The third and largest category is clarification of previously reported data—changes of a charge from $5 to $10 or changes of an insurance company from Blue Cross to Blue Cross of New York. (Warnecke suggested that this third category might be divided into trivial and nontrivial changes, the provider name changes being an example of trivial changes and changes to the charge data being nontrivial).

Finally, Holt's assumption that data changing from an unknown status to a known status is an improvement in data quality and that changes from a known status to an unknown status is not raises a very interesting issue that survey researchers often do not have to think about: whether any answer is always better than no answer or, put another way, whether gaining data is always better than losing data. There are examples in which losing data probably will leave us with a better-quality data set. For instance, the respondent reports a visit for somebody else in the family and then, in reviewing the Summary in the next interview, says, "I was wrong; I talked to my husband and he didn't go see Dr. Jones about his leg after all." On the other hand, it is also possible that given all the opportunities that our respondents had to make changes on the Summary, some of the changes that they made were not for the better, and loss of data does represent lower quality. The answers to these questions really will have to wait until we can look at the data that are coming from physicians and hospitals and use those in some way as a measure of "truth" against which we can compare first reports of data and final Summary reports. In that light, while one can calculate a "net gain" in data between the questionnaire and Summary, as Holt did in Table 13, at the present time we are unable to evaluate the implication of the result for either overall data quality or the effectiveness of the Summary.

Survey of interviewer attitudes toward selected methodological issues in the National Medical Care Expenditure Survey*

Esther Fleishman, National Opinion Research Center, University of Chicago

Marc Berk, National Center for Health Services Research

Introduction

The National Medical Care Expenditure Survey (NMCES) was designed to collect as accurate and complete data as possible on the health care utilization and expenditures of Americans for a one-year period, the year 1977. To this end a variety of data collection techniques were employed that would help ensure these results. Some of the techniques are routinely used in survey research, their effectiveness having been previously established, while others were innovative, their usefulness not as yet proven and the subject of debate.

The primary measure of the usefulness of various data collection strategies will be the accuracy of the data produced by those strategies. There are, however, numerous intangibles in the interviewing situation that such data alone cannot capture. The Interviewer Survey was designed to collect data about how certain information was obtained, how respondents reacted to certain issues, and how interviewers reacted when they themselves became the instruments of certain field procedures.

The evaluations of interviewers have seldom been the subject of systematic research. Most surveys are simply too small and employ too few interviewers to lend themselves to this type of inquiry. NMCES, however, was one of the largest survey research operations ever conducted. In all, some 400 interviewers and 25 regional field supervisors employed by RTI and NORC worked on the project, of whom 280 participated in at least five of six rounds of interviewing. Of these 280 interviewers who were deemed eligible to respond to the survey, 234 completed the questionnaire, for a response rate of 84 percent.

The primary purpose of the Interviewer Survey was to collect data pertaining to the effectiveness of selected field procedures, with particular emphasis on the way such procedures affect the quality of data obtained from poor and elderly respondents. A further goal of the survey was to examine interviewer attitudes toward selected aspects of their work experience.

In this paper we address three topics that have policy implications for health survey planners: (1) response rates, (2) quality of data in face-to-face interviewing versus telephone interviewing, and (3) methodological issues concerning the poor and the elderly.

Inevitably the question will be asked whether interviewers can be objective when their livelihood depends on successfully carrying out the task that they are being asked to evaluate. In fact, several of the procedures that were evaluated have a potential effect on interviewer income. We found, however, that time and again interviewers as a group did not respond in terms that were self-serving. As will be seen in the discussion of several items, they discriminated carefully among issues without regard to how they might personally be affected by changes in survey design.

One may also question whether interviewers have the technical expertise necessary to evaluate the effectiveness of field procedures. The research design of NMCES provides us with a unique opportunity to explore this issue. Data from the Medical Provider Survey will enable us to evaluate the accuracy of some of the data obtained in the Household Survey. To some extent, we will also be able to determine which field procedures resulted in more accurate and complete information. At that time, it may be possible to determine the accuracy of certain interviewer perceptions.

The NMCES interviewing staff

The National Medical Care Expenditure Survey required interviewers to use very complex instruments, undergo rigorous training, and ad-

*Work supported by NCHS Contract No. 233-78-2102, "National Medical Care Expenditure Survey—Methods and Analysis."

here to a demanding field schedule. It was therefore particularly important to select and recruit interviewers with the greatest possible skills. An analysis of the demographic characteristics of NMCES interviewers documents that the survey had a highly qualified staff. As shown in Table 1, 84 percent of the NMCES interviewing staff was 30 years old or over. In all, 76 percent of the NMCES interviewers had some college education, including 17 percent who had more than 16 years of schooling; 58 percent of the interviewers had at least five years of interviewing experience in survey research, census, or polling work. Only 13 percent had never interviewed before NMCES. These figures are considerably higher than those reported in a survey of census interviewers who conducted the 1975 Current Population Survey. In that study, 45 percent of the interviewers reported some college education, while 28 percent had more than six years of interviewing experience (U.S. Office of Federal Statistical Policy and Standards, 1978).

Responses to an open-ended question indicate that the majority of the interviewing staff found NMCES to be a rewarding experience. When asked whether they would want to work on NMCES again, 75 percent responded affirmatively, while 17 percent said that they were not sure. Only 8 percent of the interviewers said that they would not want to work on the study again. A desire to work on NMCES again was positively correlated with interviewing experience but negatively correlated with education. This may indicate that desire to work on NMCES depends at least partly on what other job opportunities the interviewer has available.

Table 1
Demographic characteristics of NMCES interviewers

Characteristic	*Percent*
Age:	
Under 29 years	16
30–39 years	33
40–49 years	44
50 years or over	7
Total	100
N	(231)
Highest grade completed in school:	
0–12 years	24
13–16 years	59
17 or more years	17
Total	100
N	(233)
Experience:	
0–2 years	25
3–4 years	18
5–9 years	24
10 or more years	33
Total	101[a]
N	(233)

[a] Category percentage exceeds 100% because of rounding.

Response rates

Response rates on NMCES were above 90 percent in every round. In order to attain those rates, certain strategies were employed, while certain others were considered and discarded. In asking the interviewers to comment on this issue, we presented them with nine alternative suggestions that might in a future survey of this kind be effective in contributing to high response rates. All of the items were the kind that one hears informally from supervisors and interviewers during the course of field work. The rank position of each item was determined by the percentage of interviewers who ranked an item as either first or second in effectiveness.

The order of the nine items turned out to be as follows:

	Order of effectiveness
Have no interview take more than one hour	1
Offer fuller and better explanation of study at Round 1	2
Offer the same amount of money (or only slightly more) but offer a payment at each round	3
Have all personal interviews	4
Offer bonus for participation in all rounds	5
Develop better promotional material	6
Offer more money	7 & 8
Have important people endorse the study	7 & 8
Involve entire household in explaining participation	9

Since every interviewer's performance is evaluated on response rate (among other performance measures), one might be led to believe that when asked to rank nine items in the order of their effectiveness in increasing the response rates on a future medical care expenditure survey, interviewers would place at or near the top those items that would enhance their performance. After all, interviewers do not have to concern themselves with budgets, and presumably if respondents found participation sufficiently rewarding monetarily, one would have no problem getting them to participate (except those few to whom no amount of money would appeal). The results of this exercise turned out

to be revealing in that the item "Offer [respondents] more money" tied in seventh place with "Have important people endorse the study." And conversely, the item that is directly related to how much an interviewer earns, which was phrased "Have no interview take more than one hour," placed first by a considerable margin.

Aside from helping to establish interviewer objectivity, there are other lessons to be learned from the responses to the question of what would help to increase response rates on a future survey. When interviewers are given other choices, they do not place much faith in paying respondents as a means of getting them to participate, since the interviewers ranked the items concerned with payment to respondents as 3rd, 5th, and 7th on a scale of 9. It appears that interviewers do not feel that an increase in compensation would be a major factor in increasing response rates.

What the interviewers are saying would be most effective, given the choices presented, is "Have no interview take more than one hour," but that is probably too high a price to pay, not in terms of money but in terms of data, for increasing a response rate that was already satisfactory.

Curiously, the item that ranked second in importance as a means of increasing the response rate was "Offer fuller and better explanation of study at Round 1." The explanation that was offered was a letter to each household with questions and answers about the survey on the back, where quite deliberately no mention was made of the longitudinal nature of the study; only toward the end of the Round 1 interview were respondents informed that the interview in which they had just participated was the first of seven interviews that would take place at intervals of two months over a period of about a year.

At the time that decisions were made about how to approach respondents, no one could foresee the genuine and serious commitment that thousands of families would eventually make to the project, hence the rather tentative approach about telling people that their households had been selected to participate in a multiwave panel survey. In retrospect, and in light of the interviewers' perceptions, we did not need to be so hesitant about telling respondents what we were asking of them. On the contrary, we might well have increased participation if we had stated more explicitly what actually was entailed.

Face-to-face versus telephone interviewing

NMCES utilized both face-to-face and telephone interviews. The first, second, and fifth rounds of the study were designed for face-to-face interviewing, while the third, fourth, and sixth rounds were designated as telephone rounds. Even in the telephone rounds in which the main instrument was used, that is, in Rounds 3 and 4, about 20 percent of the interviews were done face-to-face. The main reasons were that respondents had no phone, were on party lines where confidentiality could not be guaranteed, or refused to give out their phone number. In some cases where the respondents were elderly or impaired, interviewers sensed that the interview would go more smoothly if it was conducted face-to-face. In still other cases, respondents simply requested a face-to-face interview.

We asked interviewers to evaluate the relative advantages of both methods of interviewing, and although we recognize that interviewer perceptions are neither the only criteria nor even the most important criteria on which an objective evaluation of the two procedures should be based, nevertheless, there are important insights to be gained from those who were directly involved in the interviewing process.

Findings. Our analysis indicates that, given the objectives of NMCES to gather accurate and complete information about health expenditures, there is a very strong consensus among interviewers that face-to-face interviewing yields better data than does interviewing over the telephone. As shown in Table 2, 96 percent of the interviewers considered face-to-face interviewing to have been better suited for achieving the survey's objectives, while 3 percent said it made no difference and only 1 percent thought telephone interviews were preferable.

Among the interviewers, 87 percent thought respondents preferred the personal interview; and when interviewers were asked which

Table 2
Telephone versus face-to-face interviewing (Percent)

Question	Method				
	Telephone	Face-to-face	No difference	Total	N
Which method of interviewing was better suited to the gathering of accurate and complete data?	1	96	3	100	233
Which method did respondents prefer?	2	87	11	100	228
Which method did interviewers prefer?	3	84	13	100	232

method they themselves preferred, 84 percent favored face-to-face interviewing, 3 percent preferred the telephone, and 13 percent said that they had no preference. These findings take on increased significance when we look at the 37 interviewers who stated that they themselves either preferred the telephone or had no preference between the two methods. Even among this group, 86 percent thought that face-to-face interviewing yielded better data than those obtained in telephone interviews, and 62 percent of this group thought that respondents preferred face-to-face interviews.

Interviewers were also asked to state which interviewing method was more suitable for selected categories of respondents. As shown in Table 3, 93 percent thought that face-to-face interviewing was more suitable for the poorly educated, and the remainder said it made no difference. Not a single interviewer of the 234 who responded thought that telephone interviewing was the better method for poorly educated respondents. Similarly, 97 percent stated that face-to-face interviewing was more suitable for the elderly, and the remaining 3 percent said it made no difference. Face-to-face interviewing was also the clearly favored method for interviewing people with large families, low-income people, and the physically handicapped. It is perhaps significant that interviewers whose assignments were in the inner city considered face-to-face interviewing as the more suitable method on NMCES, although this must be stated with caution since only 18 out of the 234 interviewers, or 8 percent, claimed that the majority of their cases were located in the inner city.

Although interviewers generally expressed a strong preference for face-to-face interviewing, it is clear that they were able to distinguish between the various effects on different respondent groups. Thus, 34 percent preferred the telephone when interviewing college students, compared with 27 percent who favored face-to-face interviewing for this group; the rest felt it made no difference.

Table 3
More suitable method for interviewing selected types of respondents (Percent)

Respondent type	Method		No difference	None in interviewer's assignment	N
	Telephone	Face-to-face			
Respondents with large families	2	88	7	3	228
Poorly educated people	0	93	4	3	230
Well-educated people	6	45	46	2	227
College students	33	27	33	7	230
Elderly	0	97	3	0	230
Low income	1	78	17	4	229
Physically handicapped	0	79	6	15	230

The interviewers' perceptions that face-to-face interviewing on NMCES yielded better data than those obtained on the telephone is attributed by them to several factors. Interviewers stated that the social aspects of face-to-face interviewing had a strong positive effect on the data. They also felt that better data were obtained when they had the opportunity to see medication bottles, bills, and the calendar. Face-to-face interviewing was also considered the preferable method of interviewing when detailed probing was necessary and when the Summary was reviewed. The fact that respondents could see what interviewers were recording was also considered an advantage. Interviewers did not believe that the household distractions associated with face-to-face interviewing had an adverse impact on data quality.

In examining these results, we exercised considerable caution, recognizing that interviewers who at Round 1 were recruited to do face-to-face interviewing, a large majority of whom stated a personal preference for the face-to-face method, might bias their statements in favor of the personal interview. We note, however, that every one of the 18 interviewers who stated unwillingness to work on the study again also stated that face-to-face interviewing yielded better data.

Conclusions. The policy implications of these findings appear to be problematic. On the one hand, there is an overwhelming consensus among the interviewers that in order to achieve the objectives of NMCES, face-to-face interviews obtain better data than those conducted by telephone, especially for certain segments of the population. On the other hand, we note that these perceptions appear to be inconsistent with the findings of previous research. Studies by Colombotos (1969), Henson, Roth, and Cannell (1977), Rogers (1976), Locander, Sudman, and Bradburn (1976), and Klecka and Tuchfarber (1978) all indicate that the data obtained in telephone interviews are comparable to those obtained in face-to-face interviews. Most of these studies, however, used very different samples and research designs than those employed in NMCES. In the analysis of the more closely related study by Yaffe and Shapiro (1979), in which comparisons were made between telephone and face-to-face interviewing for accuracy of utilization and reporting charges, a generally higher degree of accuracy

was achieved in face-to-face interviewing.

There are two possible explanations for our findings. There is the strong possibility that they apply only to complicated, multiwave, general population surveys that place relatively high demands on participants. Of course, the second possibility is simply that the interviewers' perceptions were wrong; our findings, therefore, are suggestive but not conclusive. A more detailed analysis of the relative accuracy of the two methods can be conducted at the conclusion of the Medical Provider Survey, where accuracy and level of reporting can be ascertained based on the two methods of interviewing, but even that analysis may not be conclusive. Our findings show that the NMCES interviewers who were recruited for face-to-face interviewing and expressed a strong preference for this method found it more suitable for achieving the overall objectives of NMCES but were able to discriminate among population groups, for some of whom they considered the telephone to have been more suitable. What we do not know is whether interviewer preference for one method over another is in itself a confounding variable. For example, assuming that an analysis of the data will eventually show that face-to-face interviewing did obtain better data on NMCES, would the results be different if during the telephone rounds we had employed interviewers who had a strong preference for telephone interviewing? Unfortunately, it will not be possible to test this hypothesis for NMCES.

There are many factors to be considered when making decisions regarding the use of face-to-face versus telephone interviewing on health surveys. The factors that have been examined by other health survey researchers include the total amount of reporting, the level of reporting of sensitive data, and comparative monetary costs. The general feeling is that there are costs and benefits to both methods, but the factor that to our knowledge has heretofore not been included in any cost-benefit analysis is the question of interviewer preference.

In the multiwave mixed-method panel survey such as NMCES, a problem arises in that interviewers clearly have preferences for the method by which they interview, and these preferences are very likely to affect morale and, by extension, possibly the quality of the data. But changing interviewers by type of round clearly would not be a good idea. Aside from incurring heavy additional training costs, one would lose the very real advantages of respondent rapport that interviewers build throughout the survey year, in addition to losing the proficiency that interviewers gain over time in using very complex instruments.

In recruiting interviewers for a future survey such as NMCES, it would be well to state at the outset that although at least the initial interview will be face-to-face, subsequent rounds will be conducted by phone, so that interviewers who have strong antipathies toward telephone interviewing can be disqualified. We would further suggest that even in telephone rounds, selected groups of the population be interviewed face-to-face, not as a matter of choice by the interviewer, but as a matter of requirement set forth in the specifications.

Methodological issues concerning the poor and the elderly

In survey research, it is generally taken for granted that the study population has a certain competence to perform the role of respondents. For the most part, it is not known what the precise degree of that competence is, although it can be reasonably well established by means of pretesting. NMCES did conduct very extensive pretests, but judgments were made regarding the population as a whole and not specifically about special subgroups. In conducting the Interviewer Survey, a post hoc attempt was made to learn how interviewers perceived both the poor and the elderly in terms of a number of methodological issues. The "poor" were defined as those households in which the *main* respondent was a poor person, i.e., a member of a household in which the joint annual income in 1977 was roughly $5,000 or less. "Elderly" households were defined as those *non*poor households in which the main respondent was 65 years of age or older during the survey year.

The issues on which it was felt that interviewers could provide insight regarding the poor and the elderly were (1) face-to-face interviewing versus telephone interviewing; (2) the optimum length of intervals between rounds; (3) performance, or the lack thereof, in regard to utilizing the Summary; (4) use of the calendar/diary; (5) comprehension of the full implications of signing permission forms that authorized NMCES to obtain data from medical providers, employers, and health insurers; and (6) comprehension on the part of respondents of their role as respondents.

Findings. In a very broad sense, all of these issues dealt with the question of how competent the interviewers thought respondents were to perform the task that they were asked to perform and whether special population groups such as the poor and the elderly differed in their competence from all other respondents. If one looks at these two population groups across

the methodological issues that were examined, a general pattern emerges: On the whole, the poor were regarded as somewhat less competent respondents than the elderly, and both groups were seen as somewhat less competent than all other respondents.

However, in relation to several specific issues, there are important peculiarities to each group that merit separate examination. On the question of face-to-face versus telephone interviewing, as has been stated earlier, the interviewers were virtually unanimous, in that 97 percent thought that face-to-face interviewing was more suitable for the elderly; 78 percent also thought it more suitable for the poor. As regards frequency of interviewing (in a six-round panel survey), 18 percent of the interviewers thought that more frequent interviewing would be appropriate for the poor, and 55 percent thought so for the elderly.

A series of 12 questions was asked to establish how the interviewers thought respondents reacted to the Summaries, when one can be reasonably sure that most respondents had never seen a computer printout before, much less one that had their name and address on it. The reactions that were measured included "anxiety over personal information being in a computer," "embarrassment at not being able to understand the Summaries," and "eagerness about learning to read the Summaries."

For the purpose of building a typology of respondent competence in terms of the population subgroups "poor" and "elderly," the item that was selected to measure competence in dealing with the Summary was whether respondents corrected and updated Summary information prior to the interview. The interviewers were asked to rate whether most, some, a few, or none of each of the population subgroups did so. Table 4 indicates that 17 percent of the interviewers thought most or some of the poor corrected and updated the Summary, 24 percent thought most or some of the elderly did so, and 51 percent of the interviewers thought that most or some of all other respondents corrected and updated the Summary.

Use of the calendar/diary between interviews was defined as recording at least some health care events *and/or* saving bills, receipts, and appointment cards in the pocket of the calendar. Thus, people who for one reason or another were unable to write were given an equal opportunity of being rated as competent respondents. The findings are that during the four periods that intervened between Rounds 1 and 5, the interviewers thought that an average of 51 percent of the poor, 62 percent of the elderly, and 66 percent of all other respondents used the calendar (Table 5). A slight drop in use of the calendar was perceived for all groups from Round 1 to Round 5, the drop being 4 percent for the poor, 1 percent for the elderly, and 3 percent for all others.

Another item that in a broad sense might be said to be related to respondent competence was whether respondents understood that signing permission forms meant that their medical providers, employers, and health insurers would be contacted for additional information. The interviewers' estimate was that 81 percent of the poor, 84 percent of the elderly, and 91 percent of all other respondents understood the implications of signing permission forms (Table 6).

Finally, the question was put to the interviewers directly: "What percent of respondents understood their role as respondents, that is, kept records, reviewed the Summary with other

Table 4
Estimated percentage of reporting units that corrected and updated the Summary prior to the interview (N = 222)

	Corrected and updated Summary			
Reporting units[a]	Most	Some	Few	None
Poor	2	15	35	47
Elderly	5	19	39	37
All other	9	42	38	11

[a] A reporting unit consisted of persons related by blood, marriage, or adoption. Included were foster children, but not unmarried students aged 17–22 living away from home. The latter formed their own reporting units.

Table 5
Estimated percentage of reporting units using calendar[a]

	Use of calendar			
Reporting units	*Between Round 1 and Round 2*	*Between Round 2 and Round 3*	*Between Round 3 and Round 4*	*Between Round 4 and Round 5*
Poor	53	53	51	50
Elderly	62	62	61	61
All other	68	67	65	64

[a] Each interviewer estimated the percentage of respondents who used the calendar for each round. The figures reported are the mean values of these estimates.

Table 6
Estimated percentage of reporting units who understood implications of signing permission forms

Reporting units	*Mean of interviewers' estimated percentages*
Poor	81
Elderly	84
All other	91

members of the household, and reported fully and accurately?" The interviewers estimated that, on the average, 52 percent of the poor performed their role reasonably well and that 61 percent of the elderly and 69 percent of all other respondents did so.

The picture that emerges appears to be not so much one of vast differences in respondent performance between population groups—although in some instances the differences are significant—but rather one of respondent performance in general. For example, the fact that a 3 percent difference is perceived between the poor and the elderly and a 10 percent difference between the poor and all others regarding comprehension of the implications of signing permission forms is less problematic than the fact that a fair number of all respondents appear not fully to have realized what they were signing. Another case in point is the calendar/diary, although here we do find a more significant difference between population groups, with a 9 percent perceived difference between the poor and the elderly and a 15 percent difference between the poor and all others. But what is more significant is that even among all other respondents, the interviewers' impressions are that one-third do not appear to have made even minimal use of the calendar.

In the case of correcting and updating the Summary, we again find the elderly perceived as doing somewhat better than the poor, and even though all others are perceived as doing considerably better than both these groups, 50 percent of the interviewers thought that only a few or none of all other respondents corrected and updated the Summaries.

The two tasks that required active and independent participation by the respondents—independent participation in the sense that respondents had to perform tasks between rounds without the overt stimulus of the interviewer's presence either in person or on the telephone—were to record on the calendar and to correct and update the Summary. Both tasks were important tools in the attempt to obtain complete and accurate data; they were also the two tasks that required a fairly high degree of sophistication on the part of respondents. Yet we find that in the interviewers' estimation the tools were not utilized to the extent that they might have been, and when they were used, it was to a somewhat lesser degree by the poor and the elderly than by all others.

What lesson is to be learned from these findings by those whose task it is to implement health survey research in the future, assuming that hard data will eventually corroborate the opinions of the interviewers? Some indications seem fairly clear and relatively simple to implement. For example, it appears that on complex health questionnaires, at least the elderly should not be interviewed on the telephone. In this instance, one need only instruct interviewers that where the main respondent is 65 or older, all interviews must be conducted face-to-face.

What to do in practical terms about having respondents make greater use of memory aids, or how to make more people understand what the full implications are of signing permission forms, is more difficult. One apparently cannot take for granted respondent competence, and in pretesting the instruments, one learns about the comprehensibility of instruments and how respondents behave during the interview. However, if one is going to ask for *independent* respondent participation, that is, keeping records *between* interviews and correcting and updating the Summary *prior* to the interview, some additional stimuli are apparently needed.

Oksenberg, Vinokur, and Cannell (1977) have experimented with the concept of commitment to being a good respondent, based largely on the theories of Kurt Lewin, who was concerned with the factors that affect an individual's acceptance of a goal and the effects of goals on behavior and cognition. The general findings of their experimental study were that a commitment procedure, that is, signing an agreement to provide very accurate and complete information, is a workable interviewing technique and can lead to improved respondent performance. Based on these findings, NMCES did, in fact, have respondents as well as interviewers sign such an agreement; presumably similar effects were obtained as in the experimental study, although no controls were established retest the procedure.

The issue of respondent competence, however, is qualitatively different from commitment to being a good respondent. NMCES obtained written commitment, and open-ended responses from the interviewers attest, as do high response rates in every round, that respondents felt themselves to be part of an important research undertaking and generally acted the role of a committed study population. The question is whether, given their commitment, some people, and particularly the elderly and the poor, were cognitively able to perform the tasks that were asked of them.

In the Oksenberg, Vinokur, and Cannell (1977) experimental study, education level appeared to be positively correlated with several experimental conditions, including amount of information, precision of reported dates, and accuracy and completeness of reports, although no claims were made for a causal relationship

between education and these experimental conditions.

In addition, the findings of Fowler, who made an analysis in 1965 of reporting by educational groups in the Health Interview Survey, were reported by Cannell, Marquis, and Laurent (U.S. NCHSR, 1977:14) as follows:

> Based on systematic observation of interviewer and respondent behavior he [Fowler] concluded that less highly educated respondents needed more help from the interviewer to perform adequately. They were less skilled at the respondent role. There was also the tendency for the less educated to have less information about the purpose of the survey and what was being sought in the interview.

On the other hand, we also learn from Cannell et al. (U.S. NCHS, 1977:11-13), with reference to other major health surveys in which record checks were made to establish validity of reporting, that no particularly strong tendency was found for higher-educated respondents to report more accurately than respondents with less education.

The evidence for equating high educational level with respondent competence on health interview surveys is tenuous at best. Yet one cannot escape the intuitive sense that if supervisors and interviewers required a full day of intensive training in the use of the Summary, and that if 33 pages of specifications and illustrations were needed to explain the Summary to them, respondents with low education were at a disadvantage, regardless of their level of commitment. Of the two tasks that required independent respondent participation, that is, maintaining the calendar/diary and updating and correcting the Summary, we know only that the former has been used successfully on previous studies and has been shown to be a valuable tool for increasing the level of reporting on health surveys (Wright et al., 1976). Computer-generated Summaries of previously reported data, however, were a pioneering effort on NMCES, and until hard data become available, we have only the reports of those who designed the Summary and of those who trained supervisors and interviewers in its use, as well as the impressions of the interviewers about respondent use, as expressed in the Interviewer Survey.

Conclusions. As stated earlier, interviewers thought that actual correcting and updating of the Summaries prior to the interview was relatively low for all groups. However, in response to the item on whether respondents had Summaries available for the interview appointment, we found that 76 percent of the interviewers thought most or some of the poor had them available, 88 percent thought most or some of the elderly did so, and 92 percent of the interviewers thought that most or some of all other respondents had them available. We have then a disjunction between active and independent respondent participation, which was relatively low, and passive respondent participation, that is, "let the interviewer show me," as it were, in that interviewers report large numbers of all three population groups as having had the Summary available for the interview.

Although having the Summary available does not accomplish the intent of the Summary, one might conclude that respondents were sufficiently interested in the Summary to save and produce it for the interview, even if they were unable or unwilling to correct and update it themselves before the interview. Based on these findings, we would speculate that if the Summary can be made more comprehensible to all sectors of the population, it is probably a valuable tool for increasing the level of reporting. To health survey planners we recommend, therefore, that Summaries be simplified and that some simple didactic materials be developed. For example, a booklet with illustrations and simple instructions on how to correct and update the Summary should be used by the interviewer in explaining the Summary and should then be left with the respondent. Also, when telephoning to make an appointment for the next interview, the interviewer might suggest to the respondent that it would be helpful if the other members of the household had an opportunity to review the Summary before the interview to be sure that the information was accurate and complete.

As for increasing the use of the calendar/diary, an interim phone call from the interviewer to the respondent might help. In this call, the interviewer might tactfully inquire whether the respondent is maintaining the diary and express appreciation for doing so. We know that positive feedback from interviewer to respondent *during* the interview increases the level of reporting. It seems reasonable to extend the technique when independent participation on the part of respondents is required *between* interviews.

Discussion: Survey of interviewer attitudes toward selected methodological issues in the National Medical Care Expenditure Survey

Robert Wright, National Center for Health Statistics

As was noted at the Second Biennial Conference, and in many other places, survey researchers have been seeking for years the means of identifying and measuring errors in their estimates. Horvitz and several others have advocated the concept of Total Survey Design for optimizing the quality of estimates for a given amount of resources. During this conference we have heard and discussed a number of ways for improving survey methodology; most of these methods have talked about nonsampling kinds of errors. I think Sir Josiah Stamp of the Inland Revenue Department in England (1896–1919) probably said it best:

> The Government are very keen on amassing statistics, they collect them, add them, raise them to the nth power, take the cube root, and they prepare wonderful diagrams, but you must remember, never forget, that every one of the figures comes, in the first instance, from the village watchman who just puts down what he damn pleases.

Fleishman and Berk have done something that we probably all do, but the results are rarely published or acknowledged in any formal or systematic way. They have attempted to measure a number of the variables that shed some light on the process of interviewing and on the attitudes of the interviewers, which are a primary source of measurement variance and systematic errors in the measurement process, as discussed at the last conference by Kalsbeek and Lessler (1978).

In pretests, it has long been accepted that you observe the interviews and hold discussions with the pretest interviewers to gain some insight into what is happening with your procedures, instruments, and instructions. While the results are usually anecdotal in nature, a few cases can alter the course of a whole survey methodology. It is rare, however, for survey researchers to ask the interviewers what they think of the survey itself after completion. It might be dangerous. It is probably even more rare for the results to be published. In the Medical Economics Survey, which Yaffe reported on at an earlier session, the interviewers were asked to complete a short questionnaire and to take part in a group discussion or debriefing at the end of data collection; the results were not widely distributed, yet they had a major impact on some of the decisions in the design of NMCES and are still having an impact on the design of the new National Medical Care Utilization and Expenditure Survey. For example, it was the interviewers in that debriefing who suggested that one reason for the lower retention rates in that study was the alternating telephone/personal/telephone/personal methodology, which was causing respondent dissatisfaction. The interviewers reported that the respondents could not understand the need for personal interviews after they had once been called on the telephone.

The NMCES Interviewer Survey, reported in the paper by Fleishman and Berk, is significant because the number of interviewers involved in the survey and the potential for real insight based on this report are quite large. The paper's discussions of the effectiveness of selected field procedures and the attitudes of interviewers toward their work, their effect on response rates, face-to-face versus telephone interviews, and the comparisons for the poor and the elderly are very important. The interviewers are the front-line troops in the battle of response rate, accurate reporting, and recording of information. They must implement the sometimes bizarre procedures and instructions from those of us who are removed from the firing line.

A question that I would ask then is why the survey was restricted to the 280 interviewers who participated in at least five of the six rounds. It would seem that some attempt should be made to collect information from interviewers who left the survey early or who joined it late. Although some of the useful analysis could

not be done with those interviewers, it would be helpful to know why they quit or the problems they had in joining an existing survey where established jargon and history existed that they had to learn. Such information might have provided a different perspective on such questions as how suitable different types of interviewing methodologies were for different groups.

Since surveying the interviewers is a relatively rare event in the annals of survey research, less is known about constructing questions and determining useful information for such a survey than is known about constructing questions to solicit responses in household surveys. Some comments about the process do come to mind, however. Most questions asked of the interviewers, as of the household respondents, are susceptible to recall error. Hard recorded data to which they can refer are not readily available to the interviewers to tell, for example, how many respondents corrected and updated the Summary prior to the interview. Because responses from the interviewers are subject to recall error, it is important that such a survey be conducted as soon after completion of the survey field activities as possible. For many reasons not within the control of either Fleishman or Berk, this survey of interviewers took place several months after the end of field activities.

In addition, the questions asked of the interviewers must be carefully worded. An example is the question on ranking the importance of the various means of increasing the response rate. The interviewers were given as one of the choices, "Offer fuller and better explanation of the study at Round 1." The conclusion of the authors is that since this means of increasing the response rate was ranked very high, telling the respondent in the advance letter that the survey would consist of seven interviews and that it would last for about 15 months might have increased participation. Although this is certainly possible, the phrase "offer fuller and better explanation of the study" has no concrete definition. In fact, it is quite likely that many interviewers who ranked it first or second as a means of increasing the response rate interpreted it to mean that more explanation of the purpose of the study, more emphasis on the importance of the respondents' participation in the study, and simpler or more comprehensive instructions for completing the diary and working with the Summary were the keys. These certainly fit the concept of "fuller and better explanation of the study."

The question of suitability of interviewing method is another example where many interpretations are possible. The overall response that face-to-face interviewing is most suitable is not unexpected. The survey designers established, before the survey began, that for some contacts, a face-to-face interview was required and more suitable. Thus, a better question would have been to isolate the suitability of the method for contacts where an option existed, namely, in Rounds 3 and 4.

The proportion (34 percent) of the interviewers who preferred using the telephone to interview college students is also interesting. The concept of suitability has many facets, including the ability to get reliable data without the severe problems of trying to explain the question over the telephone. The easier access that the interviewer has to college students in terms of the distance from the interviewers and the ease of making the frequent checks that have to be made in order to catch the students at a time and place where the interview can be completed, plus a generally assumed lower frequency of health events to report, makes for less-complicated interviews and might make the telephone the medium of choice for those college students. Evidence that the less-complicated interviews are more suitable for the telephone was also pointed out by Fleishman and Berk, and data not presented in this paper show that interviewers have indicated a preference for the telephone as most suitable for interviewing young, noncollege adults, people with small or no families, and nonfamily groups.

The comments about being able to see medication bottles, bills, and the calendar also support interviewing selected groups by telephone, since the interviewers may sense that well-educated, high-income persons understand the questions and do not need the interviewers' physical presence to be good respondents.

An area where interviewer reports have a clear impact on future survey planning is the comparison of poor respondents and elderly respondents with the rest of the population. The interviewers' impressions and problems with these groups are probably not subject to differential recall error. Although there may be differences in the exact definitions used by the interviewers (and you note that no exact definition was practical since the interviewers did not have any exact data available), the general impression of differential problems in certain areas is extremely useful. Care must be exercised, however, not to be misled by a strongly presented case, which is a rare event.

As I said earlier, the interviewers are the front-line troops in the battle to obtain accurate information. All of the messing around that we do as analysts, question writers, and the like are dependent on those interviewers. No amount of effort at these tasks is effective if the interview-

ers just "put down what they damn please." It is, therefore, very important that we understand the battle as the front line sees it. A survey of interviewers is an excellent method for obtaining some of this information, and I would recommend more use of the technique and publication of the results.

Some methodological issues raised by the National Medical Care Expenditure Survey

Gail Roggin Wilensky, National Center for Health Services Research

The National Medical Care Expenditure Survey (NMCES) is a large health survey designed to provide data for a major component of the Intramural Research Program at the National Center for Health Services Research (NCHSR). It focuses on major issues of national health policy, such as the implications of alternative national health insurance proposals in terms of cost, use, and financing; the effects of Medicare and Medicaid on use and costs of personal health care; access to care; tax treatment of medical expenditures; depth and breadth of health insurance coverage; and costs of illness for different diagnostic categories. The survey was also designed to provide data relating to a variety of methodological issues in survey research. A brief discussion of the most important of these methodological analyses is the subject of this paper.

NMCES consists of three major surveys and two smaller surveys. The major surveys are (1) a survey of 13,500 randomly selected households each interviewed six times over a 15-month period during 1977–78; (2) a survey of the physicians and hospitals that provided care to household respondents during 1977—the Medical Provider Survey (MPS); and (3) a survey of insurance companies and employers responsible for the insurance coverage of respondents.

The primary source of information for the population comes from the Household Survey. The sampled families were asked a core set of questions about health services use, expenditures, and health insurance coverage for each family member. Special supplementary questions were also administered periodically throughout the year. A major innovation associated with NMCES is the use of a computer-generated Summary. Beginning with the second interview, the Summary was mailed to each of the sampled families a short time before the expected date of the next interview. A small group of families was designated as a control group and did not receive the Summary until the time of interview. The Summary contained the most salient information about medical and dental visits, charges, and health insurance coverage reported during the previous interview(s). Respondents were urged to review the Summary with other household members prior to each interview and to correct or update the Summary during the interview process. (More detailed information about the Summary and Summary review process is contained in the paper by Holt presented earlier in this session.) The household information is verified and supplemented by information from the MPS and the Health Insurance/Employer Survey. The two smaller surveys will provide data on the characteristics of physician practices and employer insurance costs.

The NMCES data are being collected by the Research Triangle Institute and its two subcontractors, National Opinion Research Center and Abt Associates. It is funded by NCHSR and cosponsored with the National Center for Health Statistics.

There are two major types of methodological studies that the NMCES data can support. The first involves basic approaches to the collection of data; the second involves alternative measurements of specific data elements. The purpose of this paper is to enumerate major research issues in each of these areas and to outline several of the planned analyses. We hope this will encourage other researchers to think about the potential of the NMCES data for methodological research and to share with the two Centers their ideas about additional research topics that can be supported by this large data base.

Basic approaches

There are several studies on basic approaches to collecting data implicit in the NMCES design. These include the following:

1. Use of a panel design in a household survey with and without an interview summary;
2. Comparison of the interview survey with a medical provider survey as a means of correcting and updating household reported visits and charge data;
3. Use of a household survey in combination with a medical provider survey relative to use of health insurance claims data;
4. Comparison of household reported health insurance coverage with the collection and coding of health insurance policies from health insurance companies and employers;
5. Use of a usual source of care survey and an uninsured persons validation survey as a means of determining false negatives;
6. Use of telephone versus personal interviews, in terms of both the interviewing mode and the proxy versus self-reporting problems; and
7. Comparative costs of obtaining different response rates in a medical provider survey.

Each approach implies a major research effort. In some cases the NMCES data are well suited for analyzing the implications of using one approach versus another. In other cases they can provide only indirect evidence on the relative merits of each approach. In all cases an attempt must be made to quantify the relative accuracy of reporting as well as the costs associated with each method of data collection, holding all other differences constant. This will eventually enable us to estimate the benefits associated with a particular method as well as its costs, thereby allowing an estimation to be made of benefit-cost ratios for the respective approaches.

Use of comparative data collection procedures. A difficulty that will affect many of the NMCES methodological studies is that the different data collection approaches were incorporated in the design to support substantive rather than methodological analyses. As a result, the data collection strategy does not provide as tight an experimental design as would have been desirable for evaluating some of the alternative methodologies. There are, on the other hand, some advantages to observing the effects of alternative methodologies over the course of a large ongoing survey rather than limiting the analyses to small case studies. Although the lack of rigor associated with the introduction of alternative data collection strategies may affect the generalizability of some of the findings, we can probably introduce sufficient control into the analyses to permit some tentative conclusions about the effects of these different strategies.

One of the most important NMCES methodological research areas relates to the Summary, since it was a major innovation in data collection strategy. It must be recognized at the outset, however, that we cannot evaluate the effectiveness of a panel design survey with and without a summary because such an evaluation would require a control group who never received a summary and who were asked questions during the interviewing process that presumed no opportunity for future updating. This is an important issue that we hope a future study, such as the National Medical Care Utilization and Expenditure Survey, will address. We can, on the other hand, assess the effectiveness of using a summary compared with using a medical provider survey as a means of correcting household survey data, that is, comparing data on medical expenditures and visits, sources of payment of the total charge, and date of the visit as recorded on the updated summary and as reported by the medical provider.

Then, after calculating the mean differences between the estimates provided by the household interview summary and the medical provider survey and the distribution of the differences across relevant domains, the issue becomes: Is the summary worth its cost? The first step toward answering this question is to estimate the cost of producing the summary. Detailed time and expense data from the survey will enable us to make these calculations. However, to establish whether this cost was justified, it is necessary to calculate what the costs and use estimates would have been had we relied solely on a panel design household survey. Are there subgroups of respondents who, because of their ability to report accurately, become candidates to be excluded from the summary process in future panel design studies? Conversely, are there other subgroups where the data from the medical provider survey produce substantially different estimates of total cost, components of charges, or numbers of visits than those provided by the households?

To the extent that such domains can be identified, it may be possible to develop research designs where use of summaries is appropriate for certain domains, use of a medical provider survey is preferable for others, and use of both is required for few, if any, domains. For example, one might expect the Medicaid population to make few substantive changes on their summaries because much of the updating occurs as a result of settlements with insurance carriers. In addition, the Medicaid population might report many of their charges as unknown since they are unlikely to receive a bill. If this expectation is true, it would suggest that the Medicaid

population should be included in a medical provider survey but might not need to receive a summary. On the other hand, households with private insurance coverage might be expected to make major changes on their summaries but report with sufficient accuracy so as not to warrant inclusion in a medical provider survey. The elderly might require a summary, because of their interactions with Medicare and private insurance companies, and also inclusion in the MPS because of problems with recall.

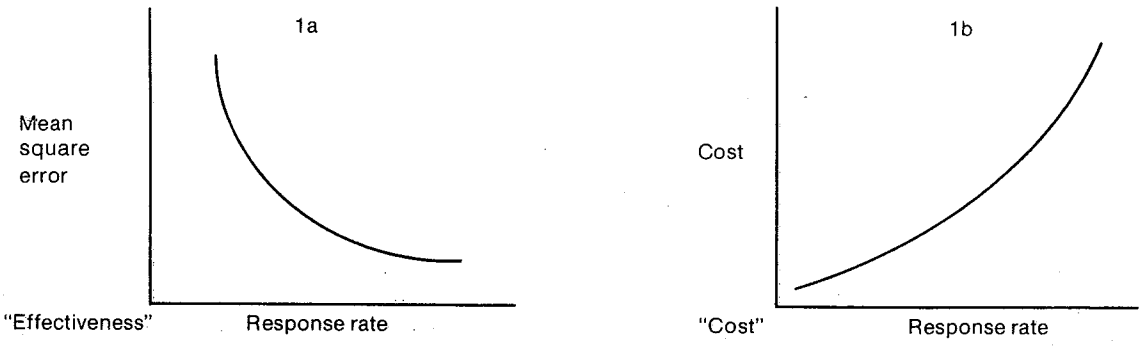
There are other tools of survey research available, such as diaries, other types of record check surveys, mode of interview, time between interviews, etc. It is time for researchers to consider choosing the combination that produces a given level of accuracy at the least cost or the highest level of accuracy for a given cost rather than applying all possible methods to all respondents, irrespective of available evidence on their reporting accuracy.

Costs of medical provider survey. Another area of methodological research that the NMCES data can support involves the implications of accepting different response rates, i.e., lower response rates on a medical provider survey. The response rate for fee-for-service physicians in our MPS was approximately 87 percent. Although the data in hand will not allow us to evaluate the effects on precision of going from 87 percent to 100 percent, we will be able to estimate the costs and at least a few of the measurable effects of stopping at a response rate that is lower than 87 percent. Costs can be estimated on a per interviewer basis in terms of the costs associated with cases completed as of particular dates. This will allow an estimation of the cumulative costs as of a particular time along with the response rate achieved as of that time. We plan to estimate the effects of stopping a medical provider survey at response rates of 70, 75, 80, and 87 percent in terms of the increase in mean square errors for critical estimators and also the impact on prototype multivariate analyses of medical care use or conditional probabilities of use in national health insurance simulations.

Figure 1 diagrams the cost-effectiveness analysis outlined above. There is general agreement about the direction of each of the curves: As the response rate increases, the mean square error (MSE) declines (Figure 1a); as the response rate increases, the costs also increase (Figure 1b). The shape of the curves in terms of slope and functional form are not known but can be estimated using some simplifying assumptions. While there may be important quantitative or qualitative effects associated with varying response rates that are not measured by (a) reductions in MSE, relative variance, or other error measures and (b) sensitivity analyses on alternative measurements of variables, this type of approach will at least give guidance to survey researchers and substantive analysts who must make decisions on how to allocate scarce resources. This approach to allocating resources according to their relative effectiveness in reducing MSE and across various components of survey design is the economist's version of Total Survey Design—a concept strongly advocated by Horvitz at the First and Second Biennial Conferences on Health Survey Research Methods.

The methodological studies just discussed present comparisons of data collection strategies as though only two or three variables at a time were changing. It is clear, however, that a great number of factors affect the relative accuracy of reporting. Since most NMCES data collection strategies were not employed in controlled experiments, these studies will require multivariate analyses rather than analyses based on simple statistics. Conceptualizing the accuracy of response as a multivariate problem would produce the following type of equation:

Figure 1



$$\frac{\text{Household value}}{\text{MPS value}} = f \text{ (length of interview, mode of interview, length of reference period, self versus proxy respondent, respondent characteristics, type of visit, etc.)}$$

The dependent variable might be the ratio of the value reported by the household to the value reported by the provider or the difference between these values. Alternatively, the dependent variable might be a comparison of a household or MPS estimate with an insurance claim estimate. Accuracy in this equation of reporting, however measured, is presumed to vary according to the length of the interview, the mode of the interview (telephone versus personal), the recall period, proxy versus self-reporting, respondent characteristics (e.g., age, sex, education), and type of visit. It is possible that some of the variables may be endogenous, requiring simultaneous estimation procedures rather than single-stage estimation. More serious consideration needs to be given to the formulation of the equation, the specification of the variables, and the appropriate estimation procedures than is possible here.

Research on specific data items

The second type of NMCES methodological study involves alternative measurements of specific data elements. Because of the size and complexity of the NMCES design, there are many opportunities for observing the effects of alternative specifications of a given data element. Three specific data element studies of particular interest are

1. Experimental occupation coding
2. Alternative definitions of medical expenditures
3. Alternative definitions of insurance coverage

Occupation coding. The NMCES occupation coding experiment is an extension of a study on respondent reporting of occupation by Garth Taylor (1976). In all SMSAs other than New York, NMCES respondents were asked to choose the category that best described their occupation. The respondents were shown a card that displayed 12 "descriptor categories" of occupations. These categories reflected the ten summary categories into which Census codes are classified, with two of the more heterogeneous categories split into two groups each. In order to avoid some of the problems identified by Taylor in his study, categories were defined in terms of examples of occupations to which we thought individuals could relate rather than through the use of the definitional descriptions used by the Census. The New York SMSA was used as an experimental group. Households in this SMSA were randomly assigned to four groups, each receiving the occupation descriptors in a different ordering. The respondents were than asked the traditional Census occupation questions so that their responses could be coded using traditional procedures. The experiment also included a control group in the New York SMSA who answered the Census questions first and then designated their occupations. The use of respondent-designated occupation codes represents a substantial reduction in coding and data processing costs. Whether this procedure represents an attractive alternative depends on its effect on reporting accuracy, which we will ascertain, and on the importance of detailed occupation coding to the research goals. The latter will obviously differ by study.

Medical expenditures. The second study involves the effects of alternative definitions of medical charges (the price of medical care). As an economist, I am particularly interested in the role of price in the demand for medical care and the effects of changes in price associated with alternative national health insurance proposals. As a survey researcher, it is not clear to me which of several different specifications represents the best measurement of price. These specifications include the following:

1. As initially reported in the interview. (This may be the best measure of "perceived price" in that it reflects what the individual thought would be the gross price [i.e., total cost] and the net price [i.e., cost to the individual] before all components were known and adjustments settled.)
2. As updated on the Summary. (This represents the "best" measurement of price as reported by the respondent.)
3. As implied by the Health Insurance Supplement. (The perceived levels of coverage reported on the Supplement provide us with a perceived coinsurance rate for 15 types of care, from which a net price can be calculated. There is concern that because the information on the Supplement does not refer to a specific event, the questions may have been too abstract to have produced reliable responses.)
4. As implied by the health insurance policy. (This measurement reflects the "actual" coinsurance rate, i.e., the share of the cost to the individual assuming the insurance plan is

administered according to the written provisions in the policy.)
5. As reported by the provider.
6. As reported in the claim.

Given these different specifications, the issue is how different the substantive findings of the research are for any given set of medical visits. For example, how different would the results be in an analysis of utilization and such patient characteristics as income, age, sex, region of country, and health insurance coverage if one specification is used rather than another, or even if some linear combination of these measures is used? The answer is clearly an empirical matter, although the choice among these measurements could be made on either theoretical or empirical grounds.

Insurance coverage. The third study involves alternative measurements of insurance coverage. The size of the insured and the uninsured population is currently a controversial topic of health policy, since in the absence of more relevant data, such as the adequacy of health insurance coverage, the size of the uninsured population is being used as an argument on behalf of national health insurance. Since a major goal of NMCES is to provide a data base for national health insurance simulations, considerable care and effort have been devoted to the measurement of health insurance coverage. Our problem, in the short run, is one of an embarrassment of riches, because we, in fact, have at least six specifications of insurance coverage. These include the following:

1. Insurance coverage as reported in the household questionnnaire, i.e., "direct report."
2. Direct report augmented by "source of payment" information. (Individuals who report no insurance or no Medicaid or Medicare but who report private insurance, Medicaid, or Medicare as paying part of their bill can be reclassified as having private or public insurance.)
3. Direct report as updated on the Summary.
4. Coverage as reported on the Health Insurance Supplement.
5. Direct report augmented by MPS source of payment information. (Medical providers were asked to report the amounts of the bill that were paid by each source of payment. Individuals whose bills were reported to have been paid by an insurance source can be reclassified as having that type of insurance.)
6. Validation of direct report coverage by the Health Insurance/Employer Survey. (Employers and insurance companies were asked to verify insurance coverage and provide copies of insurance policies for respondent reported insurance coverage. This will permit adjustment of the direct report for false positives, i.e., overreporting of coverage.)
7. Validation of absence of health insurance coverage by employers. (Employers of respondents who reported no group health insurance coverage were asked to verify the absence of such coverage. This will permit an adjustment of the direct report for false negatives, i.e., underreporting.)

Again, the methodological issue is the extent of the difference between the estimates produced by different specifications and the amount of impact of alternative measurements on both the findings on the distribution of insurance coverage and national health insurance simulations. As with the measurement of price, the magnitude of these differences is an empirical issue. For an economist, the choice among these measures is clearer theoretically than with the price variables, although recognition that all records and reports contain errors complicates the calculation of the "true" value.

The policy relevance of alternative measurements of the uninsured establishes the link between methodological and substantive analyses. By themselves, the specification of the price variable, the definition of the uninsured, and the specification of the visit variable can have a substantial impact on analyses of costs and use of care, access to care, the interaction between illness and economic well-being, and other major health policy issues. This link has so far not been well understood either by analysts or by methodologists. We hope that the methodological research done on the NMCES data, in conjunction with their substantive analysis, will help to demonstrate this linkage.

Open discussion: Session 5

Use of Summaries in NMCES

In opening the general discussion on Holt's paper, Gift asked how the situation was treated when there was a change in the report of a visit. That is, at one point in time a respondent said a visit was made; later the respondent said the visit had not taken place. Wouldn't that type of change be considered as "zero" or an absence of data rather than a change to an unknown value? Holt replied that the type of comparison coding that was done would not have measured that a respondent initially reported a visit and then later said that the visit did not occur. Questionnaire visit reports were matched on a one-for-one basis with Summary visit reports where both visit records were located. Where visit reports could not be located on the Summary, no further coding was undertaken. About 7 percent of the visits in the sample of 450 were in this category.

In response to a question from Kalsbeek about correlations between respondent characteristics and types of changes made on the Summary, Holt indicated that there has been no demographic analysis in relation to Summary changes.

Massey asked for an overall evaluation of the use of the computer-generated Summary, based on NMCES experience. He further asked why the particular Summary methodology was used and whether such an approach would be considered in a similar survey in the future.

Horvitz responded that the size of the survey dictated that the Summary process be totally automated; this approach led to a number of data processing problems. The effort could have been minimized by more selectivity of the data that were processed for inclusion on the Summary. One alternative could be to key only responses where critical data are missing, incomplete, or inconsistent. Another possibility is to incorporate computer-assisted telephone interviewing technology to secure missing data. This approach would permit error resolution on an ongoing basis independent of the primary field data collection effort.

Horvitz pointed out that many people may not have been aware that the original NMCES schedule required seven rounds of interviews—one every eight or nine weeks—during 1977. In reality, four rounds of interviews were completed. The main reason that the interviews could not be turned around as planned was the tremendous volume of data processing required to produce the Summaries and get them to the field.

Oksenberg pointed out a feature of the survey design that may have contributed to the high rates of change on the Summaries. The questionnaire and data collection procedures were designed with the knowledge that the Summaries would be used. Consequently, interviewers may not have pushed as hard to secure missing data or to resolve discrepancies. Fewer changes might have been observed on the Summaries if the interviewers had required the respondents to work harder in the initial survey rounds.

Yaffe noted that the Summary change frequencies reflect not only respondent changes but also corrections of interviewer recording errors and data processing errors.

In her paper, Holt had commented on reporting limitations of certain categories of NMCES respondents. Fowler asked what percentage of the population was comprised of persons enrolled in Medicare, Medicaid, HMOs, and PHPs, whose ability to report expenditures was affected by their delivery status. Walden estimated that 20 to 25 percent of the population would be included in these groups.

Wilensky noted that Holt's conclusion about reporting limitations was correct for recipients of direct service programs. The limitation may not apply to the Medicare population, since they should receive complete documentation of their medical transactions. The reporting limitation

may be a function of age rather than of the nature of their third-party-payment status.

Interviewer attitudes on methodological issues in NMCES

Beginning the general discussion on the Fleishman and Berk paper, Rothwell noted, with reference to the 84 percent response rate for the NMCES Interviewer Survey, that the Bureau of the Census had experienced a similar situation where an interviewer survey was conducted with a response rate of 80 percent, although the interviewers who were surveyed had achieved a response rate of 95 percent for the population that they were surveying. She then outlined an alternative approach that yielded a response rate of 100 percent and had the apparent advantage of requiring less of the interviewer in terms of generalizing about and classifying respondents. Interviewers were queried about the last respondent whom they had interviewed in a survey. The questions included on a summary sheet were specific to the immediate interview setting. A small sample of respondents was then asked the same questions about the interview that the interviewers were asked. This approach allows the interviewer to be objective about a single respondent and interview setting, while providing control by allowing comparisons of the respondents' perceptions, as well as characteristics of the respondents as reported in the main interview.

Warnecke commented that the preference expressed by interviewers for personal interviews over telephone interviews was confounded by the fact that the same interviewers were conducting both types of interviews and that the same instrument was used, which had apparently been designed primarily for personal interviews. He wondered if it would have been more economical to have hired two separate groups of interviewers, one for the personal interviews and the other for the telephone interview rounds. He further commented that the Survey Research Laboratory's experience with elderly populations showed that they preferred not to be bothered with personal interviews.

Fleishman commented that the instrument used in the telephone rounds was designed for the method limitations of telephone interviews. Zinner said that hiring two separate interviewing forces would have been extremely difficult, not only in terms of costs but also because it was impossible to predict in advance which respondents would require personal interviews.

A number of comments were made on the issue of informing respondents of survey requirements and securing their commitment to participate. Verbrugge asked if there were any tests of alternate strategies in informing respondents about the nature of an expected long-term commitment. Her experience in the Detroit Health Study showed that respondents who were not willing to make a long-term commitment tended to drop out at the onset of a study.

Axelrod reported experience in informing respondents that they "might" be recontacted in the future. He suggested that it would seem desirable to ask respondents some of the questions that were asked of the interviewers in the survey. For example, the questions on length of interview, amount and frequency of payment, and understanding of permission forms or other survey features would seem applicable. He then asked about panel attrition across NMCES interview rounds, citing previous rates of about 10 percent per round.

Horvitz commented that the NMCES rates per round never approached that level. According to Verbrugge, most studies reflect high rates of attrition at the onset, with diminishing attrition over time.

Bergner expressed concern over the practice of not informing respondents that they may be recontacted. She then described a test/retest procedure involving 100 respondents who were asked to respond to identical questions 24 hours after the initial interview. Respondents were informed at the time of the first interview that a second interview would be required 24 hours after the first. The response rates for persons in the test/retest group were essentially the same as for persons who were asked to respond to a single interview. In other panel designs, admittedly with small numbers of respondents, she has not found response rates to be different from single interview surveys. It does not seem to make a difference when respondents are told that multiple interviews will be required.

Addressing Warnecke's point about doing telephone interviews with personal interview instruments, Walden pointed out that one of the questions on the NMCES Access to Care Supplement was specifically modified for administration in telephone interviews. Responses to the NMCES telephone version can be compared with Andersen/Aday data from personal interviews conducted in 1976.

Fleishman commented that a large number of respondents were reported as having their copies of the Summary available at the time of the interview. She further noted that the Summary is not a neutral data collection instrument the way a questionnaire is. That is, the Summaries have an effect of their own—they are a personal printout that requires action on the

part of the respondent. The Summary produces a reaction in itself.

In concluding the discussion, Horvitz noted that there is a question of when you tell respondents that you want them to participate in a panel survey. The answer to this question depends in part on the types of questions that you are going to ask them. The NMCES respondents received an advance letter advising them that their participation was voluntary. At the end of the first interview, they were offered the $5 incentive and informed that more interviews were to come. It is unfortunate that no alternate procedures were tested in NMCES for differential effects based on whether or not respondents were informed of the panel interviews before or after the first interview.

NMCES methodological issues

In opening the discussion on the Wilensky paper, Horvitz observed that a further aspect of NMCES methodology centers on the fact that RTI and NORC shared the data collection responsibilities for NMCES, which permits a unique, but not original, evaluation of organizational effect. Horvitz also expressed disagreement with Oksenberg and Wilensky's thesis that one cannot look at NMCES Household Survey and Medical Provider Survey data only versus Household Survey and Summary data only.

Andersen commented on Wilensky's assumptions about the impact of response rates on mean square error. The assumptions that were stated appear to ignore the problems of item and total nonresponse. He then asked if Wilensky planned to examine different imputation procedures.

Wilensky responded that the analysts can look at how good the data would have been with differential levels of response. She indicated that there is a difference in item nonresponse versus total nonresponse and that one cannot include all nonresponse in reducing mean square error. The survey was designed from the outset to achieve a high response rate; in retrospect, this demand may not have been appropriate. The Medical Provider Survey was not an independent record check because it was derived from the Household Survey; it allowed for corrections to Household Survey values and for imputations for household reports that were not included in the record check.

Andersen further commented that there are different ways of using imputation. The practical issue of response rates depends on how nonresponse is treated. Wilensky observed that analysts will need to simulate different levels of nonresponse and the effects on estimations or adjustment procedures. Horvitz noted that a sample of nonresponders is available in terms of sample members who refused to sign permission forms. Values will have to be imputed for these people.

An interesting analysis, suggested by Sudman, would be the results of respondent conditioning effects over time. The NMCES data should be analyzed to this end. Two particular examples are the learning process related to the Summary and fatigue associated with the longer interviews.

The topic of organizational effects prompted several questions and comments. Axelrod asked what the comparative costs per interview were for each organization. Horvitz replied that because of different charging algorithms, it is difficult to make line item comparisons. The bottom line comparisons, however, have shown little difference in costs between the two organizations. Weekly field status reports for individual organizations provided competitive motivation between RTI and NORC, according to Fleishman. Preliminary MPS response rates for fee-for-service physicians were 86.5 percent for NORC and 86.4 percent for RTI.

Elinson asked who will conduct the planned comparisons of organizational effects. Horvitz indicated that he and Steve Cohen of NCHSR would make the comparisons.

Yaffe observed that competition between the two data collection contractors may have led to higher response rates than might otherwise have been achieved. According to Horvitz, the two firms achieved roughly the same results. The quality of the data collected by each firm has yet to be evaluated. The results of such a comparison may show a slightly higher quality outcome for NORC because of their more extensive data collection experience and their somewhat different orientation. RTI is a relatively new survey organization, and RTI's orientation is more toward organization and management of field operations.

It is rare to be able to make comparisons of organizational effects, Elinson noted. The Stouffer study conducted by NORC and the Gallup Organization was the subject of an independent comparison. Sirken commented that the shared weekly status reports may have removed the independent element from any comparisons, since relative standings of both organizations were documented and shared every week. Elinson further noted that the object of the contract was to get the highest response rate possible, not to compare the two organizations.

Another issue that was raised is the impact of repeated panel interviews. Sirken explained that

several years ago NCHS included a supplement in the Current Population Survey, a quarterly survey divided into monthly waves of household interviews. Two subsamples were selected. In the first, respondents were interviewed twice during the quarter; in the second, respondents were interviewed only once, but at the same time that the second set of interviews was being conducted with the first subsample.

The particular question of interest in this survey concerned whether or not people had received Salk polio shots. It was noticed that reports of having had their first Salk shot were much higher in the group interviewed twice than in the group interviewed once. Having asked that question initially clearly had an impact on the behavior of the first group in seeking medical care. It seems that the methodology of having people keep diaries and focusing their attention on their health would perhaps have the effect of making people seek medical care. We may thus be influencing behaviors that we are trying to measure. This is a terribly important issue to be faced.

In regard to NMCES, Wilensky observed that there was concern during the survey that the respondents would act differently because they were participating and being observed. An attempt to evaluate the "Hawthorne effect" will be made by looking at quarterly segments of NMCES data and comparing it to Health Interview Survey (HIS) data. By comparing different quarters of data from both surveys, variations can be detected in NMCES reports that cannot be attributed to seasonality, as reflected in the HIS data. Also, by comparing first quarter NMCES data with last quarter NMCES data, observational or participatory effects should be identified.

Sirken suggested that a possible solution to the problem would be split panel designs so that comparisons can be made among panels. Sudman further suggested the usefulness of comparing the household reports with the provider reports to allow differentiating actual changes in behavior through changes in reporting. Verbrugge spoke of a need to model conditioning effects. Although there have been a number of studies on one or two effects, there is very little literature on the whole conditioning process plus other influences on respondent reporting. A great deal of NMCES data are available for analyses along these lines. Evaluations of different effects would be interesting. There would probably be situations where analyses can identify the net effect of two different—and perhaps conflicting—effects.

The issue of the quality of data collected was raised by Bergner. In many of the papers heard at this conference and in the methodological literature, there is the issue that "more is better" in terms of reporting. The feeling is that when more data are reported, the data are better in quality. She is unaware of any systematic studies that have looked at this issue in an experimental way. The concern is that when we get more data, it is not really clear that the quality of the data is better. In listening to interviews conducted by telephone, she has become very concerned about the quality of data being collected, particularly where probing is not done in a standardized way.

Fowler recalled a study that he and Cannell did many years ago on mail surveys. With each succeeding wave of mail returns, the data got worse and the respondents were less enthusiastic about reporting. In this case, the data from the last 10 percent of the respondents made the total population estimates worse than if they had excluded those responses. A paper by Neter and Waksberg (1964b) that addresses this issue was cited by Rothwell. A rotating panel was used with different recall periods. Because of various recall biases, including telescoping, a higher level of reporting was not always better.

Recommendations

Based on the open discussion, the following recommendations emerged from the NMCES session:

1. Additional research is needed on the value in cost-benefit terms of providing respondents in health panel studies a Summary of previous information reported so that cost and other information may be completed or modified.
2. The process of reviewing and updating the Summary needs research specific to understanding the stimuli for updates that modify the originally reported data and the implications on the quality of the data.
3. The specific demographic groups for which the use of a Summary is more appropriate need to be researched and identified.
4. The cost of obtaining health provider data for validating household reports should be related to the value of the data in reducing total survey error.
5. Alternative measures of medical charges and insurance coverage require additional comparison and evaluation of both theoretical foundations and reliability of response.
6. More post-survey researching of interviewer perceptions is needed to add to our understanding of respondent limitations, instrument limitations, and limitations resulting

from the interviewer-respondent interactive process.

7. It is very important to recognize inadequately researched methodological techniques that are being implemented for the first time in health surveys and to imbed controlled evaluations in the survey design.
8. Serious consideration should be given to budgeting for methodological research of large-scale health surveys proportional to the expected survey operational budgets.
9. Methodological research for large-scale health surveys should be concerned with identifying the significant sources of error and choosing those designs that are *optimum* in terms of total survey error.

A bibliography on telephone interviewing and related matters*

D. Garth Taylor, National Opinion Research Center, University of Chicago

Adler, M.K. "The Use of the Telephone in Industrial Market Research." *Scientific Business,* 1 (February 1964):336-42.

American Telephone and Telegraph. "Notes on Distance Dialing." 1975.

Anderson, C.L. and L.J. Halford. "A Four State Comparison of Variable Sampling and Data Collection Procedures." *Pacific Sociological Review,* 13 (Summer 1970):149-55.

Assael, H. "Comparison of Brand Share Data by Three Reporting Systems." *Journal of Marketing Research,* 4 (November 1967):400-1.

Backstrom, C.H. and G.D. Hursh. *Survey Research.* Evanston, Ill.: Northwestern University Press, 1963.

Ball, D.W. "Toward a Sociology of Telephones and Telephoners." Pp. 59-75 in M. Truzzi (ed.), *Sociology and Everyday Life.* Englewood Cliffs, N.J.: Prentice-Hall, 1968.

Bennett, C.T. "A Telephone Interview: A Method for Conducting a Follow-up Study." *Mental Hygiene,* 45 (April 1961):216-20.

Berger, P.K. and J.E. Sullivan. "Instructional Set, Interview Context, and the Incidence of 'Don't Know' Responses." *Journal of Applied Psychology,* 54 (October 1970):414-16.

Blankenship, A.B. (ed.). *How To Conduct Consumer and Opinion Research: The Sampling Survey in Operation.* New York: Harper & Bros., 1946.

———. "Listed versus Unlisted Numbers in Telephone-Survey Samples." *Journal of Advertising Research,* 17 (February 1977):39-42.

———. *Professional Telephone Surveys.* New York: McGraw-Hill, 1977.

Boruch, R.F. "Maintaining Confidentiality of Data in Educational Research: A Systematic Analysis." *American Psychologist,* 26 (May 1971):413-30.

Boyd, H.W. and R. Westfall. "Interviewer Bias Revisited." *Journal of Marketing Research,* 2 (February 1965):58-63.

Brunner, G.A. and S.J. Carroll, Jr. "The Effect of Prior Telephone Appointments on Completion Rates and Response Content." *Public Opinion Quarterly,* 31 (Winter 1967-68): 52-54.

Brunner, J.A. and G.A. Brunner. "Are Voluntarily Unlisted Telephone Subscribers Really Different?" *Journal of Marketing Research,* 8 (February 1971):121-24.

Bryant, B.E. "Respondent Selection in a Time of Changing Household Composition." *Journal of Marketing Research,* 12 (May 1975):129-35.

Bushery, J.M., C.D. Cowan, and L.R. Murphy. "Experiments in Telephone-Personal Visit Surveys." Pp. 564-69 in *Proceedings,* Survey Research Methods Section, American Statistical Association, 1978.

Buzzell, R.D. and D.F. Cox. *Marketing Research and Information Systems: Text and Cases.* New York: McGraw-Hill, 1969.

Cahalan, D. "Measuring Newspaper Readership by Telephone: Two Comparisons with Face-to-Face Interviews." *Journal of Advertising Research,* 1 (December 1960):1-6.

Carter, R.E., Jr. and V.C. Troldahl. "Use of a Recall Criterion in Measuring the Educational Television Audience." *Public Opinion Quarterly,* 26 (Spring 1962):114-21.

Colombotos, J. "The Effects of Personal vs. Telephone Interviews on Socially Acceptable Responses." *Public Opinion Quarterly,* 29 (Fall 1965):457-58.

———. "Personal versus Telephone Interviews:

*The NORC Telephone Survey Committee has prepared this bibliography to aid our thinking on telephone survey methods. There are several kinds of articles listed here. The topics most thoroughly covered are (a) issues in telephone sampling, (b) screening and respondent selection, (c) quality of telephone interview data, and (d) supporting articles that are frequently referenced by articles under the other three headings. There seem to be many fewer articles describing the hardware and software of computer-assisted telephone interviewing. Some of the books on telephone interviewing contain detailed recommendations for the actual operation of telephone interviewing. Revised August 1979.

Effect on Responses." *Public Health Reports,* 84 (September 1969):773–82.

Coombs, L. and R. Freedman. "Use of Telephone Interviews in a Longitudinal Fertility Study." *Public Opinion Quarterly,* 28 (Spring 1964):112–17.

Cooper, S.L. "Random Sampling by Telephone: A New and Improved Method." *Journal of Marketing Research,* 1 (November 1964):45–48.

Cunningham, J.M., H.H. Westerman, and J. Fischhoff. "A Follow-up Study of Patients Seen in a Psychiatric Clinic for Children." *American Journal of Orthopsychiatry,* 26 (July 1956):602–10.

Dalenius, T. "The Problem of Not-At-Homes." *Statistisk Tidskrift,* N.S. 4 (April 1955):208–11.

Daniel, W.W. "Nonresponse in Sociological Surveys: A Review of Some Methods for Handling the Problem." *Sociological Methods & Research,* 3 (February 1975):291–307.

Deming, W.E. *Sample Design in Business Research.* New York: Wiley, 1960.

Dillman, D.A. *Mail and Telephone Surveys: The Total Design Method.* New York: Wiley, 1978.

———, J.G. Gallegos, and J.H. Frey. "Reducing Refusal Rates for Telephone Interviews." *Public Opinion Quarterly,* 40 (Spring 1976): 66–78.

Dohrenwend, B.S. "An Experimental Study of Directive Interviewing." *Public Opinion Quarterly,* 34 (Spring 1970):117–25.

———, J. Colombotos, and B.P. Dohrenwend. "Social Distance and Interview Effects." *Public Opinion Quarterly,* 32 (Fall 1968):410–22.

——— and B.P. Dohrenwend. "Sources of Refusals in Surveys." *Public Opinion Quarterly,* 32 (Spring 1968):74–83.

Dunkelberg, W.C. and G.S. Day. "Nonresponse Bias and Callbacks in Sample Surveys." *Journal of Marketing Research,* 10 (May 1973): 160–68.

Eastlack, J.O., Jr. and H. Assael. "Better Telephone Surveys through Centralized Interviewing." *Journal of Advertising Research,* 6 (March 1966):2–7.

Edwards, A.L. *The Social Desirability Variable in Personality Assessment and Research.* New York: Dryden Press, 1957.

Enterline, P.E. and K.G. Capt. "A Validation of Information Provided by Household Respondents in Health Surveys." *American Journal of Public Health,* 49 (February 1959): 205–12.

Falthzik, A.M. "When To Make Telephone Interviews." *Journal of Marketing Research,* 9 (November 1972):451–52.

Ferber, R. *The Reliability of Consumer Reports of Financial Assets and Debts.* Studies in Consumer Savings, No. 6. Urbana: Bureau of Economic and Business Research, University of Illinois, 1966.

Field, D.R. "The Telephone Interview in Leisure Research." *Journal of Leisure Research,* 5 (Winter 1973):51–59.

Fletcher, J.E. and H.B. Thompson. "Telephone Directory Samples and Random Telephone Number Generation." *Journal of Broadcasting,* 18 (Spring 1974):187–91.

Frankel, M.R. *Inference from Survey Samples: An Empirical Investigation.* Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan, 1971.

——— and L.R. Frankel. "Some Recent Developments in Sample Survey Design." *Journal of Marketing Research,* 14 (August 1977): 280–93.

Freeman, J. and E.W. Butler. "Some Sources of Interviewer Variance in Surveys." *Public Opinion Quarterly,* 40 (Spring 1976):79–91.

Friedman, L. and H. Friedman. "Does the Perceived Race of a Telephone Interviewer Affect the Responses of White Subjects to an Attitudes toward Negroes Scale?" Pp. 556–58 in *Proceedings,* Survey Research Methods Section, American Statistical Association, 1978.

Frisbie, B. and S. Sudman. "The Use of Computers in Coding Free Responses." *Public Opinion Quarterly,* 32 (Summer 1968):216–32.

Fry, H.G. and S. McNair. "Data Gathering by Long Distance Telephone." *Public Health Reports,* 73 (September 1958):831–35.

Gates, R. and C. McDaniel. "Improving Completion Rates by More Efficient Scheduling of Telephone Interviews." *Viewpoints—The Journal for Data Collection,* 16 (May 1976):8–10.

Glasser, G.J. and G.D. Metzger. "Random-Digit Dialing as a Method of Telephone Sampling." *Journal of Marketing Research,* 9 (February 1972):59–64.

——— and ———. "National Estimates of Nonlisted Telephone Households and Their Characteristics." *Journal of Marketing Research,* 12 (August 1975):359–61.

Goldberg, D., H. Sharp, and R. Freedman. "The Stability and Reliability of Expected Family Size Data." *Milbank Memorial Fund Quarterly,* 37 (October 1959):369–85.

Goodman, L.A. "Snowball Sampling." *Annals of Mathematical Statistics,* 32 (March 1961): 148–70.

Goudy, W.J. and H.R. Potter. "Interview Rapport: Demise of a Concept." *Public Opinion Quarterly,* 39 (Winter 1975–76):529–43.

Groves, R.M. "An Experimental Comparison of National Telephone and Personal Interview Surveys." Pp. 232–41 in *Proceedings,* Social Statistics Section, American Statistical Association, 1977.

———. "Comparing Telephone and Personal Interview Surveys." *Economic Outlook USA,* 5 (Summer 1978):49–51.
———. "An Empirical Comparison of Two Telephone Sample Designs." *Journal of Marketing Research,* 15 (November 1978):622–31.
———. "On the Mode of Administering a Questionnaire and Responses to Open-Ended Items." *Social Science Research,* 7 (September 1978):257–71.
——— and R.L. Kahn. *Surveys by Telephone: A National Comparison with Personal Interviews.* New York: Academic Press, 1979.
Haberman, P.W. and J. Elinson. "Family Income Reported in Surveys: Husbands versus Wives." *Journal of Marketing Research,* 4 (May 1967):191–94.
Hanson, R.H. and E.S. Marks. "Influence of the Interviewer on the Accuracy of Survey Results." *Journal of the American Statistical Association,* 53 (September 1958):635–55.
Hauck, M. and M. Cox. "Locating a Sample by Random Digit Dialing." *Public Opinion Quarterly,* 38 (Summer 1974):253–60.
——— and J. Goldberg. "Telephone Interviewing on the NLRB Election Study." *Survey Research,* 5 (January 1973):15–16.
Henson, R., A. Roth, and C.F. Cannell. *Personal versus Telephone Interviews and the Effects of Telephone Reinterviewing on Reporting of Psychiatric Symptomatology.* Ann Arbor: Survey Research Center, University of Michigan, 1974.
Herman, J.B. "Mixed-Mode Data Collection: Telephone and Personal Interviewing." *Journal of Applied Psychology,* 62 (August 1977): 399–404.
Hildum, D.C. and R.W. Brown. "Verbal Reinforcement and Interviewer Bias." *Journal of Abnormal and Social Psychology,* 53 (July 1956):108–11.
Hill, R.J. and N.E. Hall. "A Note on Rapport and the Quality of Interview Data." *Southern Social Science Quarterly,* 44 (December 1963): 247–55.
Hochstim, J.R. "A Critical Comparison of Three Strategies of Collecting Data from Households." *Journal of the American Statistical Association,* 62 (September 1967):976–89.
Ibsen, C.A. and J.A. Ballweg. "Telephone Interviews in Social Research: Some Methodological Considerations." *Quality and Quantity,* 8 (June 1974):181–92.
Janofsky, A.I. "Affective Self-Disclosure in Telephone versus Face to Face Interviews." *Journal of Humanistic Psychology,* 22 (Spring 1971):93–103.
Jordan, L.A., A.C. Marcus, and L.G. Reeder. "Response Styles in Telephone and Household Interviewing: A Field Experiment from the Los Angeles Health Survey." Pp. 362–66 in *Proceedings,* Survey Research Methods Section, American Statistical Association, 1978.
Josephson, E. "Screening for Visual Impairment." *Public Health Reports,* 80 (January 1965):47–54.
Judd, R.C. "Telephone Usage and Survey Research." *Journal of Advertising Research,* 6 (December 1966):38–39.
Kegeles, S.S., C.F. Fink, and J.P. Kirscht. "Interviewing a National Sample by Long-Distance Telephone." *Public Opinion Quarterly,* 33 (Fall 1969):412–19.
Kildegaard, I.C. "Telephone Trends." *Journal of Advertising Research,* 6 (June 1966):56–60.
———. "Rejoinder." *Journal of Advertising Research,* 6 (December 1966):40–41.
Kish, L. "A Procedure for Objective Respondent Selection within the Household." *Journal of the American Statistical Association,* 44 (September 1949):380–87.
———. "Studies of Interviewer Variance for Attitudinal Variables." *Journal of the American Statistical Association,*57 (March 1962):92–115.
——— and I. Hess. "On Noncoverage of Sample Dwellings." *Journal of the American Statistical Association,* 53 (June 1958):509–24.
——— and ———. "A 'Replacement' Procedure for Reducing the Bias of Nonresponse." *American Statistician,* 13 (October 1959):17–19.
Klecka, W.R. and A.J. Tuchfarber. "Random Digit Dialing as an Efficient Method for Political Polling." *Georgia Political Science Association Journal,* 2 (Spring 1974):133–51.
——— and ———. "Random Digit Dialing: A Comparison to Personal Surveys." *Public Opinion Quarterly,* 42 (Spring 1978):105–14.
Koo, H.P., J.C. Ridley, P.V. Piserchia, D.A. Dawson, C.A. Bachrach, M.I. Holt, and D.G. Horvitz. "An Experiment on Improving Response Rates and Reducing Call Backs in Household Surveys." Pp. 491–94 in *Proceedings,* Social Statistics Section, American Statistical Association, 1976.
Koons, D.A. "Current Medicare Survey, Telephone Interviewing Compared with Personal Interviews." Response Research Staff Report #74–4. Statistical Research Division, U.S. Bureau of the Census, 1974.
Lahiri, D.B. "A Method of Sample Selection Providing Unbiased Ratio Estimates." *Bulletin of the International Statistical Institute,* 33 (Part II, 1951):133–40.
Landon, E.L., Jr. and S.K. Banks. "Relative Efficiency and Bias of Plus-one Telephone Sampling." *Journal of Marketing Research,* 14 (August 1977):294–99.
Lansing, J.B., G.P. Ginsberg, and K. Braaten. *An Investigation of Response Error.* Studies in

Consumer Savings, No. 2. Urbana: Bureau of Economic and Business Research, University of Illinois, 1961.

Larsen, O.N. "The Comparative Validity of Telephone and Face-to-Face Interviews in the Measurement of Message Diffusion from Leaflets." *American Sociological Review,* 17 (August 1952):471–76.

Leuthold, D.A. and R. Scheele. "Patterns of Bias in Samples Based on Telephone Directories." *Public Opinion Quarterly,* 35 (Summer 1971): 249–57.

Locander, W.B. and J.P. Burton. "The Effect of Question Form on Gathering Income Data by Telephone." *Journal of Marketing Research,* 13 (May 1976):189–92.

———, S. Sudman, and N. Bradburn. "An Investigation of Interview Method, Threat and Response Distortion." *Journal of the American Statistical Association,* 71 (June 1976):269–75.

Lucas, W. and W. Adams. *An Assessment of Telephone Survey Methods.* Santa Monica, Cal.: Rand Corporation, 1977.

Luck, D.J., H.G. Wales, and D.A. Taylor. *Marketing Research.* Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Mahalanobis, P.C. "On Large-Scale Sample Surveys." *Philosophical Transactions of the Royal Society of London, Series B,* 231 (1946):329–451.

Market Research Society. "Response Rates in Sample Surveys: Report of a Working Party of the Market Research Society's Research and Development Committee." *Journal of the Market Research Society,* 18 (July 1976):113–42.

Marquis, K.H., J. Marshall, and S. Oskamp. "Testimony Validity as a Function of Question Form, Atmosphere and Item Difficulty." *Journal of Applied Social Psychology,* 2 (April–June 1972):167–86.

Maynes, E.S. "The Anatomy of Response Errors: Consumer Saving." *Journal of Marketing Research,* 2 (November 1965):378–87.

Meltzer, J.W. and J.R. Hochstim. "Reliability and Validity of Survey Data on Physical Health." *Public Health Reports,* 85 (December 1970):1075–86.

Mooney, H.W., B.R. Pollack, and L. Corsa, Jr. "Use of Telephone Interviewing To Study Human Reproduction." *Public Health Reports,* 83 (December 1968):1049–60.

A National Probability Sample of Telephone Households Using Computerized Sampling Techniques. Radnor, Pa.: Chilton Research Services, n.d.

Nelson, R.O. *The McMillan System of Random Digit Dialing as Used by Chilton Research Services.* Radnor, Pa.: Chilton Research Services, 1977.

Nicholls, W.L., II. *Designing Telephone Surveys for the Greater Bay Area.* Technical Report #20. Berkeley: Survey Research Center, University of California, 1977.

———. "Experiences with CATI in a Large-Scale Survey." Pp. 9–17 in *Proceedings,* Survey Research Methods Section, American Statistical Association, 1978.

Northrop, R.M. and O.L. Deniston. "Comparison of Mail and Telephone Methods To Collect Program Evaluation Data." *Public Health Reports,* 82 (August 1967):739–45.

Oakes, R.H. "Differences in Responsiveness in Telephone versus Personal Interviews." *Journal of Marketing,* 19 (October 1954):169.

Paisley, W.J. and E.B. Parker. "A Computer-Generated Sampling Table for Selecting Respondents within Households." *Public Opinion Quarterly,* 29 (Fall 1965):431–36.

Payne, S.L. "Some Advantages of Telephone Surveys." *Journal of Marketing,* 20 (January 1956):278–81.

———. "Data Collection Methods: Telephone Surveys." Pp. 2-105–23 in R. Ferber (ed.), *Handbook of Marketing Research.* New York: McGraw-Hill, 1974.

Pearl, R.B. and D.B. Levine. "A New Methodology for a Consumer Expenditure Survey." Pp. 254–59 in *Proceedings,* Business and Economic Statistics Section, American Statistical Association, 1971.

Perry, J.B., Jr. "A Note on the Use of Telephone Directories as a Sample Source." *Public Opinion Quarterly,* 32 (Winter 1968–69): 691–95.

Reingen, P.H. and J.B. Kernan. "Compliance with an Interview Request: A Foot-in-the-Door, Self-Perception Interpretation." *Journal of Marketing Research,* 14 (August 1977): 365–69.

Rich, C.L. "Is Random Digit Dialing Really Necessary?" *Journal of Marketing Research,* 14 (August 1977):300–5.

Rogers, C., M. Rogers, L. Seward, and G. Shure. *CATI User Documentation.* Los Angeles: Center for Computer-Based Behavioral Studies, University of California at Los Angeles, 1979.

Rogers, T.F. "Interviews by Telephone and in Person: Quality of Responses and Field Performance." *Public Opinion Quarterly,* 40 (Spring 1976):51–65.

Roslow, S. and L. Roslow. "Unlisted Phone Subscribers Are Different." *Journal of Advertising Research,* 12 (August 1972):35–38.

Rustemeyer, A. "Toward Development of a Computer-Assisted Telephone Interviewing System." Revised. U.S. Bureau of the Census, 1977.

——— and A. Levin. "Report on a Telephone Survey Using Computer Assistance." U.S. Bureau of the Census, 1977.

———, G.H. Shure, M.S. Rogers, and R.J. Meeker. "Computer-Assisted Telephone Interviewing: Design Considerations." Pp. 1–8 in *Proceedings,* Survey Research Methods Section, American Statistical Association, 1978.

San Augustine, A.J. and H.H. Friedman. "The Use of the Telephone Interview in Obtaining Information of a Sensitive Nature: A Comparative Study." Pp. 559–61 in *Proceedings,* Survey Research Methods Section, American Statistical Association, 1978.

Schmeideskamp, J.W. "Reinterviews by Telephone." *Journal of Marketing,* 26 (January 1962):28–34.

Shure, G.H. and R.J. Meeker. "A Minicomputer System for Multiperson Computer-Assisted Telephone Interviewing." *Behavior Research Methods & Instrumentation,* 10 (April 1978): 196–202.

Siemiatycki, J. "A Comparison of Mail, Telephone, and Home Interview Strategies for Household Health Surveys." *American Journal of Public Health,* 69 (March 1979):238–45.

Skogan, W. *The Center for Urban Affairs Random Digit Dialing Telephone Survey.* Report #M-31F (revised). Evanston, Ill.: Center for Urban Affairs, Northwestern University, n.d.

Spaeth, M. "Interviewing in Telephone Surveys." *Survey Research,* 5 (January 1973):9–13.

———. "Selected Bibliography on Telephone Interviewing." *Survey Research,* 5 (May 1973):13–14.

Stafford, J.E. "Influence of Preliminary Contact on Mail Returns." *Journal of Marketing Research,* 3 (November 1966):410–11.

Stock, J.S. "On Methods: How To Improve Samples Based on Telephone Listings." *Journal of Advertising Research,* 2 (September 1962):50–51.

Sudman, S. "New Uses of Telephone Methods in Survey Research." *Journal of Marketing Research,* 3 (May 1966):163–67.

———. "On Sampling of Very Rare Human Populations." *Journal of the American Statistical Association,* 67 (June 1972):335–39.

———. "The Uses of Telephone Directories for Survey Sampling." *Journal of Marketing Research,* 10 (May 1973):204–7.

——— and R. Ferber. "A Comparison of Alternative Procedures for Collecting Consumer Expenditure Data for Frequently Purchased Products." *Journal of Marketing Research,* 11 (May 1974):128–35.

——— and L.B. Lannom. *A Comparison of Alternative Panel Procedures for Obtaining Health Data.* Revised. Urbana: Survey Research Laboratory, University of Illinois, 1979.

Thornberry, O.T. and J.T. Massey. "Correcting for Undercoverage Bias in Random Digit Dialed National Health Surveys." Pp. 224–29 in *Proceedings,* Survey Research Methods Section, American Statistical Association, 1978.

Tigert, D.J., J.G. Barnes, and J.C. Bourgeois. "Research on Research: Mail Panel versus Telephone Survey in Retail Image Analysis." *The Canadian Marketer,* (Winter 1975):22–27.

Troldahl, V.C. and R.E. Carter, Jr. "Random Selection of Respondents within Households in Telephone Surveys." *Journal of Marketing Research,* 1 (May 1964):71–76.

Tuchfarber, A.J. and W.R. Klecka. *Random Digit Dialing: Lowering the Cost of Victimization Surveys.* Washington, D.C.: The Police Foundation, 1976.

Tull, D.S. and G.S. Albaum. *Survey Research: A Decisional Approach.* New York: Intext Educational Publishers, 1973.

——— and ———. "Bias in Random Digit Dialed Surveys." *Public Opinion Quarterly,* 41 (Fall 1977):389–95.

Turner, A.G. "An Experiment To Compare Three Interviewing Procedures in the National Crime Survey." Statistical Research Division, U.S. Bureau of the Census, 1977.

Uhl, K.P. and B. Schoner. *Marketing Research: Information Systems and Decision Making.* New York: Wiley, 1969.

U.S. Bureau of the Census. *Characteristics of Households with Telephones, March 1958.* Current Population Reports, Series, P-20, No. 95. Washington, D.C.: U.S. Government Printing Office, 1959.

———. *Characteristics of Households with Telephones, March 1960.* Current Population Reports, Series P-20, No. 111. Washington, D.C.: U.S. Government Printing Office, 1961.

———. *Characteristics of Households with Telephones, March 1965.* Current Population Reports, Series P-20, No. 146. Washington, D.C.: U.S. Government Printing Office, 1965.

———. *Response Errors in Collection of Expenditures Data by Household Interviews: An Experimental Study,* by J. Neter and J. Waksberg. Technical Paper No. 11. Washington, D.C.: U.S. Government Printing Office, 1965.

———. *Census of Housing: 1970. General Housing Characteristics, Final Report HC(1)-A.* Washington, D.C.: U.S. Government Printing Office, 1971, Tables 8 and 18.

———. *Who's Home When,* by D. Weber. Working Paper 37. Washington, D.C.: U.S. Government Printing Office, 1973.

U.S. National Center for Health Services Research. *Advances in Health Survey Research Methods: Proceedings of a National Invitational Conference, 1975.* DHEW Pub. No. (HRA) 77-3154. Rockville, Md.: NCHSR, 1977.

———. *Experiments in Interviewing Techniques: Field Experiments in Health Reporting, 1971–1977,* by C.F. Cannell, L. Oksenberg, and J.M. Converse (eds.). DHEW Pub. No. (HRA) 78-3204. Hyattsville, Md.: NCHSR, 1977.

U.S. National Center for Health Statistics. *Comparison of Hospitalization Reporting in Three Survey Procedures,* by C.F. Cannell and F. Fowler. Vital and Health Statistics, Series 2, No. 8. Washington, D.C.: U.S. Government Printing Office, 1965.

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

Weaver, C.N., S.L. Holmes, and N.D. Glenn. "Some Characteristics of Inaccessible Respondents in a Telephone Survey." *Journal of Applied Psychology,* 60 (April 1975):260–62.

Weiss, C.H. "Interaction in the Research Interview: The Effects of Rapport on Response." Pp. 17–20 in *Proceedings,* Social Statistics Section, American Statistical Association, 1970.

Weller, T. "Telephone Interviewing Procedures." *Survey Research,* 5 (January 1973): 13–14.

Wheatley, J.J. "Self-Administered Written Questionnaires or Telephone Interviews?" *Journal of Marketing Research,* 10 (February 1973):94–96.

Williams, J.A., Jr. "Interviewer Role Performance: A Further Note on Bias in the Information Interview." *Public Opinion Quarterly,* 32 (Summer 1968):287–94.

Wiseman, F. "Methodological Bias in Public Opinion Surveys." *Public Opinion Quarterly,* 36 (Spring 1972):105–8.

Woltman, H. and J. Bushery. "Results of the NCS Maximum Personal Visit-Maximum Telephone Interview Experiment." Statistical Methods Division, U.S. Bureau of the Census, 1977.

Glossary

Acquiescence—A tendency of the respondent to base his/her reply on some stimulus other than the question content. It may be stimulated by the desire to please the interviewer or the agency collecting the data or by some other cue such as an unbalanced question.

Agree-disagree—Form of question in which the respondent responds by stating his or her concurrence or nonconcurrence with a statement.

AID (Automatic Interaction Detector Program)—A computer search program that divides a sample into more homogeneous subgroups by maximizing the between sum-of-squares for the groups.

Alpha (same as Cronbach's α)—A measure of scale reliability based on comparisons of part-whole variances.

Balanced Format—Form of a question in which both alternatives or all choices are stated in the question. An unbalanced format includes only one alternative or choice.

Bias (or Net Systematic Error) of a Survey Estimate—The difference between the expected value (taken over the sampling design and the distribution of measurement errors) of the estimator and the "true" value of the parameter being estimated. This is particularly acute in surveys concerning sensitive or confidential matters and in ones where it might be expected that the estimates are consistently below or above the true population parameter. A consistent pattern of under- or overreporting will result in bias.

Bounded Recall—An interview where the respondent is reminded of what he/she reported in an earlier interview and is then asked only to report on any new events that occurred subsequent to the bounding interview.

Callback—In personal interviewing, repeat visits or calls to a sampling unit because the unit was not found at home or available for interviewing at the first or earlier calls. In mail surveys, repeat mailings to units that do not return questionnaires.

Clustering Effect—The increase in sampling variance caused by selecting several sample units from the same small geographic area. Since units in close geographic proximity are similar on some dimensions, the effective number of independent observations is reduced.

Coding Procedures—Techniques for providing unique numerical designations to data such that quantitative analysis of the data can be performed. These techniques may be used for assigning labels to survey respondents that, while allowing identification of data as coming from a single source, protect the identity of the person who is that source. The method can also be used to conceal the true value of data, especially that stored in computers, so that interpretation of the coded data is impossible and meaningless until the data are decoded.

Cohort Analysis—Statistical procedures for distinguishing between effects caused by the age(s) of a population and those caused by their living in a given historical period.

Completion Rate—See Response Rate.

Computer-Assisted Telephone Interviewing (CATI)—A telephone interviewing procedure in which cathode ray terminals (CRTs) are used to administer a questionnaire and record results.

Counting Rules—Procedures used in network sampling to determine an individual's probability of being mentioned.

Criterion Validity—A procedure for checking the accuracy of individual survey responses by using an independent source of information regarding a population being surveyed.

Cue—Some characteristics of the interview or the interviewer, the question wording, or the

interviewer's behavior, including feedback, that influence the direction of answers to one or more questions.

Diary—A written record kept concurrently by an individual respondent or household about events that would usually otherwise be difficult to remember.

Dichotomous—A random variable is said to be dichotomous if it assumes only one of two responses or values.

Error Model—A mathematical relationship that postulates the manner in which both sampling and nonsampling errors arise in the conduct and analysis of a sample survey.

Eta—An indicator of the ability of a predictor, using the categories given, to explain variation in the dependent variable (analogous to beta but based on raw rather than adjusted means).

Follow-up—A procedure whereby those members of a selected sample for whom a response is not obtained by one data collection strategy (e.g., telephone or mail) are contacted by the same or another data collection strategy in order to increase the response rate. It can also be used to designate repeated surveys among a panel of respondents.

Imputation—A procedure used to assign a value to a missing or obviously incorrect answer, based on the pattern of answers to related questions.

Interviewer Feedback—Some verbal or nonverbal communication by the interviewer in response to respondent behavior.

MCA (Multiple Classification Analysis)—A computer procedure for multiple regression using categorical independent variables.

Mean Square Error (MSE)—In a survey estimate, this is the expected value of the squared difference between the estimator and the population parameter being estimated, where the expectation is taken over the sampling design and the distribution of measurement errors.

Memory Lapse—The universally observed phenomenon that the longer ago the event occurred in the past, the more likely the respondent is to have difficulty recalling the event. This rule may not hold true where the event is associated with some dramatic period of time in the life of the respondent.

Monotonic—Refers to data that always move in the same direction or are constant with reference to time or another variable. The data never move in the opposite direction.

Multiplicity Estimator—An unbiased network estimator that weights the sample elements by the inverses of the number of enumeration units at which they are eligible to be enumerated. The information needed to determine the weight is collected in the survey from the enumeration units that report the elements.

Network Estimators—See Multiplicity Estimator.

Network Sampling—A sampling method that obtains information from a respondent about other persons inside and outside a household using well-defined kinship and friendship rules.

Nonresponse Rate—The complement of response rate. The numerator is those eligible respondents selected in a sample for whom information is not obtained because of refusals, not found at home, unavailable by reason of illness, incompetence, language difficulty, etc. The denominator is the total number of eligible respondents initially selected for the sample.

Open-ended Questions—Questions for which the response categories are unspecified as opposed to closed questions where the categories are specified by the researcher.

Overreporting—Survey responses that produce a higher estimate of the incidence of some event or characteristic than is accurate.

Panel—A study design involving reinterview or a series of diaries or questionnaires with the same sample or respondents (or household units) at two or more different times. Normally used to study changes over time, giving rise to longitudinal data.

Principal Components Analysis—A factor analysis procedure that maximizes variance of composite scores.

Proxy Respondents—Respondents who provide information about other persons, generally within the same household, in addition to or instead of providing information about themselves.

Random Digit Dialing—A procedure for obtaining a probability sample of households with telephones. Numbers are selected at random from exchanges without prior knowledge of whether they are working numbers, business numbers, or residential household numbers. The strength of the procedure is the inclusion of those households with unlisted numbers. Caution must be taken to ensure that the digits used, whether terminal or otherwise, are uniformly distributed.

Rapport—A broadly defined term used to refer to the quality of the relationship of interaction

between the interviewer and respondent. Usually this refers to characteristics of warmth and friendliness and open communication in interpersonal relationships.

Recall Period—The time period over which a respondent is required to remember what events have occurred.

Record Checks—The comparison of information provided by a respondent in a survey with information obtained from other sources, especially governmental or institutional records including census, Social Security, vital records, dispensaries, hospitals, mental health agencies, pharmacies, and municipal activities such as police and fire department functions.

Reliability—Correspondence, repeatability, or consistency between identical survey questions at two different times.

Respondent Anonymity—Situation in which the survey information is gathered in such a manner that the respondent's identity cannot in any way be linked to the information provided.

Respondent Burden—The level of demand placed on the respondent necessary to answer the questions in the survey instrument. This includes the total time demands on the respondent, the demands on his memory, difficulty in understanding the question, and possible embarrassment.

Response Rate (Completion Rate)—The percentage of an eligible sample for whom information is obtained. For an interview survey the numerator of the formula is the number of completed interviews. The denominator is the total sample size minus ineligible sample units, that is, minus those not meeting the criteria for a potential respondent as defined for that particular study.

Robust—A description of statistical estimates or conclusions that do not change as a result of small changes in assumptions or measurements.

Sampling Variance—The contribution to the total variance arising from the random selection of a sample, rather than a complete enumeration, of the population.

Self-weighting Design—A sample in which all units have the same total probability of selection, so that no additional weighting is required.

Social Desirability Bias—Answers that reflect an attempt to enhance some socially desirable characteristics or minimize the presence of some socially undesirable characteristics. The source of the expectations or values influencing answers can be the respondents themselves, the interviewers, or society as a whole; these may give rise to an acquiescent response.

Standardized Measures—Tested and validated measures of major variables, such as those dealing with illness and demographic characteristics. The use of standard measures and measuring techniques provides a basis for comparability of information from investigator to investigator.

Stratification—A design technique employed in sample surveys whereby the finite population is classified into several parts (or strata) and a random sample is independently selected from each stratum. The purpose of stratification is to reduce the sample variance.

Telescoping—A reporting error in which the time an event occurred is remembered as having been more recent than it actually was. Events may also be placed backward in time.

Total Survey Design (TSD)—A concept that implies an efficient allocation of survey resources among the different error components in order to minimize the total error of estimates.

Total Survey Error (TSE)—The aggregate of all components of error occurring in the conduct or analysis of a sample survey. Included in the total survey error are all sampling and nonsampling errors.

Underreporting—Survey responses that produce a lower estimate of the incidence of some event or characteristic than is accurate.

Validity—A valid measure is one that measures what it claims to and not something else. Validity is a continuous concept, so most measures fall between total validity and total nonvalidity. A totally valid measure is one without bias.

Weighting—A procedure used to obtain unbiased population estimates when sample units have unequal probabilities of selection, either by design or because of differential cooperation; the weights are the reciprocals of the unequal selection probabilities.

References

Acheson, R. (ed.)
1965 Comparability in International Epidemiology. Milbank Memorial Fund Quarterly 43 (April, Part 2).

Aday L. and R. Andersen
1978 "Standard measures of standard variables." Pp. 63–66 in U.S. National Center for Health Services Research, Health Survey Research Methods: Second Biennial Conference, 1977. DHEW Pub. No. (PHS) 79–3207. Hyattsville, Md.: NCHSR.

Afifi, A.A. and R.M. Elashoff
1966 "Missing observations in multivariate statistics: I. Review of the literature." Journal of the American Statistical Association 61 (September):595–604.

Allen, G.I., L. Breslow, A Weissman, and H. Nisselson
1954 "Interviewing versus diary keeping in eliciting information in a morbidity survey." American Journal of Public Health 44 (July):919–27.

Alwin, D.F.
1977 "Making errors in surveys: an overview." Sociological Methods & Research, 6 (November):131–50.

American Medical Association
1978 Critique and Comment on the Council on Wage and Price Stability's Staff Report: A Study of Physician Fees. Chicago: AMA.
1979 Profile of Medical Practice. Chicago: AMA. (Annual since 1971.)

American Statistical Association
1974 "Report on the ASA Conference on Surveys of Human Populations." American Statistician 28 (February):30–34.

Andersen, R.
1975 "The effect of measurement error in differences in the use of health services." Pp. 229–55 in R. Andersen, J. Kravits, and O.W. Anderson, Equity in Health Services. Cambridge, Mass.: Ballinger.

Andersen, R., J. Kasper, M.R. Frankel, and Associates
1979 Total Survey Error: Applications to Improve Health Surveys. San Francisco: Jossey-Bass.

Anderson, J. et al.
1978 "Performance versus capacity: a conflict in classifying function for health status measurement." Unpublished draft, Division of Health Policy, Department of Community Medicine, University of California, San Diego.

Andrew, F.M., J.N. Morgan, and J.A. Sonquist
1969 Multiple Classification Analysis: A Report on a Computer Program for Multiple Regression Using Categorical Predictors. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.

Astrachan, B.M., M. Harrow, D. Adler, L. Brauer, A. Schwartz, C. Schwartz, and G. Tucker
1972 "A checklist for the diagnosis of schizophrenia." British Journal of Psychiatry 121 (November):529–39.

Axelrod, M.
1956 "Urban structure and social participation." American Sociological Review 21 (February):13–18.

Bailar, B.A.
1975 "The effects of rotation group bias on estimates from panel surveys." Journal of the American Statistical Association 70 (March):23–30.

Bailar, B.A. and C.M. Lanphier
1978 Development of Survey Methods To Assess Survey Practices. Washington, D.C.: American Statistical Association.

Barnes, J.A.
1969 "Networks and political process." Pp.

51–76 in J.C. Mitchell (ed.), Social Networks in Urban Situations. Manchester: Manchester University Press.

Berg, R.L. (ed.)
1973 Health Status Indexes. Chicago: Hospital Research and Educational Trust.

Bergner, M., R.A. Babbitt, S. Kressel, W.E. Pollard, B.S. Gilson, and J.R. Morris
1976 "The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure." International Journal of Health Services 6 (3):393–416.

Berkanovic, E., A.C. Marcus, and L.A. Jordan
1978 "The health belief model, health orientations, and the decision to seek medical care." Manuscript in preparation, School of Public Health, University of California, Los Angeles.

Berkman, L.F. and S.L. Syme
1979 "Social networks, host resistance, and mortality: a nine-year follow-up study of Alameda County residents." American Journal of Epidemiology 109 (February):186–204.

Berle, B.B., R.H. Pinsky, S. Wolf, and H.G. Wolff
1952 "A clinical guide to prognosis in stress diseases." Journal of the American Medical Association 149 (August 30):1624–28.

Bice, T.
1976 "Measurement of attitudes." In M. Pflanz and E. Schach (eds.), Cross-national Sociomedical Research: Concepts, Methods, Practice. Stuttgart: George Thieme.

Bohrnstedt, G.W. and T.M. Carter
1971 "Robustness in regression analysis." Pp. 118–46 in H.L. Costner (ed.), Sociological Methodology 1971. San Francisco: Jossey-Bass.

Boissevain, J.
1974 Friends of Friends: Networks, Manipulators and Coalitions. Oxford: Basic Blackwell.

Boswell, D.M.
1969 "Personal crises and the mobilization of the social network." Pp. 245–96 in J.C. Mitchell (ed.), Social Networks in Urban Situations. Manchester: Manchester University Press.

Bradburn, N.M.
1978 "Respondent burden." Pp. 45–53 in U.S. National Center for Health Services Research, Health Survey Research Methods: Second Biennial Conference, 1977. DHEW Pub. No. (PHS) 79–3207. Hyattsville, Md.: NCHSR.

Bradburn, N.M. and D. Caplovitz
1965 Reports on Happiness: A Pilot Study of Behavior Related to Mental Health. Chicago: Aldine.

Brislin, R.W., W.J. Lonner, and R.M. Thorndike
1973 Cross-Cultural Research Methods. New York: Wiley.

Brodman, K., A.J. Erdmann, Jr., and H.G. Wolff
1956 Cornell Medical Index Health Questionnaire (MANUAL). New York: Cornell University Medical College, 1956.

Bushery, J.M., C.D. Cowan, and L.R. Murphy
1978 "Experiments in telephone-personal visit surveys." Pp. 564–69 in Proceedings, Survey Methods Research Section, American Statistical Association.

Campbell, D.T. and D.W. Fiske
1959 "Convergent and discriminant validation by the multitrait-multimethod matrix." Psychological Bulletin 56 (2):81–105.

Campbell, D.T. and J.C. Stanley
1966 Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally.

Cannell, C.F. and L. Monteiro
1978 "Discussion of response rates." Pp. 13–17 in U.S. National Center for Health Services Research, Health Survey Research Methods: Second Biennial Conference, 1977. DHEW Pub. No. (PHS) 79–3207. Hyattsville, Md.: NCHSR.

Cannell, C.F., L. Oksenberg, and J.M. Converse
1977 "Striving for response accuracy: experiments in new interviewing techniques." Journal of Marketing Research 14 (August):306–15.

Cassel, J.
1974 "Psychosocial processes and 'stress': theoretical formulation." International Journal of Health Services 4 (Summer):471–82.

Cobb, S.
1976 "Social support as a moderator of life stress." Journal of Psychosomatic Medicine 38 (5):300–14.

Cohen, J.
1960 "A coefficient of agreement for nominal scales." Educational and Psychological Measurement 20 (Spring): 37–46.

Colombotos, J.
1969 "Personal versus telephone interviews: effect on responses." Public Health Reports 84 (September): 773–82.

Comstock, G.W. and K.P. Partridge
1972 "Church attendance and health." Journal of Chronic Disease 25: 665–72.

Coombs, L. and R. Freedman
1964 "Use of telephone interviews in a longitudinal fertility study." Public Opinion Quarterly 28 (Spring): 112–17.

Cooper, S.L.
1964 "Random sampling by telephone—an improved method." Journal of Marketing Research 1 (November):46–48.

Costner, H.L.
1969 "Theory, deduction and rules of correspondence." American Journal of Sociology 75 (September):245–63.

Cotterill, P.
1978 "Technology, prices and competition: an analysis of the market for physician services." Working paper, American Medical Association, Chicago.

Crandell, D.L., D.C. Cook, and B.P. Dohrenwend
1971 "How psychiatric patients vs. community respondents explain the cause of their symptoms." Unpublished paper, Columbia University, 1971.

Daniel, W.W.
1975 "Nonresponse in sociological surveys: a review of some methods for handling the problem." Sociological Methods & Research 3 (February): 291–307.

Dean, A. and N. Lin
1977 "The stress-buffering role of social support," Journal of Nervous and Mental Disease 165 (December):403–17.

De Miguel, J.
1974 "A framework for the study of national health systems." Paper presented at 8th World Congress of Sociology, Toronto.

Dillman, D.A.
1978 Mail and Telephone Surveys: The Total Design Method. New York: Wiley.

Dillman, D.A. and J.H. Frey
1974 "Coming of age: interviews by telephone." Paper presented at the meetings of the Pacific Sociological Association, San Jose, California.

Dohrenwend, B.P.
1973 "Some issues in the definition and measurement of psychiatric disorders in general populations." Pp. 480–89 in Proceedings, Public Health Conference on Records and Statistics, 1972. DHEW Pub. No. 74–1214. Washington, D.C.: U.S. Government Printing Office.
1974 "Problems in defining and sampling the relevant population of stressful life events." Pp. 275–310 in B.S. Dohrenwend and B.P. Dohrenwend (eds.), Stressful Life Events: Their Nature and Effects. New York: Wiley.

Dohrenwend, B.P., E.T. Chin-Shong, G. Egri, F.S. Mendelsohn, and J. Stokes
1970 "Measures of psychiatric disorder in contrasting class and ethnic groups: a preliminary report of on-going research." Pp. 159–202 in E.H. Hare and J.K. Wing (eds.), Psychiatric Epidemiology: Proceedings of an International Symposium. London: Oxford University Press.

Dohrenwend, B.P. and D.L. Crandell
1970 "Psychiatric symptoms in community, clinic, and mental hospital groups." American Journal of Psychiatry 126 (May): 1611–21.

Dohrenwend, B.P. and B.S. Dohrenwend
1969 Social Status and Psychological Disorder: A Causal Inquiry. New York: Wiley.

Dohrenwend, B.S. and B.P. Dohrenwend (eds.)
1974 Stressful Life Events: Their Nature and Effects. New York: Wiley.

Dohrenwend, B.S. and B.P. Dohrenwend
1978 "Some issues in research on stressful life events." Journal of Nervous and Mental Disease 166 (January):7–15.

Dohrenwend, B.S., B.P. Dohrenwend, and D. Cook
1973 "Ability and disability in role functioning in psychiatric patient and non-patient groups." Pp. 337–60 in J.K. Wing and H. Häfner (eds.), Roots of Evaluation: The Epidemiological Basis for Planning Psychiatric Services. London: Oxford University Press.

Dohrenwend, B.S., L. Krasnoff, A.R. Askenasy, and B.P. Dohrenwend
1978 "Exemplification of a method for scaling life events: the Peri life events scale." Journal of Health and Social Behavior 19 (June):205–29.

Dunning, B. and D. Cahalan
1973-74 "By-mail vs. field self-administered questionnaires: an armed forces survey." Public Opinion Quarterly 37 (Winter):618-24.
Durbin, J. and A. Stuart
1951 "Differences in response rates of experienced and inexperienced interviewers." Journal of the Royal Statistical Society, Series A, 114 (Part 2): 163-95.
Einhorn, H.
Forthcoming "Learning from experience and suboptimal rules in decision making." In T. Wallsten (ed.), Cognitive Processes in Choice and Decision Behavior. Hillsdale, N.J.: Lawrence Erlbaum Associates.
Elashoff, R.M. and J.D. Elashoff
1970 "Regression analysis with missing data." Pp. 198-207 in R.L. Bisco (ed.), Data Bases, Computers, and the Social Sciences. New York: Wiley.
Elinson, J. (ed.)
1976 "Sociomedical health indicators." International Journal of Health Services 6 (3):377-538.
Elling, R.H.
1974 "Case studies of contrasting approaches to organizing for health: an introduction to a framework." Social Science and Medicine 8 (May): 263-70.
Endicott, J. and R.L. Spitzer
1972 "What! Another rating scale? The Psychiatric Evaluation Form." Journal of Nervous and Mental Disease 154 (February):88-104.
1978 "A diagnostic interview: the schedule for affective disorders and schizophrenia." Archives of General Psychiatry 35 (July): 837-44.
Fabrega, H., Jr.
1975 "The need for an ethnomedical science." Science 189 (September 19):969-75.
Fleiss, J.L., R.L. Spitzer, J. Endicott, and J. Cohen
1972 "Quantification of agreement in multiple psychiatric diagnoses." Archives of General Psychiatry 26:168-71.
Foulds, G.A.
1976 The Hierarchical Nature of Personal Illness. New York: Academic Press.
Frank, J.D.
1973 Persuasion and Healing: A Comparative Study of Psychotherapy. Revised ed. Baltimore: Johns Hopkins University Press.
Frankel, M.R.
1971 Inference from Survey Samples: An Empirical Investigation. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.
Freshnock, L.J. and L.J. Goodman
1979a "Medical group practice in the U.S.: patterns of survival between 1969 and 1975." Journal of Health and Social Behavior 20 (December):352-62.
1979b "The organization of physician services in solo and group practice." Working Paper 78-KC-502. Chicago: American Medical Association.
Fuller, C.
1974 "Effect of anonymity on return rate and response bias in a mail survey." Journal of Applied Psychology 59 (June):292-96.
Gilson, B.S., D. Erickson, C.T. Chavey, R.A. Babbitt, M. Bergner, and W.B. Carter
Forthcoming "A Chicano version of the Sickness Impact Profile." Culture, Medicine and Psychiatry.
Glasser, J. and R. Forthofer
1972 "Analysis of health service data." Unpublished paper, Department of Biometry, University of Texas, Houston.
Goodman, L.J.
1975 Differences in New York City Hospital Emergency Room Services. Ph.D. dissertation, New York University.
Goodman, L.J., E.H. Bennett III, and R.J. Odem
1977 "Current status of group medical practice in the United States." Public Health Reports 92 (September-October):430-43.
Goodman, L.J. and B.S. Eisenberg
1977 "The quality of physician data." Review of Public Data Use 5 (May): 37-44.
Gore, S.
1978 "The effect of social support in moderating the health consequences of unemployment." Journal of Health and Social Behavior 19 (June):157-65.
Goudy, W.J.
1976 "Nonresponse effects on relationships between variables." Public Opinion Quarterly 40 (Fall):360-9.
1977 Nonresponse Effects: Studies of the Failure of Potential Respondents to Reply to Survey Instruments. Exchange Bibliography No. 1236. Monticello, Ill.: Council of Planning Librarians.

Graham, T.W., B.H. Kaplan, J.C. Cornoni-Huntley, S.A. James, C. Becker, C.G. Hames, and S. Heyden
1978 "Frequency of church attendance and blood pressure evaluation." Journal of Behavioral Medicine 1 (March): 37-43.

Granovetter, M.S.
1973 "The strength of weak ties." American Journal of Sociology 78 (May): 1360-80.

Great Britain. Resource Allocation Working Party
1976 Sharing Resources for Health in England. London: Her Majesty's Stationery Office.

Great Britain. Social Survey Division
1971 Handicapped and Impaired in Great Britain, by A.I. Harris. London: Her Majesty's Stationery Office.

Groves, R.M.
1977 "An experimental comparison of national telephone and personal interview surveys." Pp. 232-41 in Proceedings, Social Statistics Section, American Statistical Association.

Groves, R.M. and R.L. Kahn
1979 Surveys by Telephone: A National Comparison with Personal Interviews. New York: Academic Press.

Gullen, W.H. and G.E. Garrison
1973 "Factors influencing physicians' response to mailed questionnaires." Health Services Reports 88 (June-July):510-14.

Gurin, G., J. Veroff, and S. Feld
1960 Americans View Their Mental Health: A Nationwide Interview Survey. New York: Basic Books.

Hauck, M. and M. Cox
1974 "Locating a sample by random digit dialing." Public Opinion Quarterly 38 (Summer):253-60.

Hawkins, D.F.
1977 "Nonresponse in Detroit Area Study surveys: a ten-year analysis." Working Papers in Methodology No. 8. Chapel Hill: Institute for Research in Social Science, University of North Carolina.

Health Services Research and Development Center, Johns Hopkins University
1977 Medical Economics Survey-Methods Study: Final Report. Submitted to Division of Health Interview Statistics, National Center for Health Statistics. Baltimore: HSRDC, Johns Hopkins University.

Henderson, S.R.
1978 "The Periodic Surveys of Physicians' design and methodology." Pp. 177-88 in American Medical Association, Profile of Medical Practice. Chicago: AMA.

Heneman, H.G., Jr. and D.G. Paterson
1949 "Refusal rates and interviewer quality." International Journal of Opinion and Attitude Research 3 (Fall): 392-98.

Henson, R., A. Roth, and C. Cannell
1974 Personal versus Telephone Interviews and the Effects of Telephone Reinterviews on Reporting of Psychiatric Symptomatology. Research report. Ann Arbor: Survey Research Center, University of Michigan.
1977 "Personal versus telephone interviews: the effects of telephone reinterviews on reporting of psychiatric symptomatology." Pp. 205-19 in U.S. National Center for Health Services Research, Experiments in Interviewing Techniques: Field Experiments in Health Reporting, 1971-1977. DHEW Pub. No. (HRA) 78-3204. Hyattsville, Md.: NCHSR. (PB 276 080, available NTIS only.)

Herman, J.B.
1977 "Mixed-mode data collection: telephone and personal interviewing." Journal of Applied Psychology 62 (August):399-404.

Hertel, B.R.
1976 "Minimizing error variance introduced by missing data routines in survey analysis." Sociological Methods & Research 4 (May):459-74.

Hochstim, J.R.
1967 "A critical comparison of three strategies of collecting data from households." Journal of the American Statistical Association 62 (September): 976-89.

Hoerner, J., Jr. and A. Haas
1979 "Survey method and attitude crystallization: a comparison of mail and phone surveys." Unpublished manuscript, Indiana University.

Holland, W.W., J. Ipsen, and J. Kostrzewski (eds.)
1979 Measurement of Levels of Health. Copenhagen: World Health Organization.

Holmes, T.H., J.R. Joffe, J.W. Ketcham, and T.F. Sheehy
1961 "Experimental study of prognosis." Journal of Psychosomatic Research 5 (October):235-52.

Holmes, T.H. and R.H. Rahe
1967 "The social readjustment rating scale." Journal of Psychosomatic Research 11 (August):213–18.

Horvitz, D. and J. Lessler
1978 "Disscussion of total survey design." Pp. 43–47 in U.S. National Center for Health Services Research, Health Survey Methods: Second Biennial Conference, 1977. DHEW Pub. No. (PHS) 79–3207. Hyattsville, Md.: NCHSR.

Hutcheson, J.D., Jr. and J.E. Prather
1977 "Assessing the effects of missing data." Pp. 279–83 in Proceedings, Social Statistics Section, American Statistical Association.

Hyman, H.H.
1954 Interviewing in Social Research. Chicago: University of Chicago Press.

Jackson, D.N. and S. Messick
1958 "Content and style in personality assessment." Psychological Bulletin 55 (July):243–52.

Jensen, L.E.
1979 "The social and economic milieu of health care provision: quality, access and cost." Journal of the American Medical Association 241 (March 30): 1345–47.

Johnson, D.R. and L.K. White
1979 "A comparison of telephone and personal interviewing for older populations." Unpublished manuscript, University of Nebraska, Lincoln.

Jordan, L.A.
1977 The Use of Covariance Structure Analysis for the Examination of Acquiescence and Other Response Style Issues in Faceted Test Data. Ph.D. dissertation, City University of New York.

Jöreskog, K.G.
1969 "A general approach to confirmatory maximum likelihood factor analysis." Psychometrika 34 (June):183–202.

Kalsbeek, W.D. and T.D. Hartwell
1977 "Head and spinal cord injuries: a pilot study of morbidity survey procedures." American Journal of Public Health 67 (November):1051–57.

Kalsbeek, W.D. and J.T. Lessler
1978 "Total survey design: effect of nonresponse bias and procedures for controlling measurement errors." Pp. 19–41 in U.S. National Center for Health Services Research, Health Survey Research Methods: Second Biennial Conference, 1977. DHEW Pub. No. (PHS) 79–3207. Hyattsville, Md.: NCHSR.

Kalsbeek, W.D. et al.
1975 Pilot Study of the Head and Spinal Cord Injury Survey. Research Monograph, Project 2554–1033. Research Triangle Park, N.C.: Research Triangle Institute.

Kanuk, L. and C. Berenson
1975 "Mail surveys and response rates: a literature review." Journal of Marketing Research 12 (November): 440–53.

Kaplan, B.
1975 "An epilogue: toward further research on family and health." Pp. 89–106 in B. Kaplan and J. Cassel (eds.), Family and Health: An Epidemiological Approach. Chapel Hill: Institute for Research in Social Science, University of North Carolina.

Kaplan, B., J. Cassel, and S. Gore
1977 "Social support and health." Medical Care 15 (May):Suppl. 47–58.

Kaplan, R.M., J.W. Bush, and C.C. Berry
1976 "Health status: types of validity and the index of well-being." Health Services Research 11 (Winter):478–507.

Katz, S. and C.A. Akpom
1976 "A measure of primary sociobiological functions." International Journal of Health Services 6 (3):493–508.

Kegeles, S.S., C.F. Fink, and J.P. Kirscht
1969 "Interviewing a national sample by long-distance telephone." Public Opinion Quarterly 33 (Fall):412–19.

Kephart, W.M. and M. Bressler
1958 "Increasing the responses to mail questionnaires: a research study." Public Opinion Quarterly 22 (Summer):123–32.

Kirscht, J.P., M.H. Becker, and J.P. Eveland
1976 "Psychological and social factors as predictors of medical behavior." Medical Care 14 (May):422–31.

Kish, L.
1965 Survey Sampling. New York: Wiley.

Kish, L. and M.R. Frankel
1970 "Balanced repeated replications for standard errors." Journal of the American Statistical Association 65 (September): 1071–94.

Klecka, W.R. and A.J. Tuchfarber
1978 "Random digit dialing: a comparison of personal surveys." Public Opinion Quarterly 42 (Spring):105–14.

Kohn, R. and K.L. White
1976 Health Care: An International Study. London: Oxford University Press.
Kosa, J., J.J. Alpert, and R.J. Haggerty
1967 "On the reliability of family health information." Social Science and Medicine 1 (July):165–81.
Krippendorff, K.
1970 "Bivariate agreement coefficients for reliability of data." Pp. 139–50 in E.F. Borgatta and C.W. Bohrnstedt (eds.), Sociological Methodology 1970. San Francisco: Jossey-Bass.
Kviz, F.J.
1977 "Toward a standard definition of response rate." Public Opinion Quarterly 41 (Summer):265–67.
Landis, J.R. and G.G. Koch
1977 "The measurement of observer agreement for categorical data." Biometrics 33 (March):159–74.
Langner, T.S.
1962 "A twenty-two item screening score of psychiatric symptoms indicating impairment." Journal of Health and Human Behavior 3 (Winter):269–76.
Larsen, O.N.
1952 "The comparative validity of telephone and face-to-face interviews in the measurement of message diffusion from leaflets." American Sociological Review 17 (August):471–76.
Laudan, L.
1977 Progress and Its Problems: Toward a Theory of Scientific Growth. Berkeley: University of California Press.
Leighton, D.C., J.S. Harding, D.B. Macklin, A.M. Macmillan, and A.H. Leighton
1963 The Character of Danger: Psychiatric Symptoms in Selected Communities. New York: Basic Books.
Lief, A. (ed.)
1948 The Common Sense Psychiatry of Dr. Adolf Meyer. New York: McGraw-Hill.
Lilienfeld, A.M.
1976 Foundations of Epidemiology. New York: Oxford University Press.
Lin, N., R.S. Simeone, W.M. Ensel, and W. Kuo
1979 "Social support, stressful life events, and illness: a model and an empirical test." Journal of Health and Social Behavior 20 (June):108–19.
Litwak, E.
1960a "Geographic mobility and extended family cohesion." American Sociological Review 25 (June): 385–94.
1960b "Occupational mobility and family cohesion." American Sociological Review 25 (February):9–21.
1961 "Voluntary associations and neighborhood cohesion." American Sociological Review 26 (April):258–71.
Locander, W., S. Sudman, and N. Bradburn
1976 "An investigation of interview method, threat and response distortion." Journal of the American Statistical Association 71 (June):269–75.
Lowenthal, M.F. and C. Haven
1968 "Interaction and adaptation: intimacy as a critical variable." American Sociological Review 33 (February): 20–30.
McKennell, A.
1970 "Attitude measurement: use of coefficient alpha with cluster or factor analysis." Sociology 4 (May):227–45.
Macmillan, A.M.
1957 "The Health Opinion Survey: technique for estimating prevalence of psychoneurotic and related types of disorder in communities." Psychological Reports 3 (September):325–29.
Mandell, L.
1974 "When to weight: determining nonresponse bias in survey data." Public Opinion Quarterly 38 (Summer): 247–52.
Market Opinion Research
1978 Issues in Health Care and Medicine. Chicago: MOR.
Market Research Society
1976 "Response rates in sample surveys: report of a working party of the Market Research Society's Research and Development Committee." Journal of the Market Research Society 18 (July): 113–42.
Markush, R.E. and R.V. Favero
1974 "Epidemiologic assessment of stressful life events, depressed mood, and psychophysiological symptoms: a preliminary report." Pp. 171–90 in B.S. Dohrenwend and B.P. Dohrenwend (eds.), Stressful Life Events: Their Nature and Effects. New York: Wiley.
Marquis, K.H.
1978a Record Check Validity of Survey Responses: A Reassessment of Bias in Reports of Hospitalizations. R-2319-HEW. Santa Monica, Calif.: Rand Corporation.
1978b "Survey response rates: some trends, causes, and correlates." Pp. 3–12 in U.S. National Center for Health Services Research, Health Survey Research Methods: Second Biennial Conference, 1977. DHEW Pub. No. (PHS) 79–3207. Hyattsville, Md.: NCHSR.

Marquis K.H. and C.F. Cannell
1969 A Study of Interviewer-Respondent Interaction in the Urban Employment Survey. Final report submitted to Manpower Administration, U.S. Department of Labor. Ann Arbor: Survey Research Center, University of Michigan.

Mechanic, D. and M. Newton
1965 "Some problems in the analysis of morbidity data." Journal of Chronic Disease 18 (June):569-80.

Medalie, J.H. and U. Goldbourt
1976 "Angina pectoris among 10,000 men: II. Psychosocial and other risk factors as evidenced by a multivariate analysis of a five-year incidence study." American Journal of Medicine 60 (May 31):910-21.

Meltzer, J.W. and J.R. Hochstim
1970 "Reliability and validity of survey data on physical health." Public Health Reports 85 (December):1075-86.

Mendenhall, R.C., R.A. Girard, and S. Abrahamson
1978 "A national study of medical and surgical specialties: I. Background, purpose, and methodology." Journal of the American Medical Association 240 (September 1):848-52.

Mendenhall, R.C., J.S. Lloyd, P.A. Repicky, J.R. Monson, R.A. Girard, and S. Abrahamson
1978 "A national study of medical and surgical specialties: II. Description of the survey instrument." Journal of the American Medical Association 240 (September 8):1160-68.

Menninger, K.
1963 The Vital Balance: The Life Process in Mental Health and Illness. New York: Viking Press.

Miller, P.V. and C.F. Cannell
1977 "Communicating measurement objectives in the survey interview." Pp. 127-51 in P. Hirsch, P.V. Miller, and F.G. Klein (eds.), Strategies for Communication Research. Beverly Hills, Calif.: Sage.

Miller, R.F.
1970 "Some ways of handling missing data: a case study." Pp. 177-97 in R.L. Bisco (ed.), Data Bases, Computers, and the Social Sciences. New York: Wiley.

Minor, M., J. Mullan, and J.D. Loft
1976 A Special Report on Cycles 1 and 2 of the National Ambulatory Medical Care Survey: A Methodological Evaluation and Analysis of Response Rates. Report to the National Center for Health Statistics. Chicago: National Opinion Research Center. Litho.

Mitchell, J.C.
1969 "The concept and use of social networks." Pp. 1-50 in J.C. Mitchell (ed.), Social Networks in Urban Situations. Manchester: Manchester University Press.

Moore, J.W.
1971 "Mexican Americans and cities: a study in migration and the use of formal resources." International Migration Review 5 (Fall):292-308.

Moriwaki, S.Y.
1973 "Self disclosure, significant others and psychological well being in old age." Journal of Health and Social Behavior 14 (September):226-32.

Muller, C.F., A. Waybur, and E.R. Weinerman
1952 "Methodology of a family health study." Public Health Reports 67 (November):1149-56.

Murphy, H.B.M.
1974 "Two stress measures in three cultures—their prognostic efficiency, significance and incongruities." In D. Leigh, J. Noorbakhsh, and C. Isadi (eds.), International Symposium on Epidemiological Studies in Psychiatry, Tehran.

Myers, J.
1976 "Future research in mental disease." Paper presented at the annual meetings of the American Sociological Association, New York.

Myers, J.K., J.J. Lindenthal, and M.P. Pepper
1975 "Life events, social integration and psychiatric symptomatology." Journal of Health and Social Behavior 16 (December):421-27.

Myers, J.K., J.J. Lindenthal, M.P. Pepper, and D.R. Ostrander
1972 "Life events and mental status: a longitudinal study." Journal of Health and Social Behavior 13 (December): 398-406.

Neter, J. and J. Waksberg
1964a "Conditioning effects from repeated household interviews." Journal of Marketing 28 (April):51-56.
1964b "A study of response errors in expenditure data from household interviews." Journal of the American Statistical Association 59 (March): 18-55.

Nuckolls, K.B., J. Cassel, and B.H. Kaplan
1972 "Psychosocial assets, life crisis and the prognosis of pregnancy." American Journal of Epidemiology 95 (May):431–41.

Nunnally, J.C.
1967 Psychometric Theory. New York: McGraw-Hill.

Oksenberg, L., A. Vinokur, and C.F. Cannell
1977 "The effects of commitment to being a good respondent on interviewer performance." Pp. 75–108 in U.S. National Center for Health Services Research, Experiments in Interviewing Techniques: Field Experiments in Health Reporting, 1971–1977. DHEW Pub. No. (HRA) 78-3204. Hyattsville, Md.: NCHSR. (PB 276 080, available NTIS only.)

O'Muircheartaigh, C.A.
1977 "Response errors." Pp. 193–239 in C.A. O'Muircheartaigh and C. Payne (eds.), The Analysis of Survey Data. Vol. 2: Model Fitting. New York: Wiley.

Oppenheim, A.N.
1966 Questionnaire Design and Attitude Measurement. New York: Basic Books.

Palmore, E. and C. Luikart
1972 "Health and social factors related to life satisfaction." Journal of Health and Social Behavior 13 (March): 68–80.

Parsons, T.
1958 "Definitions of health and illness in the light of American values and social structure." Pp. 165–87 in E.G. Jaco (ed.), Patients, Physicians and Illness: Sourcebook in Behavioral Science and Medicine. New York: Free Press of Glencoe.

Patrick, D.L.
1976 "Constructing social metrics for health status indexes." International Journal of Health Services 6 (Summer):443–53.
1979 Health and Care of the Physically Handicapped in Lambeth. London: Department of Community Medicine, St. Thomas's Hospital Medical School.

Patrick, D.L. and J. Elinson
1979 "Methods of sociomedical research." Pp. 437–59 in H.E. Freeman, S. Levine, and L.G. Reeder (eds.), Handbook of Medical Sociology. Englewood Cliffs, N.J.: Prentice-Hall.

Pearlin, L.I. and J.S. Johnson
1977 "Marital status, life-strains and depression." American Sociological Review 42 (October):704–15.

Peart, A.F.W.
1952 "Canada's Sickness Survey: review of methods." Canadian Journal of Public Health 43 (October):401–14.

Perrin, E.B., E.B. Harkins, and M.M. Marini
1978 Evaluation of the Reliability and Validity of Data Collection in the USC Medical Activities and Manpower Projects—Final Report. Seattle: Battelle Health and Population Study Center.

Rabkin, J.G. and E.L. Struening
1976 "Life events, stress, and illness." Science 194 (December 3):1013–20.

Radloff, L.S.
1977 "The CES-D Scale: a self-report depression scale for research in the general population." Applied Pscyhological Measurement 1 (Summer):385–401.

Rahe, R.
1975 "Epidemiologic studies of life change and illness." International Journal of Psychiatry in Medicine 6:133–46.

Reeder, L.G.
1976 "Recent literature concerning the use of the telephone in survey research." Institute for Social Science Research, University of California at Los Angeles.
1977 "Summary and conclusions." Pp. 1–3 in U.S. National Center for Health Services Research, Advances in Health Survey Research Methods: Proceedings of a National Invitational Conference, 1975. DHEW Pub. No. (HRA) 77–3154. Rockville, Md.: NCHSR. (PB 262 230, available NTIS only.)

Reissman, C.K.
1979 "Interviewer effects in psychiatric epidemiology: a study of medical and lay interviewers and their impact on reported symptoms." American Journal of Public Health 69 (May):485–91.

Robins, L.N.
1963 "The reluctant respondent." Public Opinion Quarterly 27 (Summer): 276–86.

Robinson, J.P.
1972 "Television's impact on everyday life: some cross-national evidence." Pp. 410–31 in Television and Social Behavior: A Technical Report to the Surgeon General's Scientific Advisory

Committee on Television and Social Behavior. Rockville, Md.: National Institute of Mental Health.

1977 How Americans Use Time: A Social-Psychological Analysis of Everyday Behavior. New York: Praeger.

Rogers, T.F.

1976 "Interviews by telephone and in person: quality of responses and field performance." Public Opinion Quarterly 40 (Spring): 51–65.

Roghmann, K.J. and R.J. Haggerty

1972 "The diary as a research instrument in the study of health and illness behavior." Medical Care 10 (March-April):143–63.

Roper, B.W.

1971 An Extended View of Public Attitudes toward Television and Other Mass Media. New York: Television Information Office.

Ross, C.E. and J. Mirowsky II

1979 "A comparison of life-event-weighting schemes: change, undesirability, and effect-proportional indices." Journal of Health and Social Behavior 20 (June):166–77.

Rothwell, N. and G. Bridge

1978 "Discussion of respondent burden." Pp. 55–61 in U.S. National Center for Health Services Research, Health Survey Research Methods: Second Biennial Conference, 1977. DHEW Pub. No. (PHS) 79-3207. Hyattsville, Md.: NCHSR.

Schmale, A.H.

1972 "Giving up as a final common pathway to changes in health." Pp. 20–40 in Z.J. Lipowski (ed.), Psychosocial Aspects of Physical Illness. Advances in Psychosomatic Medicine, Vol. 8. Basel: S. Karger.

Schmiedeskamp, J.W.

1962 "Reinterviews by telephone." Journal of Marketing 26 (January): 28–34.

Schofield, W.

1964 Psychotherapy: The Purchase of Friendship. Englewood Cliffs, N.J.: Prentice-Hall.

Schuman, H. and S. Presser

1977 "Question wording as an independent variable in survey analysis." Sociological Methods & Research 6 (November):151–70.

Schwartz, C.C., B.M. Astrachan, and J.K. Myers

1973 "Comparing three measures of mental status: a note on the validity of estimates of psychological disorder in the community." Journal of Health and Social Behavior 14 (September): 265–73.

Seiler, L.H.

1973 "The 22-item scale used in field studies of mental illness: a question of method, a question of substance, and a question of theory." Journal of Health and Social Behavior 14 (September): 252–64.

Shapiro, S., R. Yaffe, R.R. Fuchsberg, and H.C. Corpeño

1976 "Medical Economics Survey-Methods Study: design, data collection, and analytical plan." Medical Care 14 (November):893–912.

Sheatsley, P.B. and J.D. Loft

1977 Expansion of the National Ambulatory Medical Care Survey To Include Data on Product-Related Accidents and Illnesses. Report to the National Center for Health Statistics. Chicago: National Opinion Research Center. Litho.

Sheatsley, P.B., S. Scharf, and J.D. Loft

1977 A Report on a Feasibility Study To Extend the National Ambulatory Medical Care Survey to Hospital Outpatient Clinics. Report to the National Center for Health Statistics. Chicago: National Opinion Research Center. Litho.

Shure, G.H. and R.J. Meeker

1978 "A minicomputer system for multiperson computer-assisted telephone interviewing." Behavior Research Methods and Instrumentation 10 (April):196–202.

Siemiatycki, J.

1979 "A comparison of mail, telephone, and home interview strategies for household health surveys." American Journal of Public Health 69 (March): 238–45.

Sirken, M.G.

1974 "The counting rule strategy in sample surveys." Pp. 119–23 in Proceedings, Social Statistics Section, American Statistical Association.

1975 "Discussion of medical provider surveys of neurological conditions." Pp. 55–58 in Proceedings, Social Statistics Section, American Statistical Association.

1979 "Sample survey estimators of drug use." Unpublished manuscript.

Sirken, M.G., B.I. Graubard, and M.J. McDaniel

1978 "National network surveys of diabetes." Pp. 631–35 in Proceedings, Survey Research Methods Section, American Statistical Association.

Sirken, M.G. and P.N. Royston
1976 "Design effects in retrospective mortality surveys." Pp. 773–77 in Proceedings, Social Statistics Section, American Statistical Association.
1977 "Counting rule bias in household survey of deaths." Pp. 347–51 in Proceedings, Social Statistics Section, American Statistical Association.

Sloan, F., J. Cromwell, and J. Mitchell
1977 A Study of Administrative Costs in Physicians' Offices and Medicaid Participation: Final Report. Cambridge, Mass.: Abt Publications.

Spitzer, R.L., J. Endicott, J.L. Fleiss, and J. Cohen
1970 "The psychiatric status schedule: a technique for evaluating psychopathology and impairment in role functioning." Archives of General Psychiatry 23:41–55.

Spitzer, R.L., J. Endicott, and E. Robins
1978 "Research diagnostic criteria: rationale and reliability." Archives of General Psychiatry 35 (June):773–82.

Sprately, E.
1973 "An analysis of the efficiency with which Health Interview Survey data were collected during the first three quarters of 1972." Unpublished trainee report, U.S. National Center for Health Statistics.

Srole, L., T.S. Langner, S.T. Michael, M.K. Opler, and T.A.C. Rennie
1962 Mental Health in the Metropolis: the Midtown Manhattan Study. New York: McGraw-Hill.

Stephens, R.C.
1979 "Comparison of telephone and face-to-face interviewing techniques among older respondents." Unpublished manuscript, University of Texas, Houston.

Sudman, S.
1966 "Quantifying interviewer quality." Public Opinion Quarterly 30 (Winter):664–67.
1967 Reducing the Cost of Surveys. Chicago: Aldine.
1976 Applied Sampling. New York: Academic Press.

Sudman, S. and N.M. Bradburn
1974 Response Effects in Surveys: A Review and Synthesis. Chicago: Aldine.

Sudman, S. and L.B. Lannom
1979 A Comparison of Alternative Panel Procedures for Obtaining Health Data. Revised. Urbana: Survey Research Laboratory, University of Illinois.

Sudman, S., W. Wilson, and R. Ferber
1976 The Cost-Effectiveness of Using the Diary as an Instrument for Collecting Health Data in Household Surveys. Report to the Bureau of Health Services Research and Evaluation. Revised. Urbana: Survey Research Laboratory, University of Illinois.

Sumner, J.
1978 The 1976 LAMAS Frame and Master Sample: Technical Description. Los Angeles: Institute for Social Science Research, University of California at Los Angeles.

Taylor, D.G.
1976 "The accuracy of respondent-coded occupation." Public Opinion Quarterly 40 (Summer):245–55.

Theodore, C.N. and G.E. Sutter
1967 "A report on the first periodic survey of physicians." Journal of the American Medical Association 202 (November 6):516–24.

Thornberry, O.T., Jr. and J.T. Massey
1978 "Correcting for undercoverage bias in random digit dialed national health surveys." Pp. 224–29 in Proceedings, Survey Research Methods Section, American Statistical Association.

Turner, R.
1961 "Inter-week variations in expenditure recorded during a two-week survey of family expenditure." Applied Statistics 10 (November):136–46.

U.S. Bureau of the Census
1968 Methodology of Consumer Expenditure Surveys, by R.B. Pearl. Working Paper No. 27. Washington, D.C.: U.S. Bureau of the Census.

U.S. National Center for Health Services Research
1977 Experiments in Interviewing Techniques: Field Experiments in Health Reporting, 1971–1977, by C.F. Cannell, L. Oksenberg, and J.M. Converse (eds.). DHEW Pub. No. (HRA) 78–3204. Hyattsville, Md.: NCHSR. (PB 276 080, available NTIS only.)

U.S. National Center for Health Statistics
1965a Comparison of Hospitalization Reporting in Three Survey Procedures, by C.F. Cannell and F. Fowler. Vital and Health Statistics, Series 2, No. 8. Washington, D.C.: U.S. Government Printing Office.
1965b Design of Sample Surveys To Esti-

mate the Prevalence of Rare Diseases: Three Unbiased Estimates, by Z.W. Birnbaum and M.G. Sirken. Vital and Health Statistics, Series 2, No. 11. Washington, D.C.: U.S. Government Printing Office.

1965c Health Interview Responses Compared with Medical Records, by E. Balmuth. Vital and Health Statistics, Series 2, No. 7. Washington, D.C.: U.S. Government Printing Office.

1965d Reporting of Hospitalization in the Health Interview Survey, by C.F. Cannell, G. Fisher, and T. Bakker. Vital and Health Statistics, Series 2, No. 6. Washington, D.C.: U.S. Government Printing Office.

1967 Interview Data on Chronic Conditions Compared with Information Derived from Medical Records, by W.G. Madow. Vital and Health Statistics, Series 2, No. 23. Washington, D.C.: U.S. Government Printing Office.

1972 Reporting Health Events in Household Interviews, by A. Laurent, C.F. Cannell, and K.H. Marquis. Vital and Health Statistics, Series 2, No. 49. Washington, D.C.: U.S. Government Printing Office.

1974 National Ambulatory Medical Care Survey: Background and Methodology, United States—1976-72, by J.B. Tenney, K.L. White, and J.W. Williamson. Vital and Health Statistics, Series 2, No. 61. Washington, D.C.: U.S. Government Printing Office.

1975 Health Interview Survey Procedure, 1957-74. Vital and Health Statistics, Series 1, No. 11. Washington, D.C.: U.S. Government Printing Office.

1977 A Summary of Studies of Interviewing Methodology, by C.F. Cannell, K.H. Marquis, and A. Laurent. Vital and Health Statistics, Series 2, No. 69. Washington, D.C.: U.S. Government Printing Office.

U.S. Office of Federal Statistical Policy and Standards

1978 An Error Profile: Employment as Measured by the Current Population Survey, by C.A. Brooks and B.A. Bailar. Statistical Policy Working Paper 3. Washington, D.C.: U.S. Government Printing Office.

U.S. Public Health Service

1962 Methodology in Two California Health Surveys, by H.W. Mooney. Public Health Monograph No. 70. Washington, D.C.: U.S. Government Printing Office.

U.S. Social Security Administration, Office of Research and Statistics

1977 1975 Net Incomes and Work Patterns of Physicians in Five Medical Specialties, by N. Thorndike. Research and Statistics Note, No. 13. Washington, D.C.: SSA.

Verbrugge, L.M.

1978 "Health diaries." Pp. 271-76 in Proceedings, Survey Research Methods Section, American Statistical Association.

1979 "Female illness rates and illness behavior: testing hypotheses about sex differences in health." Women and Health 4 (Spring): 61-79.

1980 "Health diaries." Medical Care 18 (January):73-95.

Walden, D.C.

1975 "The use of a panel design in survey research to obtain health care utilization and expenditure data: the experience of the Federal Employees Health Benefits Program Utilization Study." Pp. 273-80 in Proceedings, Public Health Conference on Records and Statistics, 1974. DHEW Pub. No. (HRA) 75-1214. Rockville, Md.: National Center for Health Statistics.

Way, P.O., L.E. Jensen, and L.J. Goodman

1978 "Foreign medical graduates and the issue of substantial disruption of medical services." New England Journal of Medicine 299 (October 5): 745-51.

Weissman, M., J.K. Myers, and P.S. Harding

1978 "Psychiatric disorders in a U.S. urban community: 1975-76." American Journal of Psychiatry 135 (April): 459-62.

Werner, J., W. Wendling, and N. Budde

1979 "Determinants of the physician's location choice." Working paper, American Medical Association, Chicago.

Wilcox, K.R., Jr.

1963 Comparison of Three Methods for the Collection of Morbidity Data by Household Survey. Ph.D. dissertation, Department of Epidemiology, University of Michigan.

Williams, E.

1977 "Experimental comparisons of face-to-face and mediated communication: a review." Psychological Bulletin 84 (September):963-76.

Williams, R.G.A.
1979 "Theories and measurement in disability." Epidemiology and Community Health 33:32-47.

Wiseman, F.
1972 "Methodological bias in public opinion surveys." Public Opinion Quarterly 36 (Spring):105-8.

Wood, P.H.N.
1975 Classification of Impairments and Handicaps. WHO/ICD9/REV. CONF./75.13. Geneva: World Health Organization.

Wright, R.A., R.H. Beisel, J.D. Oliver, and M.C. Gerzowski
1976 "The use of a multiple entry diary in a panel study on health care expenditure." Pp. 848-52 in Proceedings, Social Statistics Section, American Statistical Association.

Yaffe, R. and S. Shapiro
1979 "Reporting accuracy of health care utilization and expenditures in a household survey as compared with provider records and insurance claims records." Paper presented at spring meetings of the Biometrics Society, Eastern North American Region, New Orleans.

Yaffe, R., S. Shapiro, R.R. Fuchsberg, C.A. Rohde, and H.C. Corpeño
"Medical Economics Survey-Methods Study: cost-effectiveness of alternative survey strategies." Medical Care 16 (August):641-59.

Yett, D.
1977 "Validation of health manpower data." Pp. 194-202 in Proceedings, Public Health Conference on Records and Statistics, 1976. DHEW Pub. No. (HRA) 77-1214. Rockville, Md.: National Center for Health Statistics.

Conference participants

Lu Ann Aday
Center for Health Administration Studies
University of Chicago
5720 S. Woodlawn Avenue
Chicago, IL 60637

Ronald Andersen
Center for Health Administration Studies
University of Chicago
5720 S. Woodlawn Avenue
Chicago, IL 60637

Frank M. Andrews
Survey Research Center
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106

Morris Axelrod
Survey Research Laboratory
Department of Sociology
Arizona State University
Tempe, AZ 85281

Marilyn Bergner
Department of Health Services
School of Public Health and Community Medicine
University of Washington
Seattle, WA 98195

Marc Berk
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

Lisa F. Berkman
Department of Epidemiology and Public Health
Yale School of Medicine
60 College Street
New Haven, CT 06520

Laurence G. Branch
Department of Preventive and Social Medicine
Harvard Medical School
25 Shattuck Street
Boston, MA 02115

Charles F. Cannell
Survey Research Center
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106

Steven B. Cohen
National Center for Health Services Research
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Diana Cook
Department of Psychiatry
Columbia University
New York, NY 10032

Alfred Dean
Department of Psychiatry
Albany Medical College
New Scotland Avenue
Albany, NY 12208

Joseph L. de la Puente
Health Services Research Methods/Evaluation Cluster
National Center for Health Services Research
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Charlene E. Depner
Survey Research Center
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106

Carole D. Dillard
National Center for Health Services Research
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Barbara S. Dohrenwend
Division of Sociomedical Sciences
School of Public Health
Columbia University
600 West 168th Street
New York, NY 10032

Bruce P. Dohrenwend
Department of Psychiatry
Columbia University
New York, NY 10032

Jack Elinson
Division of Sociomedical Sciences
School of Public Health
Columbia University
60 Haven Avenue
New York, NY 10032

Walter M. Ensel
Department of Sociology
State University of New York at Albany
Albany, NY 12222

Lynn A. Evans
Department of Community Medicine
Baylor College of Medicine
Texas Medical Center
Houston, TX 77030

Esther Fleishman
National Opinion Research Center
817 Broadway
New York, NY 10003

Floyd J. Fowler, Jr.
Center for Survey Research
University of Massachusetts
100 Arlington Street
Boston, MA 02116

Matilda Frankel
Survey Research Laboratory
University of Illinois at Urbana-Champaign
1005 W. Nevada Street
Urbana, IL 61801

Howard E. Freeman
Institute for Social Science Research
University of California at Los Angeles
Los Angeles, CA 90024

Deborah Freund
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

Robert R. Fuchsberg
Division of Health Interview Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Helen C. Gift
Bureau of Economic and Behavioral Research
American Dental Association
211 E. Chicago Avenue, Suite 2001
Chicago, IL 60611

Louis J. Goodman
Health Services Research and Development
American Medical Association
535 N. Dearborn Street
Chicago, IL 60610

Bernard G. Greenberg
School of Public Health
University of North Carolina
Chapel Hill, NC 27514

Robert M. Groves
Survey Research Center
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106

Lawrence Haber
Office of Federal Statistical Policy and Standards
2001 S Street, N.W.
Washington, DC 20230

Ruth S. Hanft
Office of the Assistant Secretary for Health
Department of Health, Education, and Welfare
703-H Hubert Humphrey Building
Washington, DC 20201

Elizabeth B. Harkins
Health and Population Study Center
Battelle Human Affairs Research Centers
4000 N.E. 41st Street
Seattle, WA 98105

Benjamin S.H. Harris
Survey Operations Center
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Mimi Holt
Survey Operations Center
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Daniel G. Horvitz
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Morton Israel
Office of Biostatistics
Department of Health
125 Worth Street
New York, NY 10013

Robert A. Israel
National Center for Health Statistics
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Lynn E. Jensen
Center for Health Services Research and Development
American Medical Association
535 N. Dearborn Street
Chicago, IL 60610

Charles D. Jones
Statistical Methods Division
Bureau of the Census
Washington, DC 20233

Gareth Jones
Statistics Canada
Rm. A234, 8 Temporary Building
Carling Avenue
Ottawa, Ontario, Canada

Lawrence A. Jordan
System Development Corporation
2500 Colorado Avenue
Santa Monica, CA 90406

William D. Kalsbeek
Department of Biostatistics
School of Public Health
University of North Carolina
Chapel Hill, NC 27514

Judith Kasper
National Center for Health Services Research
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

William M. Kitching
National Center for Health Services Research
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Nan Lin
Department of Sociology
State University of New York at Albany
1400 Washington Avenue
Albany, NY 12222

John D. Loft
National Opinion Research Center
University of Chicago
6030 S. Ellis Avenue
Chicago, IL 60637

Alfred C. Marcus
School of Public Health
University of California at Los Angeles
Los Angeles, CA 90024

Kent H. Marquis
The Rand Corporation
1700 Main Street
Santa Monica, CA 90406

James T. Massey
Division of Health Interview Statistics
National Center for Health Statistics
3700 East-West Highway
Hyattsville, MD 20782

Linda S. McCleary
National Center for Health Services Research
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Ian McDowell
Epidemiology and Community Medicine
Royal Ottawa Hospital
Faculty of Medicine of University of Ottawa
1145 Carling Avenue
Ottawa, Ontario, Canada K1Z 7K4

Samuel M. Meyers
National Center for Health Services Research
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Peter V. Miller
Institute of Communications Research
University of Illinois at Urbana-Champaign
222B Armory
Champaign, IL 61820

Lois Monteiro
Sociology and Community Health
Box 1916
Brown University
Providence, RI 02912

Marco Montoya
Health Services Research Study Section
National Center for Health Services Research
Center Building
3700 East-West Highway
Hyattsville, MD 20782

R. Paul Moore
Research Triangle Institute
P.O. Box 12194
Research Triangle Park, NC 27709

Lois Oksenberg
National Center for Health Services Research
3700 East-West Highway
Hyattsville, MD 20782

Donald L. Patrick
Department of Community Medicine
St. Thomas's Hospital Medical School
University of London
London SE1 7EH, England

Clyde Pope
Health Services Research Center
Kaiser Foundation Hospitals
4610 S.E. Belmont Street
Portland, OR 97215

Sharon Reeder
School of Nursing
Center for the Health Sciences
University of California at Los Angeles
Los Angeles, CA 90024

Gerald Rosenthal
National Center for Health Services Research
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Louis F. Rossiter
National Center for Health Services Research
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

George S. Rothbart
Center for Policy Research
475 Riverside Drive
New York, NY 10027

Naomi D. Rothwell
Statistical Research Division
Bureau of the Census
Washington, DC 20233

Sam Shapiro
Health Services Research and Development Center
Johns Hopkins Medical Institutions
624 N. Broadway
Baltimore, MD 21205

Patrick E. Shrout
School of Public Health
Columbia University
600 West 168th Street
New York, NY 10032

Eleanor Singer
500 Journalism Building
Columbia University
New York, NY 10027

Monroe G. Sirken
National Center for Health Statistics
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Marian Solomon
Office of Planning and Evaluation
Department of Health, Education, and Welfare
419E Hubert Humphrey Building
Washington, DC 20201

Seymour Sudman
Survey Research Laboratory

University of Illinois at Urbana-Champaign
1005 W. Nevada Street
Urbana, IL 61801

D. Garth Taylor
National Opinion Research Center
6030 S. Ellis Avenue
Chicago, IL 60637

Owen Thornberry
National Center for Health Statistics
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Lois M. Verbrugge
Survey Research Center
University of Michigan
P.O. Box 1248
Ann Arbor, MI 48106

Daniel C. Walden
National Center for Health Services Research
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Thomas T.H. Wan
Department of Sociology
University of Maryland
Baltimore County Campus
5401 Wilkins Avenue
Baltimore, MD 21228

Richard B. Warnecke
Survey Research Laboratory
4011 Behavioral Sciences Building
University of Illinois at Chicago Circle
Chicago, IL 60608

Gail R. Wilensky
National Center for Health Services Research
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Sherman R. Williams
National Center for Health Services Research
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Robert Wright
National Center for Health Statistics
Department of Health, Education, and Welfare
3700 East-West Highway
Hyattsville, MD 20782

Richard Yaffe
Health Services Research and Development Center
Johns Hopkins Medical Institutions
624 N. Broadway
Baltimore, MD 21205

Pearl Zinner
National Opinion Research Center
817 Broadway
New York, NY 10003

Current NCHSR Publications

National Center for Health Services Research publications of interest to the health community are available on request to NCHSR, Publications and Information Branch, 3700 East-West Highway, Room 7-44, Hyattsville, MD 20782 (telephone: 301/436-8970). Mail requests will be facilitated by enclosure of a self-addressed, adhesive-backed mailing label. These publications also are available for sale through the National Technical Information Service (NTIS), Springfield, VA 22161 (telephone: 703-487-4650). PB and HRP numbers in parentheses are NTIS order numbers. Publications which are out of stock in NCHSR are indicated as available only from NTIS. Prices may be obtained from the NTIS order desk on request.

Research Digests

The *Research Digest Series* provides overviews of significant research supported by NCHSR. The series describes either ongoing or completed projects directed toward high priority health services problems. Issues are prepared by the principal investigators performing the research, in collaboration with NCHSR staff. Digests are intended for an interdisciplinary audience of health services planners, administrators, legislators, and others who make decisions on research applications.

(HRA) 76-3144	Evaluation of a Medical Information System in a Community Hospital (PB 264 353, available NTIS only)
(HRA) 76-3145	Computer-Stored Ambulatory Record (COSTAR) (PB 268 342)
(HRA) 77-3160	Program Analysis of Physician Extender Algorithm Projects (PB 264 610, available NTIS only)
(HRA) 77-3161	Changes in the Costs of Treatment of Selected Illnesses, 1951-1964-1971 (HRP 0014598, available NTIS only)
(HRA) 77-3163	Impact of State Certificate-of-Need Laws on Health Care Costs and Utilization (PB 264 352, available NTIS only)
(HRA) 77-3164	An Evaluation of Physician Assistants in Diagnostic Radiology (PB 266 507, available NTIS only)
(HRA) 77-3166	Foreign Medical Graduates: A Comparative Study of State Licensure Policies (PB 265 233, available NTIS only)
(HRA) 77-3171	Analysis of Physician Price and Output Decisions (PB 273 312)
(HRA) 77-3173	Nurse Practitioner and Physician Assistant Training and Deployment (PB 271 000, available NTIS only)
(HRA) 77-3177	Automation of the Problem-Oriented Medical Record (PB 266 881, available NTIS only)
(PHS) 78-3190	Uncertainties of Federal Child Health Policies: Impact in Two States (PB 283 202)
(PHS) 80-3229	Responses of Canadian Physicians to the Introduction of Universal Medical Care Insurance: The First Five Years in Quebec (PB 80-137 979)
(PHS) 79-3231	Israel Study of Socialization for Medicine (PB 293 887)
(PHS) 79-3235	AAMC Longitudinal Study of Medical School Graduates of 1960 (PB 294 689)
(PHS) 79-3238	Some Effects of Quebec Health Insurance (PB 294 097)
(PHS) 79-3261	Medical Education Financing: Issues and Options (PB 80-134 851)

Research Summaries

The *Research Summary Series* provides rapid access to significant results of NCHSR-supported research projects. The series presents executive summaries prepared by the investigators. Specific findings are highlighted in a more concise form than in the final report. The *Research Summary Series* is intended for health services administrators, planners, and other research users who require recent findings relevant to immediate programs in health services.

(HRA) 77-3162	Recent Studies in Health Services Research, Vol. I (July 1974 through December 1976) (PB 266 460)
(HRA) 77-3176	Quality of Medical Care Assessment Using Outcome Measures (PB 272 455)
(HRA) 77-3183	Recent Studies in Health Services Research, Vol. II (CY 1976) (PB 279 198)
(PHS) 78-3187	Criterion Measures of Nursing Care Quality (PB 287 449)

(PHS) 77-3188 Demonstration and Evaluation of a Total Hospital Information System (PB 271 079)

(PHS) 78-3192 Assessing the Quality of Long-Term Care (PB 293 473)

(PHS) 78-3193 Optimal Electrocardiography (PB 281 558)

(PHS) 78-3201 A National Profile of Catastrophic Illness (PB 287 291)

(PHS) 79-3223 Medical Direction in Skilled Nursing Facilities (PB 300 845)

(PHS) 79-3230 Per-Case Reimbursement for Medical Care (PB 294 688)

(PHS) 79-3236 Nurse Practitioners and Physician Assistants: A Research Agenda (PB 294 084)

(PHS) 80-3244 Quality of Ambulatory Care

(PHS) 79-3247 A Demonstration of PROMIS

(PHS) 79-3248 Effects of the 1974-75 Recession on Health Care for the Disadvantaged (PB 80-138 449)

(PHS) 79-3250 Effects and Costs of Day Care and Homemaker Services for the Chronically Ill: A Randomized Experiment (PB 301 171)

(PHS) 80-3265 How a Medical Information System Affects Hospital Costs: The El Camino Hospital Experience

(PHS) 80-3266 Developing a Quality Assurance Strategy for Primary Care (PB 80-162 878)

Research Reports

The *Research Report Series* provides significant research reports in their entirety upon the completion of the project. Research Reports are developed by the principal investigators who conducted the research, and are directed to selected users of health services research as part of a continuing NCHSR effort to expedite the dissemination of new knowledge resulting from its project support.

(HRA) 77-3152 How Lawyers Handle Medical Malpractice Cases (HRP 0014313)

(HRA) 77-3159 An Analysis of the Southern California Arbitration Project, January 1966 through June 1975 (HRP 0012466)

(HRA) 77-3165 Statutory Provisions for Binding Arbitration of Medical Malpratice Cases (PB 264 409, available NTIS only)

(HRA) 77-3184 1960 and 1970 Spanish Heritage Population of the Southwest by County (PB 280 656, available NTIS only)

(HRA) 77-3189 Drug Coverage under National Health Insurance: The Policy Options (PB 272 074)

(PHS) 78-3204 Experiments in Interviewing Techniques: Field Experiments in Health Reporting (PB 276 080, available NTIS only)

(PHS) 79-3210 Telehealth Handbook: A Guide to Telecommunications Technology for Rural Health Care, (PB 292 557, available NTIS only)

(PHS) 78-3211 Emergency Medical Technician Performance Evaluation (PB 285 961)

(PHS) 79-3217-1 Evaluation of Child Abuse and Neglect Demonstration Projects, 1974-77, Vols. 1 and 2 (PB 278 438 and 278 439; vols. 1-12, PB 278 437, the set)

(PHS) 78-3219 Needed Research in the Assessment and Monitoring of the Quality of Medical Care (PB 288 826)

(PHS) 79-3237 A Cost-Effective Approach to Cervical Cancer Detection (PB 295 515)

(PHS) 79-3245 Costs and Benefits of Electronic Fetal Monitoring: A Review of the Literature (PB 294 690)

(PHS) 79-3251 Computer Applications in Health Care (PB 300 838)

(PHS) 79-3258 Effects and Costs of Day Care and Homemaker Services for the Chronically Ill; A Randomized Experiment (PB 80-138 100)

(PHS) 80-3263 Health Status, Medical Care Utilization, and Outcome: An Annotated Bibliography of Empirical Studies; Parts I-IV (PB 284 997, available NTIS only)

Research Proceedings

The *Research Proceedings Series* extends the availability of new research announced at key conferences, symposia and seminars spnsored or supported by NCHSR. In addition to papers presented, publications in this series include discussions and responses whenever possible. The series is intended to help meet the information needs of health services providers and others who require direct access to concepts and ideas evolving from the exchange of research results.

(HRA) 77-3154 Advances in Health Survey Research Methods (PB 262 230, available NTIS only)

(HRA) 77-3181 NCHSR Research Conference Report on Consumer Self-Care in Health (PB 273 811)

(HRA) 77-3186 International Conference on Drug and Pharmaceutical Services Reimbursement (PB 271 386)

(PHS) 78-3195 Emergency Medical Services: Research Metodology (PB 279 096)

(PHS) 78-3207 Health Survey Research Methods, Second Biennial Conference (PB 293 492)

(PHS) 78-3208 Drug Coverage Under National Health Insurance (PB 293 468)

(PHS) 79-3209 Health Services Research in Puerto Rico (PB 292 326)

(PHS) 80-3215 Cost Accounting for Pharmaceutical Services (PB 80-157 936)

(PHS) 79-3216 Medical Technology: The Culprit Behind Health Care Costs? (PB 299 408)

(PHS) 79-3225-1 Emergency Medical Services Research Methodology: Workshop 1 (PB 294 048)

(PHS) 79-3225-2 Emergency Medical Services Research Methodology: Workshop 2 (PB 80-142 292)

(PHS) 78-3227 Effects of the Payment Mechanism on the Health Care Delivery System (PB 291 231)

(PHS) 79-3228 A National Conference on Health Policy, Planning, and Financing the Future of Health Care for Blacks in America (PB 292 559)

(PHS) 79-3233 Emergency Medical Services Systems as a Health Services Research Setting (PB 297 102)

(PHS) 79-3254 Medical Technology (PB 80-149 511)

(PHS) 79-3256 Sharing Health Care Costs (PB 80 162 795)

(PHS) 79-3257 Health Facility Reuse, Retrofit, and Reconfiguration (PB 80-142 383)

Research Management

The *Research Management Series* describes programmatic rather than technical aspects of the NCHSR research effort. Information is represented on the NCHSR goals, research objectives, and priorities; in addition, this series contains lists of grants and contracts, and administrative information on funding. Publications in this series are intended to bring basic information on NCHSR and its programs to research planners, administrators, and others who are involved with the allocation of research resources.

(PHS) 79-3220 Emergency Medical Services Systems Research Projects, 1978 (PB 292 558)

(PHS) 80-3271 Emergency Medical Services Systems Research Projects Abstracts, 1979

NHCES

The *National Health Care Expenditures Study Series* presents information and analyses on critical national health policy issues. Basic data were obtained from the National Medical Care Expenditure Survey, a statistical picture of how health services are used and paid for. Data Previews give preliminary estimates of key measures.

(PHS) 80-3276 Data Preview 1: Who are the Uninsured?

(PHS) 80-3275 Data Preview 2: Charges and Sources of Payment for Dental Visits with Separate Charges

Policy Research

The *Policy Research Series* describes findings from the research program that have major significance for policy issues of the moment. These papers are prepared by members of the staff of NCHSR or by independent investigators. The series is intended specifically to inform those in the public and private sectors who must consider, design, and implement policies affecting the delivery of health services.

(HRA) 77-3182 Controlling the Cost of Health Care (PB 266 885)